LECTURE NOTES IN ELECTRICAL ENGINEERING 69

Vincent Acary · Olivier Bonnefon
Bernard Brogliato

Nonsmooth Modeling and Simulation for Switched Circuits

Springer

Lecture Notes in Electrical Engineering

Volume 69

For other titles published in this series, go to
www.springer.com/series/7818

Vincent Acary · Olivier Bonnefon ·
Bernard Brogliato

Nonsmooth Modeling and Simulation for Switched Circuits

Springer

Vincent Acary
INRIA Rhône-Alpes
avenue de l'Europe 655
38334 Saint-Ismier
France
vincent.acary@inrialpes.fr

Bernard Brogliato
INRIA Rhône-Alpes
avenue de l'Europe 655
38334 Saint-Ismier
France
bernard.brogliato@inrialpes.fr

Olivier Bonnefon
INRIA Rhône-Alpes
avenue de l'Europe 655
38334 Saint-Ismier
France
olivier.bonnefon@inrialpes.fr

ISSN 1876-1100 e-ISSN 1876-1119
ISBN 978-90-481-9680-7 e-ISBN 978-90-481-9681-4
DOI 10.1007/978-90-481-9681-4
Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: 2010938003

Cover design: VTEX, Vilnius

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The major aim of this monograph is to show that the nonsmooth dynamics framework (involving keywords like complementarity problems, piecewise-linear characteristics, inclusions into normal cones, variational inequalities, multivalued characteristics) is a convenient and efficient way to handle *analog* switched circuits. It has been long known in the circuits community that such nonsmooth switched systems are difficult to simulate numerically, for various reasons that will be recalled. In parallel the simulation of nonsmooth mechanical systems (*i.e.* mainly mechanical systems with nonsmooth interface or contact laws, like unilateral constraints, impacts, Coulomb's friction and its many extensions, *etc.*) has been the object of a lot of research studies (see for instance the recent monographs Acary and Brogliato 2008 and Studer 2009). This field has now reached a certain degree of maturity, and has proved to be a quite useful and efficient approach in many areas of mechanics. Here we would like to show that the tools that have been employed in the contact mechanics context, can be successfully extended to the simulation of analog switched circuits. To the best of our knowledge it is the first time that such extensive numerical simulations using the nonsmooth dynamics framework for analog switched circuits are presented and published.

Montbonnot

Vincent Acary
Olivier Bonnefon
Bernard Brogliato

Acknowledgements

The material presented in this book is the result of on-going works led in the BIPOP research team of the INRIA Grenoble Rhône-Alpes, and is directly linked with the development of the open-source software package SICONOS (http://siconos.gforge.inria.fr/). In particular the numerical results owe a lot to Pascal Denoyelle, who worked in the BIPOP research team as an expert engineer from March 2005 to February 2008.

Contents

Part I Theoretical Framework

Part III Numerical Simulations

List of Abbreviations

ACEF	Automatic Circuit Equations Formulation
API	Application Programming Interface
BCE	Branch Constitutive Equation
BDF	Backward Differentiation Formulas
CCP	Cone Complementarity Problem
DAE	Differential Algebraic Equation
KCL	Kirchhoff’s Current Laws
KVL	Kirchhoff’s Voltage Laws
LCP	Linear Complementarity Problem
LTI	Linear Time Invariant
MCP	Mixed Complementarity Problem
MDE	Measure Differential Equation
MDI	Measure Differential Inclusion
MLCP	Mixed Linear Complementarity Problem
MNA	Modified Nodal Analysis
NCP	Nonlinear Complementarity Problem
NSDS	NonSmooth Dynamical Systems
ODE	Ordinary Differential Equation
OSNSP	Onestep NonSmooth Problem
STA	Sparse Tableau Analysis

List of Algorithms

List of Figures

List of Tables

Part I
Theoretical Framework

This part is mainly dedicated to introduce the basic mathematical tools which are needed to correctly understand what the nonsmooth dynamical systems (NSDS) approach is. Some basic ingredients from convex analysis, complementarity theory, variational inequalities, are indeed necessary if one wants to go further. In this part we will also present a short history of *nonsmooth* models and their numerical simulation in electrical circuits. They are better known as *piecewise-linear* representations in the circuits community.

Chapter 1
Introduction to Switched Circuits

> *Good models should describe the real physics only as far as needed and should not carry too much additional ballast, which would slow down the numerical processes necessary to solve them, but would also obscure the desired results. . . It is a matter of fact that the more compact mathematical formulations yield at the end the better numerical codes.*
>
> *C. Glocker in* Glocker (2005)

1.1 Simple Examples of Switched Circuits

This section is dedicated to introduce simple circuits that contain electronic devices with a nonsmooth current/voltage characteristic. Examples are RLC circuits with so-called ideal diodes, ideal Zener diodes, ideal switches. The main peculiarities of their dynamics are highlighted through detailed analysis. The parallel with simple nonsmooth mechanical systems is made. Last, but not least, the numerical method that will be used in the remainder of the book is introduced.

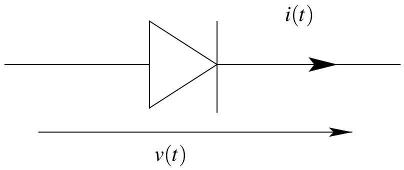
1.1.1 Diode Modeling

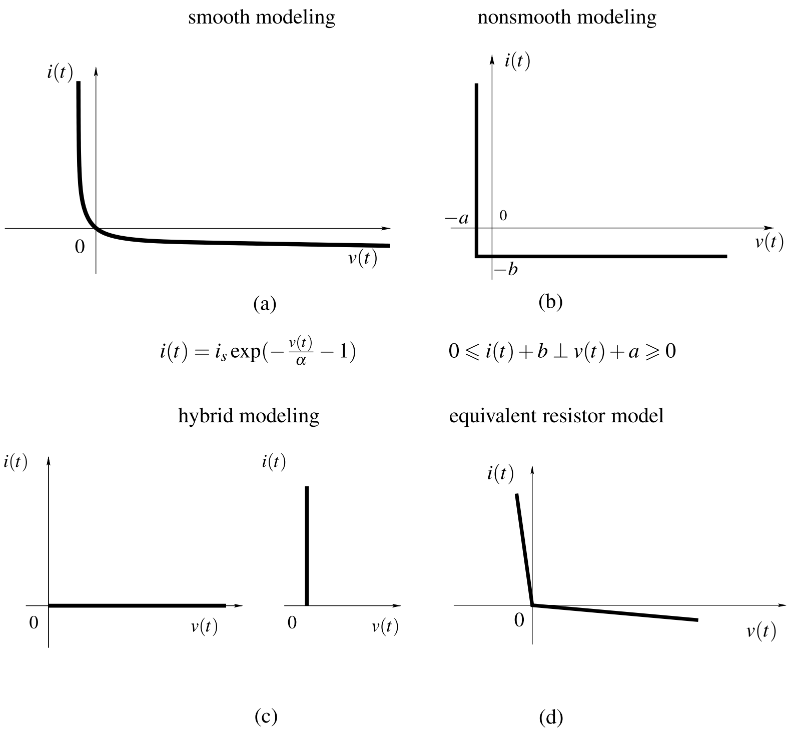
The diode represents a basic electronic device and its modeling is a central issue. Several models of the diode in Fig. 1.1 may be used, as depicted in Fig. 1.2. Each one of these models possesses some drawbacks and some advantages.

The first model in Fig. 1.2(a) is the Shockley law, that describes accurately the diode behaviour. Its drawback is that it introduces stiffness in the dynamics, hence it may considerably slow down the simulations. The model of Fig. 1.2(c) separately considers the two modes of the diode with conditional "if" and "then" statements. In a circuit the number of modes grows exponentially with the number of ideal components such as ideal diodes. Such a "hybrid" modeling approach then quickly becomes untractable and yields simulation times which are not acceptable (see Chap. 8 for numerical results on the buck converter, using a hybrid simulator). The model in

V. Acary et al., *Nonsmooth Modeling and Simulation for Switched Circuits,*
Lecture Notes in Electrical Engineering 69,
DOI 10.1007/978-90-481-9681-4_1, © Springer Science+Business Media B.V. 2011

Fig. 1.1 Diode symbol





(a)

$$i(t) = i_s \exp(-\frac{v(t)}{\alpha} - 1)$$

(b)

$$0 \leqslant i(t) + b \perp v(t) + a \geqslant 0$$

(c)

$$\begin{array}{rl} \mathsf{off} = & \mathsf{s} < 0 \\ v(t) = & \textbf{if} \;\; \mathsf{off} \;\; \textbf{then} \;\; -\mathsf{s} \;\; \textbf{else} \;\; 0 \\ i(t) = & \textbf{if} \;\; \mathsf{off} \;\; \textbf{then} \;\; 0 \;\; \textbf{else} \;\; \mathsf{s} \end{array}$$

(d)

$$v(t) = \begin{cases} -R_{\mathsf{on}}\, i(t) \text{ if } v(t) < 0 \\ -R_{\mathsf{off}}\, i(t) \text{ if } v(t) \geqslant 0 \\ R_{\mathsf{on}} \ll 1 \quad R_{\mathsf{off}} \gg 1 \end{cases}$$

Fig. 1.2 Four models of diodes

Fig. 1.2(d) is a piecewise-linear approximation of the ideal diode of Fig. 1.2(b). Its drawback is also that it introduces stiffness in the dynamics.

The model that will be mostly chosen in the present work, is the ideal diode model of Fig. 1.2(b) with $a > 0$ and $b > 0$. This nonsmooth model possesses the advantage that it keeps the main physical features of the diode (the "on"–"off" property with possible residual current and voltage), and allows one to avoid stiffness issues so that the solvers are well-conditioned. This model is not only nonsmooth, but it is *multivalued* since $i(t)$ may take any value in $[-b, +\infty)$ when $v(t) = -a$. Also,

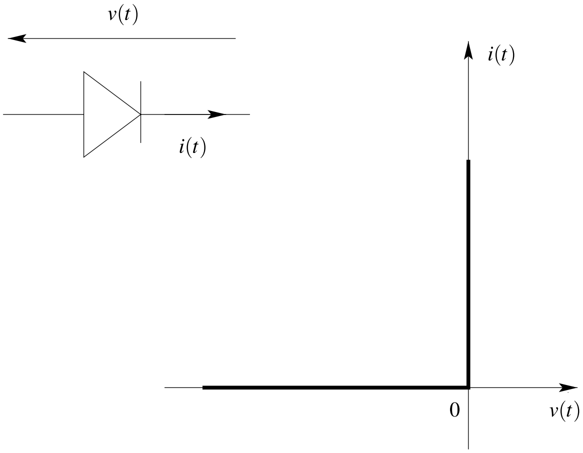


Fig. 1.3 The ideal diode voltage/current law

contrary to the model of Fig. 1.2(c) it is not purely logical (boolean), it is an analog model. This analog model is represented by so-called *complementarity relations*

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

whose meaning is that

$$\begin{array}{ll} \text{if } i(t) + b > 0 & \text{then } v(t) + a = 0, \\ \text{and if } v(t) + a > 0 & \text{then } i(t) + b = 0. \end{array} \tag{1.2}$$

When inserted into a dynamical circuit with resistors, capacitors, inductors, this will lead to a specific class of nonsmooth dynamical systems.

These arguments will be illustrated in this book through concrete simulation experiments and comparisons between various software packages.

Remark 1.1 One may also adopt the convention of Fig. 1.3 to define the voltage/current law of the diode in the case of vanishing residual current and voltage. This does not influence the results, however.

1.1.2 An RCD Circuit

Let us consider the circuit of Fig. 1.4, that is composed of a resistor R, a voltage source $u(t)$, an ideal diode, and a capacitor C mounted in series. The current through the circuit is denoted as $x(\cdot)$, and the charge of the capacitor is denoted as

$$z(t) = \int_0^t x(s)ds.$$

The dynamical equations are:

$$\begin{cases} \dot{z}(t) = -\frac{u(t)}{R} - \frac{1}{RC}z(t) + \frac{1}{R}v(t), \\ 0 \leqslant v(t) \perp w(t) = \frac{u(t)}{R} - \frac{1}{RC}z(t) + \frac{1}{R}v(t) \geqslant 0 \end{cases} \tag{1.3}$$

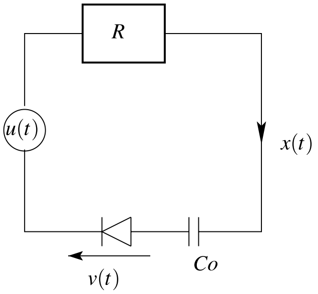


Fig. 1.4 A circuit with an ideal diode, a resistor, a capacitor and a voltage source

for all $t \in \mathbb{R}^+$. It is noteworthy that the complementarity conditions in (1.3) do not involve any unilateral constraint on the state $z(t)$. Therefore one may expect that a solution of (1.3) starting at $z(0) \in \mathbb{R}$ will be a continuous function of time. More precisely, at each time t one can solve the linear complementarity problem

$$0 \leqslant v(t) \perp \frac{u(t)}{R} - \frac{1}{RC} z(t) + \frac{1}{R} v(t) \geqslant 0,$$

whose solution depends on the sign of $\frac{u(t)}{R} - \frac{1}{RC} z(t)$. The solution $v(t)$ is unique and is given by

$$v(t) = R \max\left[0, -\frac{u(t)}{R} + \frac{1}{RC} z(t)\right].$$

Inserting it in the first line of (1.3) yields:

$$\dot{z}(t) = -\frac{u(t)}{R} - \frac{1}{RC} z(t) + \max\left[0, -\frac{u(t)}{R} + \frac{1}{RC} z(t)\right]. \tag{1.4}$$

The $\max(\cdot)$ function is not differentiable (it has a corner at 0) but it is continuous (actually, since $\max[0, x]$ is the projection of x on the convex cone $\mathbb{R}^+$ it is a Lipschitz continuous function of x). We deduce that the right-hand-side of (1.4) is Lipschitz continuous in z and u. As such (1.4) can be recast into an Ordinary Differential Equation (ODE) with Lipschitz vector field. Provided $u(t)$ possesses some basic regularity properties, this system has continuously differentiable solutions and uniqueness of the solutions holds.

The circuit of Fig. 1.4 has a single-valued dynamics despite the diode defines a multivalued voltage/current law.

Let $[0, T]$, $T > 0$ be the integration interval, and $t_0 = 0 < t_1 < \cdots < t_{n-1} < t_n = T$, with $t_{i+1} - t_i = h > 0$ the time step, $n = \frac{T}{h}$. For a continuous function $f(\cdot)$ the value f_k denotes $f(t_k)$. The backward Euler scheme used to discretize (1.3) is given by:

$$\begin{cases} z_{k+1} = z_k - h\frac{u_{k+1}}{R} - \frac{h}{RC} z_{k+1} + \frac{h}{R} v_{k+1}, \\ 0 \leqslant v_{k+1} \perp w_{k+1} = \frac{u_{k+1}}{R} - \frac{1}{RC} z_{k+1} + \frac{1}{R} v_{k+1} \geqslant 0. \end{cases} \tag{1.5}$$

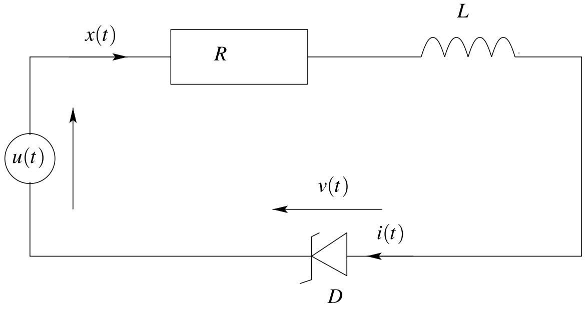


Fig. 1.5 A circuit with an ideal Zener diode, a resistor, an inductor and a voltage source

After some manipulations one finds:

$$0 \leqslant v_{k+1} \perp w_{k+1} = \left(1+\frac{h}{RC}\right)^{-1}\left[-h\frac{u_{k+1}}{R}+z_k\right]+\frac{1+h}{R}v_{k+1} \geqslant 0 \quad (1.6)$$

which is a Linear Complementarity Problem (LCP) one has to solve at each step to advance the algorithm from step k to step $k+1$. Plugging the obtained v_{k+1} into the first line of (1.5) allows one to calculate z_{k+1}. From a numerical point of view there is consequently no particular difficulty with this system, however one has to keep in mind that it is nonsmooth and this will affect greatly the possible order of the standard integration methods for ODEs (see Acary and Brogliato 2008 for details).

1.1.3 An RLZD Circuit

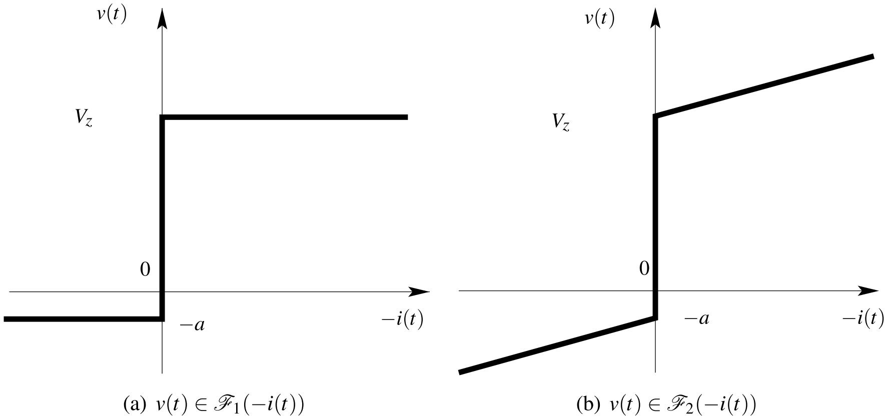
We now consider the circuit of Fig. 1.5 that contains a Zener diode. Two current/voltage laws of Zener diodes are depicted in Fig. 1.6. The model of Fig. 1.6(a) is an ideal Zener diode with some residual voltage $a > 0$. The second one may be seen as a kind of regularization of the ideal model. Indeed looking at the voltage/current law that is obtained by inverting the graphs of Fig. 1.6, one sees that the characteristic of the ideal model possesses two vertical branches, which are replaced by two linear branches with slope $\gg 1$ in the piecewise-linear model of Fig. 1.6(b). Both models are multivalued at $i(t) = 0$.

For the time being we shall denote the current/voltage laws corresponding to both models as $v(t) \in \mathscr{F}_1(-i(t))$ and $v(t) \in \mathscr{F}_2(-i(t))$ respectively. Applying the Kirchhoff's Voltage Laws (KVL) in the circuit we obtain the dynamical equations, where $x(t) = i(t)$:

$$\begin{cases} \dot{x}(t) = -\frac{R}{L}x(t) + \frac{u(t)}{L} + \frac{v(t)}{L}, \\ v(t) \in \mathscr{F}_i(-x(t)), \quad i = 1 \text{ or } i = 2. \end{cases} \quad (1.7)$$

It is noteworthy that the multifunctions $\mathscr{F}_i(\cdot)$ are quite similar to the sign multifunction

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



(a) $v(t) \in \mathscr{F}_1(-i(t))$ (b) $v(t) \in \mathscr{F}_2(-i(t))$

Fig. 1.6 Two models of Zener diode

Indeed one has

$$\mathscr{F}_1(x) = \begin{cases} \{V_z\} & \text{if } x > 0, \\ \{-a\} & \text{if } x < 0, \\ [-a, V_z] & \text{if } x = 0, \end{cases}$$

and

$$\mathscr{F}_2(x) = \begin{cases} \{\alpha x + V_z\} & \text{if } x > 0, \\ \{\beta x - a\} & \text{if } x < 0, \\ [-a, V_z] & \text{if } x = 0, \end{cases}$$

for some $\alpha > 0$, $\beta > 0$.[1] The multivalued part of the graph $(v(t), -i(t))$ may be necessary to merely guarantee the existence of a static equilibrium for (1.7). Let x^* denote the fixed point of (1.7). Then

$$-Rx^* + u(t) \in \mathscr{F}_i(x^*). \tag{1.8}$$

If $u(t) \in [-a, V_z]$, x^* necessarily belongs to the multivalued part of the graph in Fig. 1.6. Thus we get that the equilibrium inclusion in (1.8) implies that $x^* = 0$ and $u(t) \in \mathscr{F}_i(0) = [-a, V_z]$. The Zener diode multivaluedness at $i = 0$ allows a time-varying $u(t)$ while the system stays at its equilibrium $i^* = 0$. This is equivalent, from a mathematical point of view, to mechanical systems with Coulomb's friction where contact keeps sticking for tangential forces that stay inside the friction cone.

The implicit Euler method for (1.7) writes as:

$$\begin{cases} x_{k+1} = x_k - h\frac{R}{L}x_{k+1} + h\frac{u_{k+1}}{L} + \frac{h}{L}v_{k+1}, \\ v_{k+1} \in \mathscr{F}_i(-x_{k+1}), \quad i = 1 \text{ or } i = 2. \end{cases} \tag{1.9}$$

At each step one therefore has to solve the generalized equation with unknown x_{k+1}:

$$0 \in x_{k+1} - \left(1 + h\frac{R}{L}\right)^{-1}\left[x_k + h\frac{u_{k+1}}{L} + \frac{h}{L}\mathscr{F}_i(-x_{k+1})\right]. \tag{1.10}$$

[1] In the sequel we shall often neglect the brackets $\{a\}$ to denote the singletons.

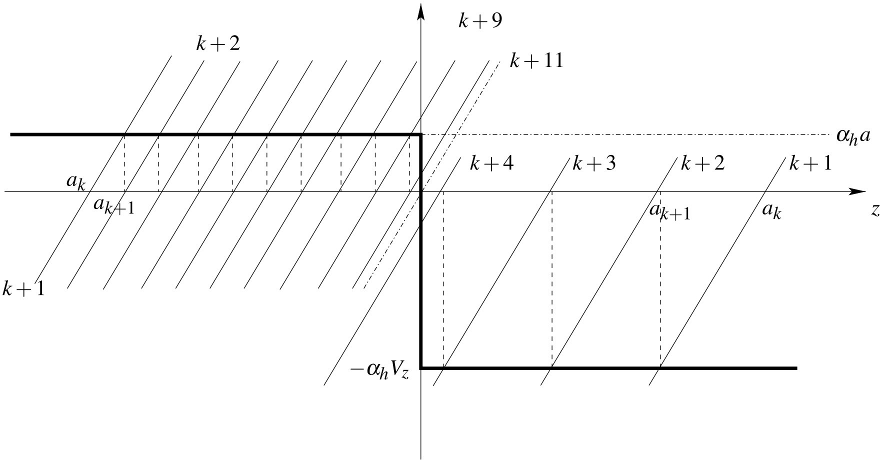


Fig. 1.7 Iterations of the backward Euler method for (1.7)

The first problem is to determine whether this has a unique solution whatever the data at step k. The answer is yes, as can be checked graphically. Indeed solving (1.10) boils down to calculating the intersection between two graphs: the graph of the single-valued mapping $z \mapsto z - a_k$ with

$$a_k = \left(1 + h\frac{R}{L}\right)^{-1}\left[x_k + h\frac{u_{k+1}}{L}\right],$$

and the graph of the multivalued mapping $z \mapsto -\alpha_h \mathscr{F}_i(-z)$ with

$$\alpha_h = \left(1 + h\frac{R}{L}\right)^{-1}\frac{h}{L}.$$

Few iterations are depicted in Fig. 1.7 for the case where $u(t) \equiv 0$, when $a_k < 0$ and $a_k > 0$. In both cases one sees that after a finite number of steps the solution is calculated to be zero. This occurs at step $k+11$ for the first case and at step $k+5$ for the second case. One says that a sliding motion (or a sliding mode) has occurred in the system, where the solution converges in finite-time towards a "switching surface" and then stays on it.

From a more general mathematical point of view, the mathematical argument that is behind the existence and the uniqueness of the intersection, is the *maximal monotonicity* of the multifunctions $z \mapsto \mathscr{F}_i(-z)$. Roughly speaking, in the plane monotonicity means that the graph of the multivalued mapping never decreases, and maximality means that the gap at the discontinuity at $z = 0$ is completely filled-in. This is the case for the graphs in Fig. 1.6.

1.1.4 An RCZD Circuit

Let us consider the circuit of Fig. 1.5 where the inductor is replaced by a capacitor C. Let us denote

$$x(t) = \int_0^t i(s)ds,$$

the charge of the capacitor. We shall this time adopt a different convention for the voltage/current law of the Zener diode, as shown in Fig. 1.8. The dynamical equations are:

$$\begin{cases} \dot{x}(t) = \frac{-1}{RC}x(t) + \frac{1}{R}(u(t) - v(t)), \\ v(t) \in \mathscr{F}_z(i(t)). \end{cases} \tag{1.11}$$

In view of the state variable definition the second line of (1.11) rewrites as $v(t) \in \mathscr{F}_z(\dot{x}(t))$, *i.e.*:

$$0 \in v(t) - \mathscr{F}_z\left(\frac{-x(t)}{RC} + \frac{u(t)}{R} - \frac{v(t)}{R}\right). \tag{1.12}$$

Letting

$$y \triangleq \frac{-x(t)}{RC} + \frac{u(t)}{R} - \frac{v(t)}{R}, \quad \text{and} \quad B \triangleq \frac{-x(t)}{RC} + \frac{u(t)}{R},$$

we can rewrite (1.12) as:

$$-\frac{y}{R} + \frac{B}{R} \in \mathscr{F}_z(y) \tag{1.13}$$

that is a generalized equation with unknown $v(t)$. Solving (1.13) boils down to finding the intersection of two graphs as depicted in Fig. 1.9. Clearly the solution is unique for any value of B if $R > 0$ (the intersection is depicted in three cases in Fig. 1.9 at the points **a**, **b** and **c**, where **b** is in the multivalued part of the graph of the mapping $\mathscr{F}_z$). Once again this is due to the maximal monotonicity of the multivalued mapping $\mathscr{F}_z(\cdot)$. Calculations yield $\frac{x(t)}{C} - u(t) > V_z \Rightarrow v^* = -V_z$ for **c**, $\frac{x(t)}{C} - u(t) < 0 \Rightarrow v^* = 0$ for **a**, and $0 < \frac{x(t)}{C} - u(t) < V_z \Rightarrow v^* \in [-V_z, 0]$ for **b**.

The implicit discretization of (1.11) is given by:

$$\begin{cases} x_{k+1} = x_k - \frac{h}{RC}x_{k+1} + \frac{h}{R}(u_{k+1} - v_{k+1}), \\ v_{k+1} \in \mathscr{F}_z(i_{k+1}), \end{cases} \tag{1.14}$$

where

$$i_{k+1} = \frac{x_{k+1} - x_k}{h}.$$

To advance the algorithm one has to solve the generalized equation:

$$0 \in v_{k+1} - \mathscr{F}_z\left(-\frac{1}{RC}x_{k+1} + \frac{1}{R}u_{k+1} - \frac{1}{R}v_{k+1}\right). \tag{1.15}$$

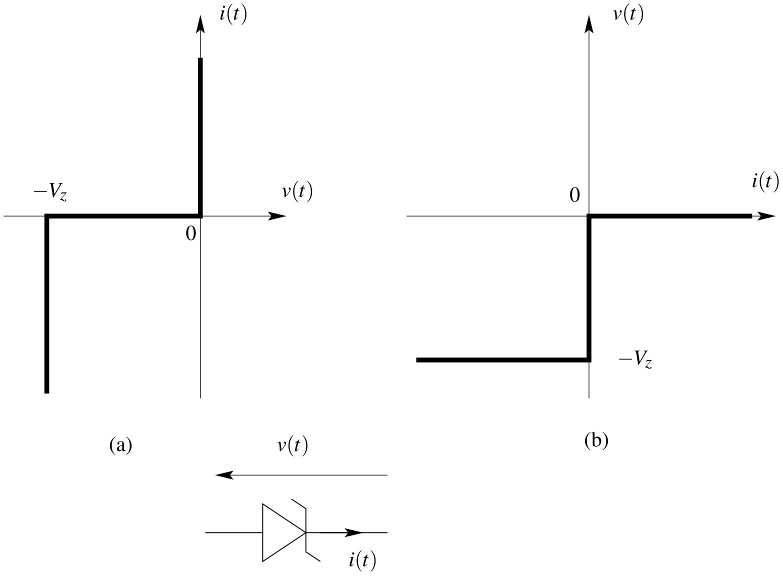


Fig. 1.8 Zener diode voltage/current law

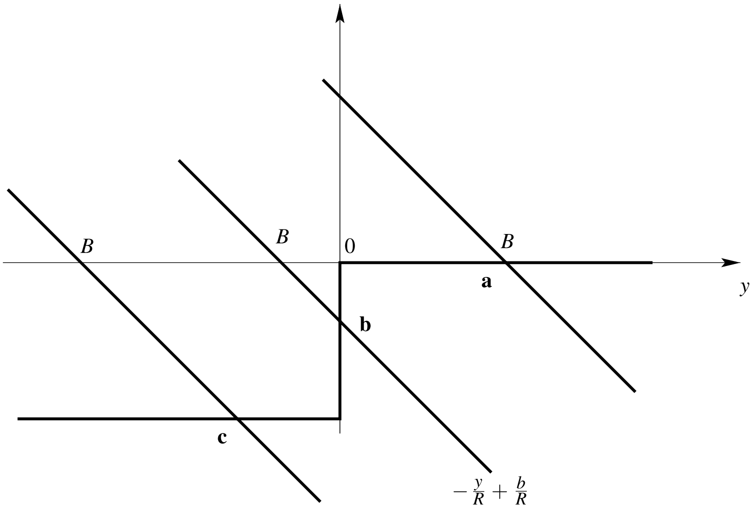


Fig. 1.9 Solving the generalized equation (1.12)

One computes that

$$x_{k+1} = -\frac{1}{RC}\left(1+\frac{h}{RC}\right)^{-1}\left(x_k+\frac{h}{R}u_{k+1}\right)+\frac{h}{R^2C}\left(1+\frac{h}{RC}\right)^{-1}\upsilon_{k+1}.$$

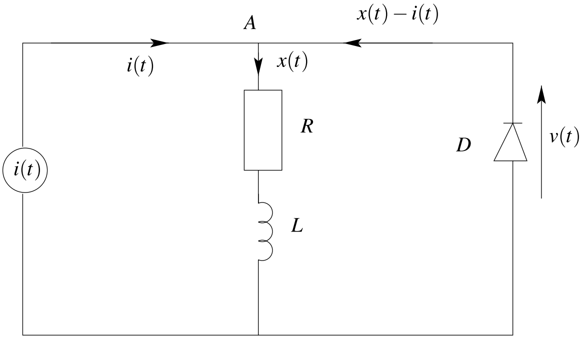


Fig. 1.10 A circuit with an ideal diode, a resistor, an inductor and a current source

Inserting this value into (1.15) one gets a generalized equation which has a unique solution v_{k+1} for similar reasons as above, provided h is small enough so that

$$\frac{1}{R} - \frac{h}{R^2C}\left(1 + \frac{h}{RC}\right)^{-1} > 0.$$

> The problems in (1.6), (1.10), (1.15) and (1.18) are the Onestep NonSmooth Problem (OSNSP) to be solved at each step of the Euler scheme. When the number of variables grows they have to be solved numerically. The design of good solvers for OSNSP is a central topic of research.

We do not present numerical results on these systems that may exhibit sliding modes. The reader is referred to Part III and especially Chaps. 7 and 8 where many simulation results are presented.

1.1.5 An RLD Circuit

Let us consider the circuit of Fig. 1.10, that is composed of an ideal diode with zero residual current and voltage, mounted in parallel with an inductor/resistor (L/R) and a current source $i(t)$. The current through the inductor/resistor is denoted as $x(t)$. The application of Kirchhoff's law for the current at node A and for the voltage in the resistor/inductor/diode loop, yields the following dynamics:

$$\begin{cases} \dot{x}(t) = -\frac{R}{L}x(t) + v(t), \\ 0 \leqslant w(t) = x(t) - i(t) \perp v(t) \geqslant 0 \end{cases} \tag{1.16}$$

for all $t \geqslant 0$, where $\frac{v(t)}{L}$ is the voltage across the diode. We shall see in Sect. 2.5.5 another way to write the dynamics in (1.16), using some basic convex analysis, and which is at the roots of Moreau's time-stepping method.

For the time being, let us propose a time-discretization of (1.16). A backward Euler algorithm for (1.16) is:

$$\begin{cases} x_{k+1} = x_k - h\frac{R}{L}x_k + hv_{k+1}, \\ 0 \leqslant w_{k+1} = x_{k+1} - i_{k+1} \perp v_{k+1} \geqslant 0. \end{cases} \tag{1.17}$$

Inserting the value of x_{k+1} in the complementarity condition $0 \leqslant x_{k+1} - i_{k+1} \perp v_{k+1} \geqslant 0$ one obtains

$$0 \leqslant \left(1 - h\frac{R}{L}\right)x_k - i_{k+1} + hv_{k+1} \perp v_{k+1} \geqslant 0, \tag{1.18}$$

which is an LCP with unknown v_{k+1}. Since $h > 0$ it may be deduced by simple inspection that such a problem always possesses a unique solution.

Notice now from the complementarity condition of (1.16) that the state $x(t)$ is unilaterally constrained as $x(t) \geqslant i(t)$ for all $t \geqslant 0$. Suppose that at some time $t \geqslant 0$ this constraint is violated (it may for instance be that initially $x(0) < i(0)$). A jump in $x(\cdot)$ is necessary at t to continue the integration on the right of t (otherwise, this dynamical system will possess a solution on $[0, t)$, and not on $[0, +\infty)$). In passing this shows that the dynamical system in (1.16) should contain a third ingredient, in a similar way as mechanical systems with impacts: a state reinitialization law. We shall come back to this issue in Sect. 2.4.3.2 in a more general setting.

Let us assume now that at step k one has $x_k - i_k = -\delta$ for some $\delta > 0$. If h is small enough and the signal $i(t)$ is continuous, then $-h\frac{R}{L}x_k + x_k - i_{k+1}$ is negative as well (actually, we could directly suppose that $x_k - i_{k+1} < 0$). Therefore the solution of the linear complementarity problem is $v_{k+1} = \frac{1}{h}(h\frac{R}{L}x_k - x_k + i_{k+1}) > 0$. Inserting this value in the first line of (1.17) one obtains:

$$x_{k+1} = i_{k+1}. \tag{1.19}$$

So at step $k+1$ one has $x_{k+1} - i_{k+1} = 0$ while $v_{k+1} > 0$: the complementarity condition is satisfied. The state has jumped and the jump value is $x_{k+1} - x_k = i_{k+1} - i_k + \delta$, that is close to δ if the time-step is very small and $i(t)$ does not vary too much on $[t_k, t_{k+1}]$. For instance, if $i(t)$ is null (no current source) the backward Euler scheme computes $x_{k+1} = 0$ while $x_k < 0$.

The basic implicit Euler method automatically computes state jumps for inconsistent states.

This suggests that the backward Euler scheme in (1.17) approximates the dynamics in (1.16) with an additional ingredient:

$$x(t^+) = i(t^+) + \max[0, x(t^-) - i(t^+)]. \tag{1.20}$$

The notation $f(t^+)$ generically denotes $\lim_{\tau \to t, \tau > t} f(\tau)$ for a function $f(\cdot)$ which possesses a discontinuity of the first kind at t and a right-limit at t (same for the left-limit).

Since the state $x(\cdot)$ may jump, the derivative of $x(\cdot)$ has to be understood in the distributional sense at a jump time t, *i.e.* a Dirac measure $(x(t^+) - x(t^-))\delta_t$. From a rigorous mathematical point of view, the dynamics in (1.16) has to be rewritten as an *equality of measures*. If the state jumps at time t, a continuous variable $v(t)$ is not sufficient to correctly model the dynamics. A measure di should be considered such that

$$di = v(t)dt + \sigma\delta_t \tag{1.21}$$

with

$$\begin{aligned}\sigma\delta_t &= (x(t^+) - x(t^-))\delta_t\\ &= \{i(t^+) - x(t^-) + \max[0, x(t^-) - i(t^+)]\}\delta_t.\end{aligned} \tag{1.22}$$

The dynamics (1.16) written as a measure differential equation is

$$dx = -\frac{R}{L}x(t)dt + di, \tag{1.23}$$

which can be decomposed into a smooth dynamics:

$$\dot{x}(t) = -\frac{R}{L}x(t) + v(t), \tag{1.24}$$

almost everywhere, and a jump dynamics at time t:

$$x(t^+) - x(t^-) = \sigma. \tag{1.25}$$

Physically the voltage across the diode approximates such a Dirac measure and the current $x(t)$ varies abruptly. Notice that if $x(t^-) - i(t^+) < 0$ then σ has a magnitude $|\sigma| = i(t^+) - x(t^-) > 0$, and if $x(t^-) - i(t^+) > 0$ then σ has a magnitude $|\sigma| = 0$. In the second case there is logically no state jump. The measure di is a non negative measure for all $t \geqslant 0$, and it can be checked that the complementarity condition is respected at the jump times, provided it is written with right-limits:

$$0 \leqslant x(t^+) - i(t^+) \perp di \geqslant 0.$$

Clearly (1.20) may be rewritten as

$$x(t^+) - i(t^+) = \mathrm{proj}[\mathbb{R}^+; x(t^-) - i(t^+)], \tag{1.26}$$

where $\mathrm{proj}[K; x]$ denotes the projection of the vector x on the convex set K, that is equivalent to:

$$x(t^+) = i(t^+) + \mathrm{argmin}_{z\in\mathbb{R}^+}\frac{1}{2}[z - x(t^-) + i(t^+)]^2. \tag{1.27}$$

As we shall see later this will be generalized to more complex circuits (see Sect. 2.4.3.2), and looks like an inelastic impact law in mechanics.

State inequality constraints (also named unilateral constraints) induce state jumps in circuits. These jumps should obey some physical rule.

Remark 1.2 There is a slight difference between the scheme in (1.17) and the one in (1.5), which is fully implicit. However changing the first line of (1.17) to $x_{k+1} = x_k - h\frac{R}{L}x_{k+1} + hv_{k+1}$ does not influence much the presented results.

In the numerical practice one prefers to use the slack variable $\sigma_k \approx di((t_k, t_{k+1}])$ which corresponds to an impulse over the time interval. This yields the following scheme:

$$\begin{cases} x_{k+1} = x_k - h\frac{R}{L}x_k + \sigma_{k+1}, \\ 0 \leqslant w_{k+1} = x_{k+1} - i_{k+1} \perp \sigma_{k+1} \geqslant 0. \end{cases} \tag{1.28}$$

Inserting the value of x_{k+1} into the complementarity condition, one obtains:

$$0 \leqslant \left(1 - h\frac{R}{L}\right)x_k - i_{k+1} + \sigma_{k+1} \perp \sigma_{k+1} \geqslant 0. \tag{1.29}$$

This choice has to major advantages:

1. The value σ_{k+1} is homogeneous to an impulse. It remains a finite value when a jump occurs on a time-interval when h vanishes. One does not want to try to approximate unbounded quantities like a Dirac measure. In fact, it is simply *impossible* to numerically approximate a Dirac measure! However it is possible to approximate the measure of a bounded interval of time by a Dirac measure, because such a quantity is bounded. This is precisely what is done when one computes $\sigma_k = di((t_k, t_{k+1}])$: this quantity is to be thought of as the measure of an interval of integration $[t_k, t_{k+1})$ by the Dirac measure di.
2. This permits to obtain an LCP whose "matrix" is equal to 1, not to h. The LCP is better conditioned. In higher dimensional systems this will be translated into matrices which do not tend to become singular as h vanishes.

Such arguments are more clearly explained within the framework of measure differential inclusions, see Sect. 2.5.5 in Chap. 2.

When state jumps occur, the mathematical formulation of the dynamics has to be adapted. A Measure Differential Equation (MDE) has to be written. The numerical time-integration should also be adapted taking into account that a numerical pointwise evaluation of measures is a nonsense.

There is a second issue to be solved, besides possible state jumps (*electrical impacts*): what happens when the trajectory evolves on the boundary of the admissible domain $\{x \in \mathbb{R} \mid x - i(t) \geqslant 0\}$, *i.e.* when $x(t) = i(t)$ on some time interval $[a, b)$, $a < b$? Usually a non zero v will be necessary to keep the state on the boundary, otherwise the unilateral constraint would be violated on the right of $t = a$, because the dynamics $\dot{x} = -\frac{R}{L}x(t)$ implies that $x(t)$ decreases.

In order to calculate the suitable value for $v(t)$ on $[a, b)$, let us remark that since $x(t) - i(t) = 0$ for all $t \in [a, b)$ then one must have $\dot{x}(t) - \frac{di}{dt}(t) \geqslant 0$ on $[a, b)$. Moreover the complementarity condition has to be satisfied as well between the right-derivative $\dot{x}(t^+)$ and v, for if $\dot{x}(t^+) - \frac{di}{dt}(t) > 0$ at some $t \in [a,b)$ then $x(t) > i(t)$ in a right-neighborhood of t (the only assumption needed here is that $\dot{x}(\cdot)$ and $\frac{di}{dt}(\cdot)$ be right-continuous functions). Therefore one can write $0 \leqslant \dot{x}(t^+) - \frac{di}{dt}(t) \perp v(t) \geqslant 0$ on $[a,b)$, which yields:

$$0 \leqslant -\frac{R}{L}x(t) + v(t^+) - \frac{di}{dt}(t) \perp v(t^+) \geqslant 0. \tag{1.30}$$

This is an LCP with unknown $v(t^+)$. In fact since we have assumed that $x(t) = i(t)$ on $[a,b)$ we are solely interested in studying what happens on the right of $t = b$. However (1.30) is true on the whole of $[a,b)$. If $-\frac{R}{L}x(t) - \frac{di}{dt} < 0$ then $v(t^+) = \frac{R}{L}x(t) + \frac{di}{dt} > 0$ on $[a,b)$: the voltage $v(t)$ keeps the trajectory on the boundary of the admissible domain, in a way quite similar to the contact force in mechanics.

> The voltage $v(\cdot)$ across the diode plays the role of a Lagrange multiplier associated with the constraint $w(t) = x(t) - i(t) \geqslant 0$.

Since we suppose that $x(\cdot)$ and $i(\cdot)$ are continuous functions, we may rewrite (1.30) at $t = b$ as:

$$0 \leqslant -\frac{R}{L}i(b) + v(b^+) - \frac{di}{dt}(b^+) \perp v(b^+) \geqslant 0. \tag{1.31}$$

One sees that all depends on the values of $i(t)$ and its derivative at $t = b$. If $-\frac{R}{L}i(b) - \frac{di}{dt}(b^+) > 0$ then $v(b^+) = 0$, while if $-\frac{R}{L}i(b) - \frac{di}{dt}(b^+) \leqslant 0$ then $v(b^+) = \frac{R}{L}i(b) + \frac{di}{dt}(b^+) \geqslant 0$. In the first case one obtains after insertion of $v(t^+)$ in (1.16):

$$\dot{x}(b^+) - \frac{di}{dt}(b^+) = -2\left(\frac{R}{L}i(b) + \frac{di}{dt}(b^+)\right) > 0, \tag{1.32}$$

from which one deduces that the trajectory detaches from the constraint boundary, in other words $x(t) > i(t)$ in a right neighborhood of $t = b$. In the second case one gets:

$$\dot{x}(b^+) - \frac{di}{dt}(b^+) = 0, \tag{1.33}$$

while $v(b^+) \geqslant 0$, so that the trajectory stays on the boundary on the right of $t = b$. In both cases the complementarity conditions in (1.16) are verified. Obviously once the trajectory no longer lies on the boundary of the admissible domain, then (1.30) is no longer valid.

From a numerical point of view, this situation is taken into account by the time-stepping scheme in (1.17) and also by the time-stepping scheme (1.28) at the impulse level.

Let us recapitulate the above developments. The circuit in Fig. 1.10 with an ideal diode, involves unilateral constraints on the state and complementarity conditions between two *slack* variables $w(t)$ and $v(t)$. These features, which are closely linked to the modeling approach, imply some peculiarities of the dynamics: state jumps may occur, and the trajectory may evolve on the boundary of the admissible domain. In the first case, the dynamics involves a measure di which contains positive Dirac measures, in the second case the variable v is a positive function.

Example 1.3 (State jumps simulation) Let us illustrate the above on the following example ($R = L = 1$ in (1.16)).

$$\begin{cases} \dot{x}(t) = -x(t) + v(t), \\ 0 \leqslant v(t) \perp w(t) = x(t) - i(t) \geqslant 0, \\ x(t^+) = i(t^+) + \max[0, x(t^-) - i(t^+)] \quad \text{at jump times}, \end{cases} \tag{1.34}$$

where

$$i(t) = \begin{cases} 0 & \text{for all } t \in [0, 5), \\ 2 & \text{for all } t \in (5, 10), \\ -2 & \text{for all } t \geqslant 10. \end{cases}$$

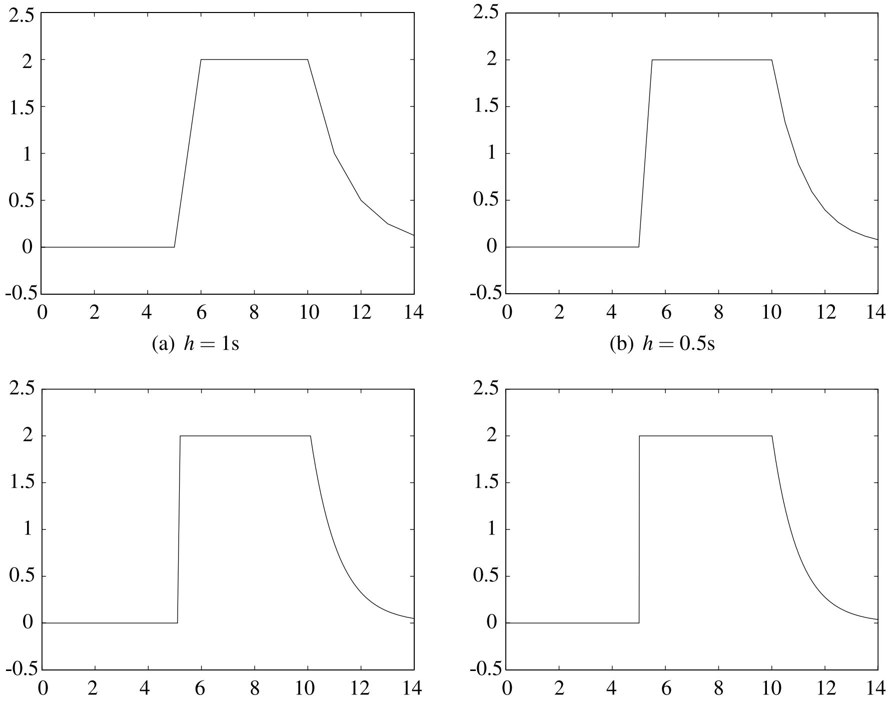
Let $x(0^-) = -2$. The analytical solution of (1.34) with this value for the current source is:

$$x(t) = \begin{cases} x(0^+) = 0, \; x(t) = 0, \; v(t) = 0 & \text{on } t \in (0, 5), \\ x(5^+) = 2, \; x(t) = 2, \; v(t) = 2 & \text{on } t \in [5, 10), \\ x(10^+) = 2, \; x(t) = 2e^{10-t}, \; v(t) = 0 & \text{on } t \in [10, +\infty). \end{cases} \tag{1.35}$$

Hence $x(\cdot)$ jumps initially and at $t = 5$, and on $5 < t \leqslant 10$ the voltage $v(t)$ keeps the solution on the boundary $x(t) = i(t)$. Mathematically speaking, the dynamics in (1.34) at $t = 0$ has to be written as an equality of measures: $(x(0^+) - x(0^-))\delta_0 = 2\delta_0 = \sigma\delta_0$, where the diode voltage is a Dirac measure at $t = 0$. The same applies at $t = 5$. Physically a peak of voltage at the inductor ports implies a peak of voltage across the diode, by Kirschhoff's law which holds at state jumps.

Various solutions are simulated with the above implicit Euler method and are depicted in Figs. 1.11 and 1.12. The numerical method (1.28) computes the values of σ_{k+1} which is the impulse of the Dirac over $(t_k, t_{k+1}]$, and one may recover the values of v_{k+1} by dividing by h. This is what is plotted in Fig. 1.12, where one sees that the smaller h, the higher and more narrow peak. It is nevertheless crucial to keep in mind at this stage that a numerical method cannot, strictly speaking, guarantee any kind of convergence of a discrete signal towards a Dirac measure: this is simply meaningless. Only the convergence of its impulse has a rigorous meaning.

At $t = 10$ there is no state jump, however the multiplier $v(t)$ possesses a jump that comes from the complementarity conditions. Figure 1.13 depicts the approximation of the impulse of $v(\cdot)$ on intervals $(t_k, t_{k+1}]$, *i.e.* the value σ_{k+1}. It is apparent from Fig. 1.13 that the discrete-time signal σ_{k+1} converges at $t = 5$ as $k \to +\infty$ to the analytical value of the Dirac measure magnitude, equal to 2. Outside $t = 5$ the



(a) $h = 1\text{s}$

(b) $h = 0.5\text{s}$

(c) $h = 0.1\text{s}$

(d) $h = 10^{-2}\text{s}$

Fig. 1.11 Solutions of (1.34): the state x_k vs. time t_k

impulse σ_{k+1} tends to zero as expected. Since $v(t)$ is a function of bounded variations outside $t = 5$, we have

$$\sigma_{k+1} \approx \int_{(t_k,t_{k+1}]} v(t)dt, \tag{1.36}$$

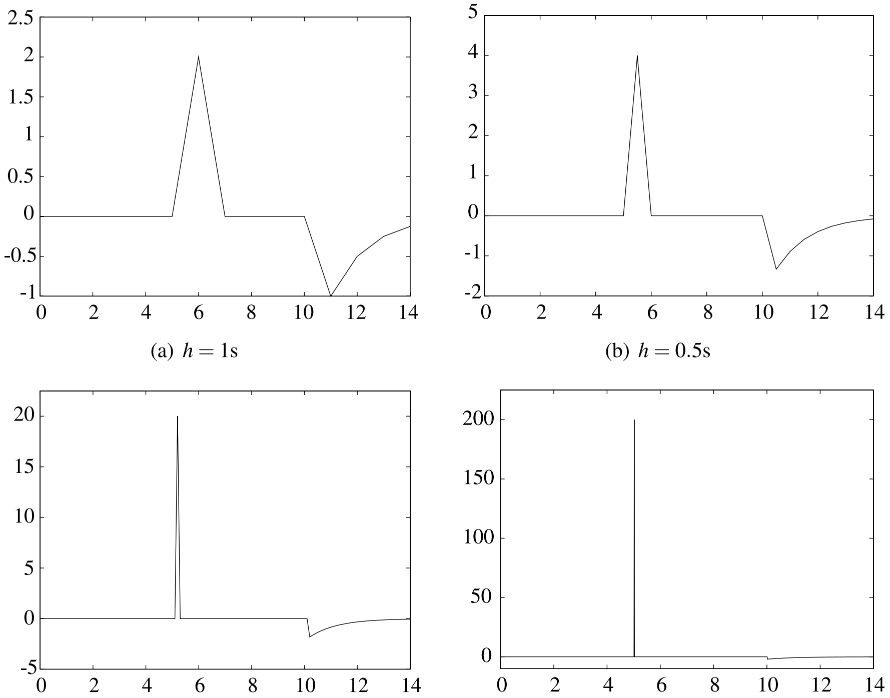
and

$$\lim_{|t_{k+1}-t_k|\to 0} \int_{(t_k,t_{k+1}]} v(t)dt = 0. \tag{1.37}$$

Remark 1.4 The fact that we write the interval as $(t_k, t_{k+1}]$ may be explained by the rules for differential measures, see Sect. A.5.

These simulations have been done using the automatic circuit equation formulation module described in Chap. 6 and the software package SICONOS.

Remark 1.5 The implicit Euler method approximates well the current reinitialization in (1.20) (equivalently in (1.26) or (1.27)). In practice one does not use such fully implicit algorithms, but so-called θ-methods. One has to be careful with the discretization with a θ-method because a wrong discretization of the Lagrange multiplier $v(t)$ may yield absurd results (see Acary and Brogliato 2008, Sect. 1.1.6.2).



(a) $h = 1$s

(b) $h = 0.5$s

(c) $h = 0.1$s

(d) $h = 10^{-2}$s

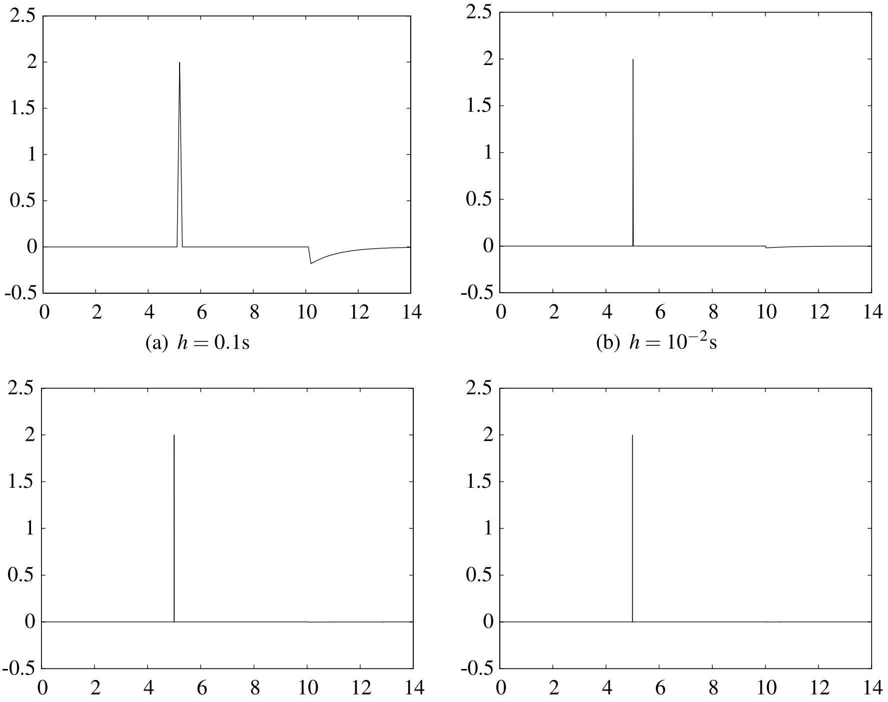
Fig. 1.12 Solutions of (1.34): the multiplier v_k vs. time t_k

1.1.6 More Examples: Order-Two and Order-Three Circuits

The circuits of the foregoing section have a state vector of dimension 1. Circuits with dimension 2 and dimension 3 state vector are presented now. First let us consider the circuits of Figs. 1.14 and 1.15, and let $x_1(t)$ be the charge of the capacitors and $x_2(t)$ be the current through the inductors. The variable $v(t)$ may represent either a current or a voltage, depending on the circuit. Their dynamical equations are summarized as follows:

$$(\mathbf{a}) \quad \begin{cases} \dot{x}_1(t) = x_2(t) - \frac{1}{RC}x_1(t) - \frac{v(t)}{R}, \\ \dot{x}_2(t) = -\frac{1}{LC}x_1(t) - \frac{v(t)}{L}, \\ 0 \leqslant v(t) \perp w(t) = \frac{v(t)}{R} + \frac{1}{RC}x_1(t) - x_2(t) \geqslant 0, \end{cases} \tag{1.38}$$

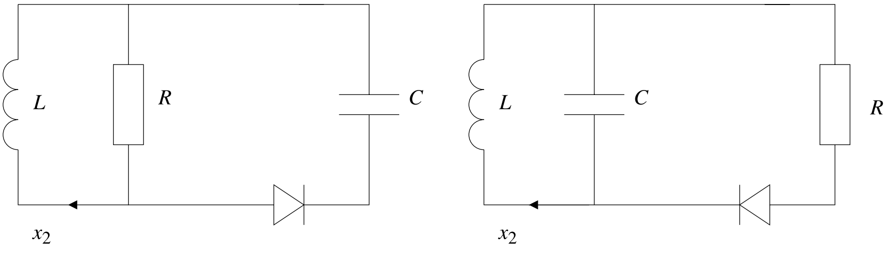
$$(\mathbf{b}) \quad \begin{cases} \dot{x}_1(t) = -x_2(t) + v(t), \\ \dot{x}_2(t) = \frac{1}{LC}x_1(t), \\ 0 \leqslant v(t) \perp w(t) = \frac{1}{C}x_1(t) + Rv(t) \geqslant 0. \end{cases} \tag{1.39}$$



(a) $h = 0.1$s (b) $h = 10^{-2}$s

(c) $h = 10^{-3}$s (d) $h = 10^{-4}$s

Fig. 1.13 Solutions of (1.34): the impulse σ_k vs. time t_k



(a) (b)

Fig. 1.14 RLC circuits with an ideal diode

Since $R > 0$, the circuits of Fig. 1.14 have a single-valued dynamics despite the diode defines a multivalued voltage/current law.

One sees that in each case (1.38) and (1.39) the complementarity relations define an LCP with a unique solution whatever the values of $x_1(t)$ and $x_2(t)$, because

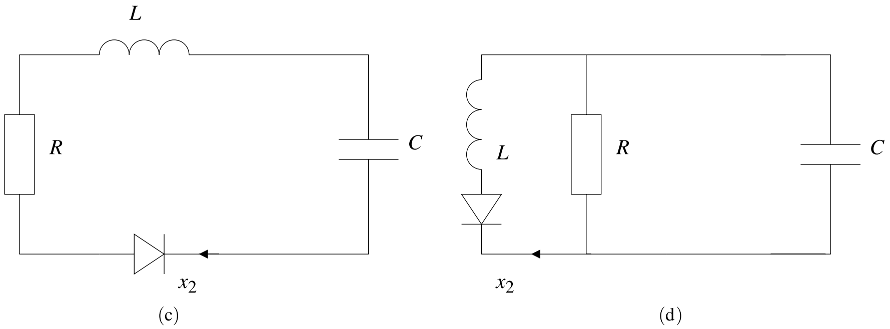


Fig. 1.15 RLC circuits with an ideal diode

the LCP matrices are equal to $\frac{1}{R}$ and R respectively. A similar manipulation as in Sect. 1.1.2 may be done for these two circuits.

$$\textbf{(c)} \quad \begin{cases} \dot{x}_1(t) = x_2(t), \\ \dot{x}_2(t) = -\frac{R}{L}x_2(t) - \frac{1}{LC}x_1(t) - \frac{v(t)}{L}, \\ 0 \leqslant v(t) \perp w(t) = -x_2(t) \geqslant 0, \end{cases} \tag{1.40}$$

$$\textbf{(d)} \quad \begin{cases} \dot{x}_1(t) = x_2(t) - \frac{1}{RC}x_1(t), \\ \dot{x}_2(t) = -\frac{1}{LC}x_1(t) - \frac{v(t)}{L}, \\ 0 \leqslant v(t) \perp w(t) = -x_2(t) \geqslant 0. \end{cases} \tag{1.41}$$

One may replace the ideal diode by a Zener diode, in which case the dynamical structure is the same, except that the complementarity conditions are replaced by $v(t) \in \mathscr{F}_z(w(t))$, where $\mathscr{F}_z(\cdot)$ is a multivalued mapping. More generally, one may introduce any nonsmooth electronic device with piecewise-linear voltage/current law into the dynamics, keeping the same system's structure. More will be said on the circuit (1.41) in Sect. 2.5.10.

Let us now consider the circuit in Fig. 1.16 which contains a Zener and an ideal diodes, with the conventions of Figs. 1.8 and 1.3. Its dynamics is given by:

$$\begin{cases} \dot{x}_1(t) = x_2(t), \\ \dot{x}_2(t) = -\frac{1}{L_1C}x_1(t) - \frac{R_1+R_3}{L_1}x_2(t) + \frac{R_1}{L_1}x_3(t) - \frac{1}{L_1}v_1(t) - \frac{1}{L_1}v_2(t), \\ \dot{x}_3(t) = \frac{R_1}{L_2}x_2(t) - \frac{R_1+R_2}{L_2}x_3(t) + \frac{1}{L_2}v_1(t) - \frac{1}{L_2}u(t), \\ v_1(t) \in \mathscr{F}_z(x_2(t) - x_3(t)), \\ 0 \leqslant v_2(t) \perp x_2(t) \geqslant 0, \end{cases} \tag{1.42}$$

where $u(t)$ is a voltage source, $x_1(t)$ is the charge of the capacitor, *i.e.* $x_1(t) = \int_0^t x_2(s)ds$, $x_2(t)$ is the current through the diode, $x_3(t)$ is the current through the inductor L_2, $v_1(t)$ is the Zener diode voltage and $v_2(t)$ is the ideal diode voltage. This is an order three dynamical system, where the multivalued part is due to the Zener and the ideal diodes characteristics.

Remark 1.6 When analysing the circuit in (1.16) we have seen that state jumps are necessary to enable one to integrate the dynamics over $[0, +\infty)$. The jumps obey

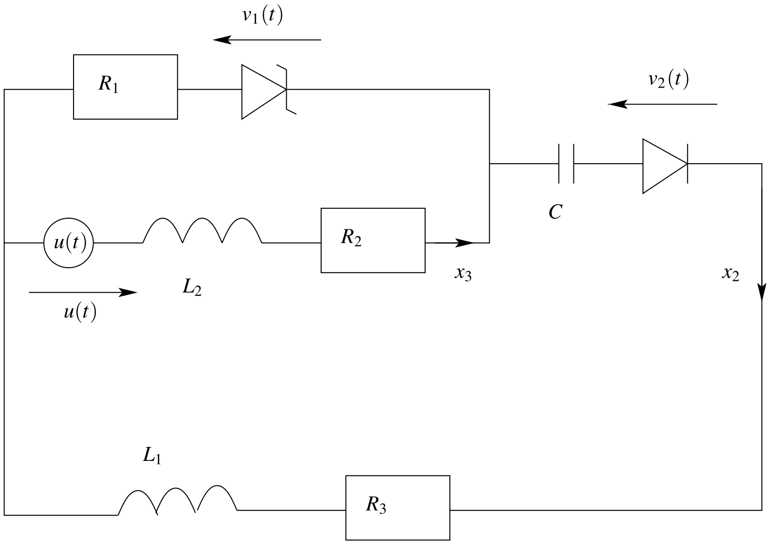


Fig. 1.16 A circuit with Zener and ideal diodes

the rule in (1.20). A similar behaviour can be observed on the circuits (1.40), (1.41) and (1.42). Indeed the nonsmooth constraints in these circuits involve unilateral constraints on the state. If the initial data in (1.40) and (1.41) is such that $x_2(0^-) > 0$ (respectively < 0 in (1.42)) then an initial jump is needed to bring $x_2(0^+)$ to a non positive value (respectively non negative in (1.42)). See Sect. 2.4.3.2 for details on state jump rules for electrical circuits (*electrical impacts*).

The implicit Euler discretization of (1.42) proceeds exactly as in the above cases. Let us write (1.42) compactly as

$$\begin{cases} \dot{x}(t) = Ax(t) - Bv(t) + Eu(t), \\ v(t) \in \mathscr{F}(w(t)), \end{cases} \tag{1.43}$$

where

$$v(t) = (v_1(t) \quad v_2(t))^T, \qquad w(t) = \begin{pmatrix} x_2(t) - x_3(t) \\ x_2(t) \end{pmatrix} = Cx(t), \tag{1.44}$$

and the matrices A, B, C, E are easily identified.[2] One obtains:

$$\begin{cases} x_{k+1} = (I_3 - hA)^{-1}[x_k - hBv_{k+1} + hEu_{k+1}], \\ w_{k+1} = Cx_{k+1}, \\ v_{k+1} \in \mathscr{F}(w_{k+1}), \end{cases} \tag{1.45}$$

[2] See Chap. 2 and (2.23) for details on how to obtain the multivalued mapping $\mathscr{F}(\cdot)$ from the nonsmooth part of (1.42).

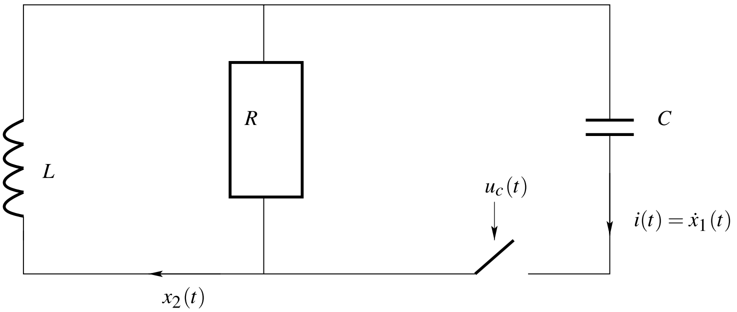


Fig. 1.17 A circuit with an ideal switch

where I_3 is the 3×3 identity matrix. Similarly to (1.6), (1.10), (1.15) and (1.18), the problem in (1.45) is a Onestep NonSmooth Problem (OSNSP) to be solved at each step. A fundamental property of the OSNSP is the uniqueness of solutions x_{k+1} at step k. For the time-being we shall admit that the OSNSP (1.45) has a unique solution for any data at step k. The proof will require some convex analysis and is therefore postponed to Sect. 2.5.12 of Chap. 2.

1.1.7 A Circuit with an Ideal Switch

The circuits that are introduced in the foregoing sections, all include some nonsmooth electronic device that implies some kind of switches in the dynamics. Not all switches are equivalent, however. This section is dedicated to present a novel kind of switch that is widely spread in electronics.

Let us now consider the circuit of Fig. 1.17, that is composed of an inductor, a resistor, a capacitor and a switch whose voltage/current law is given by:

$$u(t) = \begin{cases} R_{\mathsf{on}} i(t) & \text{if } u_c(t) > 0, \\ R_{\mathsf{off}} i(t) & \text{if } u_c(t) < 0, \end{cases} \tag{1.46}$$

where $i(t)$ is the current through the capacitor, and $u_c(t)$ is the voltage signal that triggers the switch. The ideal switch corresponds to $R_{\mathsf{on}} = 0$ and $R_{\mathsf{off}} = +\infty$. Let $x_1(t) = \int_0^t i(s)ds$ be the charge of the capacitor, and $x_2(t)$ be the current through the inductor. It is noteworthy that the switch model in (1.46) is a nonsmooth model. Indeed the switch induces a jump in the variable $u(t)$ when $u_c(t)$ changes its sign at time t, and the jump is equal to $|(R_{\mathsf{on}} - R_{\mathsf{off}})i(t)|$. In the framework of this book, we shall naturally embed the switching law in (1.46) into a multivalued model of the form $u(t) \in \mathscr{F}_s(x_1(t), x_2(t), u_c(t))$. It will be seen below why this is in fact a necessary step. The dynamical equations are:

$$\begin{cases} \dot{x}_1(t) = \frac{-1}{RC} x_1(t) + x_2(t) - \frac{u(t)}{R}, \\ \dot{x}_2(t) = \frac{-1}{LC} x_1(t) - \frac{u(t)}{L}, \\ u(t) \in \mathscr{F}_s(i(t), u_c(t)). \end{cases} \tag{1.47}$$

The multifunction $\mathscr{F}_s(i(t), u_c(t))$ may be represented by a set of complementarity relations given in (4.40). Inserting (1.46) into (1.47) we obtain the following *piecewise-linear* dynamical system:

$$
\begin{array}{ll}
(\Sigma_{\mathrm{on}}) \begin{cases} \dot{x}_1(t) = \frac{-1}{(R+R_{\mathrm{on}})RC}x_1(t) + \frac{R}{R+R_{\mathrm{on}}}x_2(t) \\ \dot{x}_2(t) = -(\frac{1}{LC} + \frac{R_{\mathrm{on}}}{(R+R_{\mathrm{on}})LC})x_1(t) - \frac{R_{\mathrm{on}}R}{L(R+R_{\mathrm{on}})}x_2(t) \end{cases} & \text{if } u_c(t) > 0, \\
(\Sigma_{\mathrm{off}}) \begin{cases} \dot{x}_1(t) = \frac{-1}{(R+R_{\mathrm{off}})RC}x_1(t) + \frac{R}{R+R_{\mathrm{off}}}x_2(t) \\ \dot{x}_2(t) = -(\frac{1}{LC} + \frac{R_{\mathrm{off}}}{(R+R_{\mathrm{off}})LC})x_1(t) + \frac{R_{\mathrm{off}}R}{L(R+R_{\mathrm{off}})}x_2(t) \end{cases} & \text{if } u_c(t) < 0.
\end{array}
\tag{1.48}
$$

For the time-being the switching condition is of the exogenous type since $u_c(\cdot)$ is a function of time. However $u_c(\cdot)$ may be equal to a function of the state (x_1, x_2) in which case the switches are state-dependent. From the point of view of the mathematical analysis of the dynamical system, this is quite important because state-dependent switches are more difficult. In view of (1.48) the multivalued function $\mathscr{F}_s(\cdot)$ is given by:

$$
u(t) \in \begin{cases} R_{\mathrm{on}}(\frac{-1}{(R+R_{\mathrm{on}})RC}x_1(t) + \frac{R}{R+R_{\mathrm{on}}}x_2(t)) & \text{if } u_c(t) > 0, \\ [u_m(t), u_M(t)] & \text{if } u_c(t) = 0, \\ R_{\mathrm{off}}(\frac{-1}{(R+R_{\mathrm{off}})RC}x_1(t) + \frac{R}{R+R_{\mathrm{off}}}x_2(t)) & \text{if } u_c(t) < 0, \end{cases}
\tag{1.49}
$$

with $u_m = \min[a, b]$ and $u_M = \max[a, b]$, where

$$
a = R_{\mathrm{on}}\left(\frac{-1}{(R+R_{\mathrm{on}})RC}x_1(t) + \frac{R}{R+R_{\mathrm{on}}}x_2(t)\right),
$$

and

$$
b = R_{\mathrm{off}}\left(\frac{-1}{(R+R_{\mathrm{off}})RC}x_1(t) + \frac{R}{R+R_{\mathrm{off}}}x_2(t)\right).
$$

In the numerical practice one often chooses $R_{\mathrm{on}} \ll 1$ and $R_{\mathrm{off}} \gg 1$. The limit case of an ideal switch is when $R_{\mathrm{on}} = 0$ and $R_{\mathrm{off}} = +\infty$. Then the dynamics in (1.48) becomes:

$$
\begin{array}{ll}
(\Sigma_{\mathrm{on}}) \begin{cases} \dot{x}_1(t) = \frac{-1}{RC}x_1(t) + x_2(t) \\ \dot{x}_2(t) = \frac{-1}{LC}x_1(t) \end{cases} & \text{if } u_c(t) > 0, \\
(\Sigma_{\mathrm{off}}) \begin{cases} \dot{x}_1(t) = 0 \\ \dot{x}_2(t) = -(\frac{1}{LC} + \frac{1}{RC})x_1(t) + x_2(t) \end{cases} & \text{if } u_c(t) < 0.
\end{array}
\tag{1.50}
$$

The value of the state variable $x_1(\cdot)$ in (Σ_{off}) is given by its value just before the switch between (Σ_{on}) and (Σ_{off}) occurs. It appears clearly from both (1.48) and (1.50) that there is a jump in the vector field of this circuit. We are therefore facing a dynamical system whose dynamics belongs to discontinuous piecewise-linear systems, written as:

$$
\dot{x}(t) = \begin{cases} A_1x(t) & \text{if } u_c(t) > 0, \\ A_2x(t) & \text{if } u_c(t) < 0, \end{cases}
\tag{1.51}
$$

or, if some state feedback is introduced in the switching conditions:

$$\dot{x}(t) = \begin{cases} A_1 x(t) & \text{if } x(t) \in \chi_1, \\ A_2 x(t) & \text{if } x(t) \in \chi_2, \end{cases} \tag{1.52}$$

where χ_1 and χ_2 are disjoint subsets of $\mathbb{R}^2$ which may be assumed to cover $\mathbb{R}^2$, *i.e.* $\chi_1 \cup \chi_2 = \mathbb{R}^2$. The matrices A_1 and A_2 are easily identifiable. Precisely, one has to know how the system is defined on the boundary between χ_1 and χ_2, since the right-hand-side of the system is discontinuous on this boundary. This is especially crucial if the boundary defines an attractive surface: the two vector fields point outside their respective domains of application when the state attains the boundary surface S_{12} between χ_1 and χ_2. One has to define what is the dynamics *on* S_{12}. Usually one resorts to the theory of Filippov's differential inclusions, see Chap. 2, Sect. 2.4.4.

Let us come back to another type of switching circuit, for instance circuit (b) of Fig. 1.14, whose dynamics is in (1.39):

$$\begin{cases} \dot{x}_1(t) = -x_2(t) + v(t), \\ \dot{x}_2(t) = \frac{1}{LC} x_1(t), \\ 0 \leqslant v(t) \perp w(t) = \frac{1}{C} x_1(t) + Rv(t) \geqslant 0. \end{cases} \tag{1.53}$$

The nonsmooth part of (1.53) is the LCP $0 \leqslant v(t) \perp w(t) = \frac{1}{C} x_1(t) + Rv(t) \geqslant 0$, whose solution $v(t)$ may be found by inspection:

$$v(t) = \max\left[0, \frac{-1}{RC} x_1(t)\right]. \tag{1.54}$$

The $\max(\cdot)$ function is not differentiable at zero, however it is continuous everywhere. This means that the dynamics in (1.53), which we rewrite as:

$$\begin{cases} \dot{x}_1(t) = -x_2(t) + \max[0, \frac{-1}{RC} x_1(t)], \\ \dot{x}_2(t) = \frac{1}{LC} x_1(t), \end{cases} \tag{1.55}$$

has a piecewise-linear continuous right-hand-side. The switching surface is given by $x_1 = 0$, and the dynamics is perfectly defined on it. The dynamics of the circuit (b) of Fig. 1.14 is an ordinary differential equation.

The two circuits in (1.48) and (1.53) belong to the class of piecewise-linear dynamical systems. However, a major discrepancy between them is that (1.48) has a discontinuous right-hand-side, whereas the right-hand-side of (1.53) is continuous. Mathematically, the first one will be embedded into differential inclusions, while the second one is an ordinary differential equation.

It is a general fact that one always has to be careful when a switching occurs in a dynamical system. The underlying mathematical structure of the system depends on the switching rules. The circuits of Figs. 1.14(b) and 1.17 possess different switching rules, and they are of a different mathematical nature.

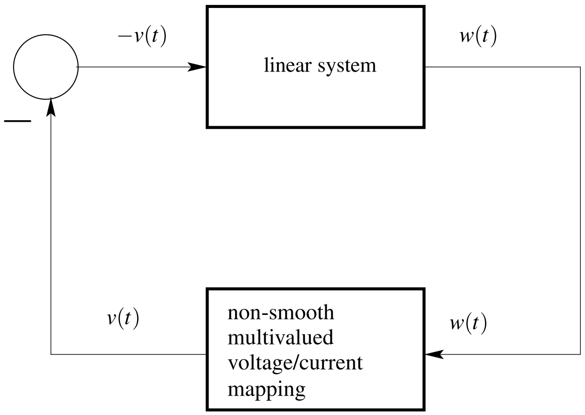


Fig. 1.18 Circuits as feedback systems

1.2 A Unified Dynamical Framework: Lur'e Dynamical Systems

All the above nonsmooth circuits with no external excitation may be recast into the general framework of Fig. 1.18, that is made of a smooth part in negative feedback with a multivalued part (called in the Systems and Control community a Lur'e system, that is a widely used formalism to study the so-called absolute stability). The same applies for the discretized systems as depicted in Fig. 1.19 (see *e.g.* (1.45)). It is noteworthy that the bouncing ball can also be interpreted this way, see Fig. 2.26. The interest of this interconnection is that if the linear part is dissipative with supply rate $\langle v(t), w(t)\rangle$ and if the multivalued part is maximal monotone, then the overall dynamical system is well-posed and stable.

Consider as an example (1.40). The linear part of the interconnection is:

$$\begin{cases} \dot{x}_1(t) = x_2(t), \\ \dot{x}_2(t) = -\frac{R}{L}x_2(t) - \frac{1}{L}x_1(t) + \frac{1}{L}(-v(t)). \end{cases} \tag{1.56}$$

The multivalued nonsmooth part is:

$$0 \leqslant v(t) \perp w(t) = -x_2(t) \geqslant 0. \tag{1.57}$$

For the discretized case see for instance (1.17). See also Remark 2.79.

The feedback branch contains a static multivalued operator $v(t) \in \mathscr{F}(w(t))$ where $w(t)$ may take various values depending on the circuit. For instance in (1.7) one has $w(t) = -x(t)$, and in (1.11) one has $w(t) = \frac{-x(t)}{RC} + \frac{u(t)}{R} - \frac{v(t)}{R}$. The multivalued mapping $w(t) \mapsto v(t)$ for the systems in (1.16) and (1.3) is given by the complementarity conditions $0 \leqslant v(t) \perp w(t) \geqslant 0$. As pointed out above this defines a multivalued mapping since $v(t)$ may take any non negative value when $w(t) = 0$. We shall see in Chap. 2 that the complementarity mapping can be expressed as an inclusion as well for some multivalued mapping $\mathscr{F}(\cdot)$. It follows from these simple examples that the variable $w(t)$ possesses the generic form $w(t) = Cx(t) + Dv(t) + Fu(t)$ for suitable matrices C, D, F. The linear system block is easily identified in the examples, and takes the form $\dot{x}(t) = Ax(t) + Bv(t) + Eu(t)$ for suitable matrices A, B, E. It is noteworthy that this dynamical framework, that is

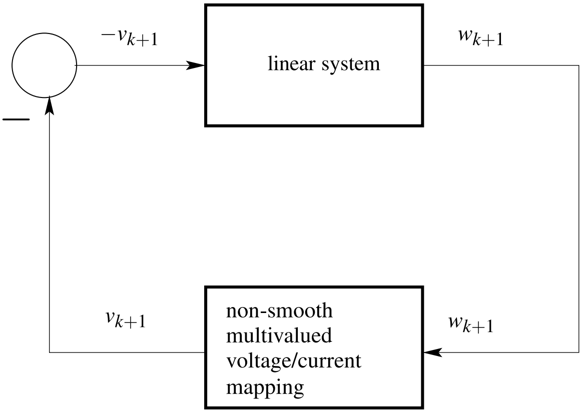


Fig. 1.19 Discretized circuits as feedback systems

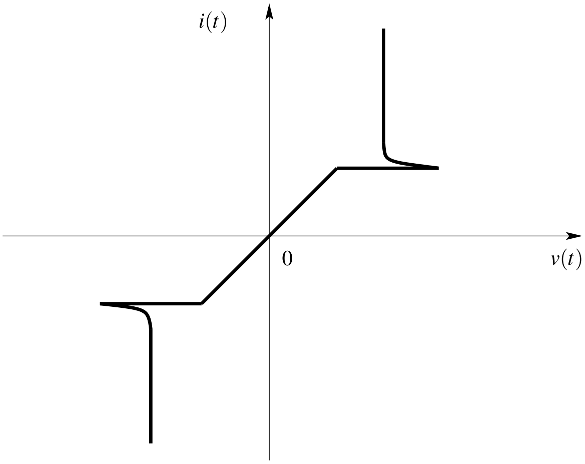


Fig. 1.20 A piecewise-nonlinear DIAC model (Diode for Alternative Current)

part of what we shall call later the nonsmooth dynamics systems (NSDS) approach, allows for nonlinear electronic devices (for instance, nonlinear resistors, capacitors and inductors, or nonlinear nonsmooth devices like the DIAC in Fig. 1.20).

The discrete-time system is advanced from step k to step $k+1$ by solving a OSNSP that is given by the feedback branch in Fig. 1.19.

1.3 An Aside on Nonsmooth Mechanics: The Bouncing Ball

The analogies between mechanics and electricity are well-known. We will not in this book provide an exhaustive presentation of all possible analogies between electrical circuits and mechanical systems on the one hand, and between electronic devices and mechanical components on the other hand. It is however interesting to point out some discrepancies and some analogies between circuits and mechanical systems

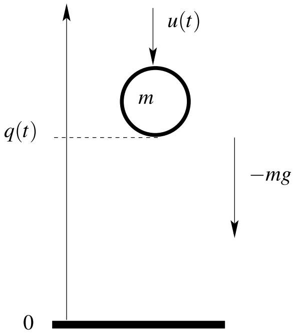


Fig. 1.21 The bouncing-ball

through a simple example: the bouncing ball. The dynamics of the bouncing ball that is represented in Fig. 1.21 may be written as follows:

$$\begin{cases} m\ddot{q}(t) + mg + u(t) = \lambda(t), \\ 0 \leqslant \lambda(t) \perp w(t) = q(t) \geqslant 0, \\ \dot{q}(t^+) = -e\dot{q}(t^-) \quad \text{if } q(t) = 0 \text{ and } \dot{q}(t^-) \leqslant 0. \end{cases} \tag{1.58}$$

The first two lines represent the dynamics outside the impact times, *i.e.* either when the ball is not in contact with the ground $q(t) > 0$, or when it is in persistent contact with the ground, *i.e.* $q(t) = 0$ on some interval $[t_1, t_2]$, $t_2 > t_1$. The third line is an impact law, which reinitializes the ball's velocity according to a restitution rule known as Newton's restitution law. The physical parameter $e \in [0, 1]$ is the restitution coefficient. The reader is referred to Acary and Brogliato (2008) for more details on the bouncing ball's dynamics. The dynamics in (1.58) is quite similar to the dynamics in (1.40) and (1.41), or in (1.34). There is, however, an important discrepancy. In the electrical circuits, the state jumps have two origins: either a "bad" initial value which does not respect the unilateral constraint, or a discontinuous exogenous excitation (like $i(t)$) in (1.34). In the bouncing ball the velocity jumps originate from a more intrinsic reason: if the conditions of the third line of (1.58) are satisfied at some t, then there is no function $\lambda(t)$ which may keep the set $\{q \in \mathbb{R} \mid q \geqslant 0\}$ invariant. A Dirac measure that produces a velocity sign reversal is necessary. This is in fact due to a higher relative degree between $w(\cdot)$ and $\lambda(\cdot)$. In the case of the above circuits the relative degree is equal to one, but for (1.58) it is equal to 2, since one needs to differentiate $w(\cdot)$ twice to recover λ.[3] See also Remark 2.82 for an explanation on the fundamental difference between the bouncing ball and circuits with a relative degree equal to one.

On a time interval $[t_1, t_2]$, $t_2 > t_1$ where $q(t) = 0$, one may compute $\lambda(t)$ using a reasoning similar to what we did for the circuit in (1.16). The fact that $q(t) = \dot{q}(t) = 0$ on $[t_1, t_2]$ and $0 \leqslant \lambda(t) \perp w(t) = q(t) \geqslant 0$ implies, under the right-continuity of the acceleration $\ddot{q}(\cdot)$, that $0 \leqslant \lambda(t) \perp \ddot{q}(t) \geqslant 0$. Indeed if $\ddot{q}(t) < 0$

[3] In a system $\dot{x}(t) = Ax(t) + B\lambda(t)$, $w(t) = Cx(t) + D\lambda(t)$, with $\lambda(t) \in \mathbb{R}$, $w(t) \in \mathbb{R}$, the relative degree r is the integer $\geqslant 0$ such that $r = 0$ if $D \neq 0$, $r \geqslant 1$ if $D = 0$ and $CA^{i-1}B = 0$ for all $i < r$, while $CA^{r-1}B \neq 0$.

for some $t \in [t_1, t_2]$ then $\dot{q}(\cdot)$ and $q(\cdot)$ both become negative on the right of t (see Proposition 7.1.1 in Glocker 2001), which is not permitted. On the other hand if $\ddot{q}(t) > 0$ then they both become positive on the right of t, so that the orthogonality has to be satisfied between the acceleration and the contact force. Then replacing the acceleration by its value one obtains

$$0 \leqslant \lambda(t) \perp \frac{1}{m}\lambda(t) - g - \frac{1}{m}u(t) \geqslant 0, \tag{1.59}$$

which is an LCP with unknown $\lambda(t)$, and which possesses a unique solution whatever $u(t)$ (see Theorem 2.43 in the next chapter). This solution, when inserted back into the dynamics, determines the sign of the acceleration and whether or not the contact is kept or lost.

It will be shown in the next chapter, particularly in Sect. 2.5.13, that the bouncing ball dynamics also lends itself to an interconnection as in Fig. 1.18.

1.4 Conclusions

Several simple circuits and a simple mechanical system have been analysed in the previous sections. They can all be embedded into the class of differential inclusions

$$\dot{x}(t) \in G(x(t), t) \tag{1.60}$$

where $G : \mathbb{R}^n \times \mathbb{R}^+ \to \mathbb{R}^n$ is some multivalued mapping. The circuits of Figs. 1.10, 1.4, 1.14 and 1.15, can be embedded into the class of linear complementarity systems:

$$\begin{cases} \dot{x}(t) = Ax(t) + Bv(t) + Eu(t), \\ 0 \leqslant v(t) \perp w(t) = Cx(t) + Dv(t) + Fu(t) \geqslant 0, \\ \text{State jump law}, \end{cases} \tag{1.61}$$

where $x(t) \in \mathbb{R}^n$, $v(t) \in \mathbb{R}^m$, $u(t) \in \mathbb{R}^l$, and the matrices have suitable dimensions.[4] The reader will easily identify the matrices for each circuit. The major discrepancy between (1.16) and (1.3) is that in the former $D = 0$, while in the latter $D = \frac{1}{R} > 0$. In terms of *relative degree* between the "input" λ and the "output" w, the systems in (1.40) and (1.41) have a relative degree $r = 1$ while the systems in (1.3), (1.38) and (1.39) have a relative degree $r = 0$. The relative degree is taken here in the sense of Systems and Control Theory, *i.e.* it is equal to the number of times one needs to differentiate the "output" so that the "input" appears explicitly. The bouncing ball system has a relative degree $r = 2$. It is clear from these worked examples that r has a strong influence on the nature of the solutions of the circuits: roughly speaking, when $r = 0$ solutions are continuous, and for $r \geqslant 1$ solutions may contain the Dirac measure and its derivatives up to the order $r - 1$, see Acary et al. (2008).

[4] Usually one writes λ instead of v, since this slack variable has the mathematical meaning of a Lagrange multiplier.

As we shall see in Chap. 2, close links exist between complementarity conditions and inclusions into normal cones to convex sets. This means that we will be able, under certain conditions, to interpret systems like (1.61) as differential inclusions whose multivalued (or set-valued) part is a normal cone to a convex set. The advantages for doing so are that some developments made in this section, will become extremely clear in the framework of differential inclusions. However a prerequisite is some understanding of basic convex analysis. The Zener diode circuit will be embedded into another type of differential inclusions, called *Filippov's differential inclusions*.

One important feature of the modeling approach and of the associated numerical method which are presented in this chapter through simple examples, is that one does not consider, when studying switching circuits, topological changes of the circuit depending on the switches status (open or closed). There is a single state vector that remains unchanged whatever the system's configuration.

It is noteworthy that this is independent on the type of numerical method (time-stepping or event-driven algorithms), despite we shall deal only with time-stepping methods in this book.

1.5 Historical Summary

The following material does not pretend to be exhaustive. It only aims at providing the reader with some rough information on the history of nonsmooth circuit theory. Modeling electrical circuits devices with piecewise-linear, nonsmooth components started some decades ago. It may be traced back in the early seventies with some works on piecewise-linear resistive networks (Fujisawa et al. 1972). Chua and his co-workers (Kang and Chua 1978; Chua and Ying 1983; Chua and Dang 1985) introduced some canonical form representations of continuous piecewise-linear functions and applied it to resistive circuits. More recent works in this spirit may be found in Wen et al. (2005), Parodi et al. (2005), Yamamura and Machida (2008), Repetto et al. (2006), and Carbone and Palma (2006). A comparative study between various approaches is made in Kevenaar and Leenaerts (1992). These approaches do not consider multivalued characteristics, nor unilateral effects. In other words, in case of planar characteristics, there are no vertical branches. Switched-capacitor networks have been studied by various authors (Vlach et al. 1984, 1995; Huang and Liu 2009; Bedrosian and Vlach 1992; Vlach and Opal 1997; Zhu and Vlach 1995). The fact that there may exist "electrical impacts" due to inconsistent initial data, and that suitable numerical techniques should be found to cope with such issues, is pointed out in Vlach et al. (1995) and Vlach and Opal (1997). Complementarity was introduced for the modeling and the analysis of piecewise-linear circuits in Stevens and Lin (1981), van Bokhoven and Jess (1978), van Bokhoven (1981), and van Eijndhoven (1984). Later on van Stiphout (1990), Vandenberghe et al. (1989), and Leenaerts (1999) analyzed piecewise-linear circuits with the complementarity approach. See the book of Leenaerts and Van Bokhoven (1998) for a good account to all these works. In particular Leenaerts (1999) introduces the backward Euler method for linear complementarity systems. More recently one may find several studies of nonsmooth circuits (*i.e.* circuits with nonsmooth devices) that significantly improve and enlarge the scope of the previous works. The analysis (well-posedness, stability, and numerical analysis) of linear complementarity systems, which model linear circuits with ideal diodes, transistors, switches, power converters, and of relay systems, has been investigated in Camlibel et al. (2002a, 2002b), Heemels et al. (2000, 2001), Camlibel (2001), Enge and Maisser (2005), Vasca et al. (2009), and Batlle et al. (2005). In particular backward Euler time-stepping methods as introduced in Leenaerts (1999) have been studied for linear complementarity systems in Camlibel et al. (2002a) with convergence results. There is a strong analogy between such implicit Euler scheme and Moreau's catching up algorithm that was designed for sweeping processes of order one and two, see Acary and Brogliato (2008, §1.4.3.5). The first convergence proofs for Moreau's catching up algorithm are due to Monteiro Marques in the eighties and may be found in the book of Monteiro Marques (1993). See also Heemels and Brogliato (2003) and Brogliato (2003) for surveys on complementarity systems. Interesting works dealing with the analogy between nonsmooth mechanical devices and nonsmooth electrical devices may be found in Glocker (2005) and Moeller and Glocker (2007). In Glocker (2005) a numerical time-stepping scheme is proposed that is quite close to the time-stepping methods analyzed in this book. The buck DC-DC converter is modeled in Glocker (2005) as a Lagrangian system with inertia matrix the matrix of the inductances, where the state variables are the generalized charges and currents in the fundamental loops, and in Moeller and Glocker (2007) as a Lagrangian system with inertia matrix the matrix of the capacitances, where the state variables are the nodal fluxes and voltages. This shows that in the nonsmooth framework also there is a strong analogy between mechanical and electrical systems. Other works may be found in Yuan and Opal (2003), Chung and Ioinovici (1994), Opal (1996), De Kelper et al. (2002), and Liu et al. (1993), that witness the intense activity in this field. The analysis of circuits with multivalued nonsmooth devices, using variational inequalities and Moreau's superpotentials, is proposed in Addi et al. (2007, 2010), Goeleven (2008), Goeleven and Brogliato (2004), and Brogliato and Goeleven (2005). The interpretation and the analysis of nonsmooth circuits as the negative feedback interconnection of passive and multivalued monotone operators (Lur'e systems) has been done in Brogliato (2004), Brogliato et al. (2007), and Brogliato and Goeleven (2010). The extension of all previous works to the case where the solutions may contain not only Dirac measures, but derivatives of the Dirac (which are Schwarz's distributions) was considered in Acary et al. (2008), where the modeling, well-posedness and numerical analysis is presented for such distribution differential inclusions. This may find applications in electrical circuits when some feedback controllers that augment the relative degree between the complementarity variables, is applied.

The issue of numerical simulation and its corollary, *i.e.* the development of software packages, has been also a very active field during the past twenty years. It has been noticed many times that simulators of the SPICE family are not suitable for the simulation of circuits whose solutions are not smooth enough (Maffezzoni et al. 2006; Wang et al. 2009; Mayaram et al. 2000; Maksimovic et al. 2001; Valsa and Vlach 1995; Biolek and Dobes 2007; Lukl et al. 2006). This has motivated the development of many simulators for such analog nonsmooth circuits (SCISIP: Lukl et al. 2006, SWANN: Valsa and Vlach 1995, WATSCAD: Bliss et al. 1992, CPPSIM[5]), most of them being old and no longer maintained, or dedicated to a very narrow class of switched systems. Commercial software packages for the simulation of analog, switched circuits are also numerous. Among them we may cite those of the SPICE family: NGSPICE (http://ngspice.sourceforge.net/), ELDO (from Mentor Graphics), SMASH (from Dolphin), VIRTUOSO SPECTRE (from Cadence), SABER (from Synopsis). Other packages based on hybrid simulators are PLECS (from Plexim), PSIM (from Powersys). It is also worth citing the package LMGC90 that has been developed in Montpellier (France).[6] Despite LMGC90 is dedicated to the simulation of granular matter (and is therefore far from circuits applications), it is based on the NSDS approach with Moreau-Jean's time-stepping algorithm. Comparisons between the results obtained with the INRIA platform SICONOS and some of these tools are presented later in this book.

This quite short summary proves that the field of modeling, analysis and simulation of nonsmooth circuits has been and is still a very active field of research.

[5] http://www.cppsim.com/index.html.

[6] http://www.lmgc.univ-montp2.fr/~dubois/LMGC90/index.html.

Chapter 2
Mathematical Background

This chapter is devoted to present the mathematical tools which are used in this book to analyze the nonsmooth circuits and their time-discretizations. This chapter does not aim at being exhaustive. The unique objective is that the book be sufficiently self-contained and that all the mathematical notions which are the foundations of the nonsmooth dynamical systems that are presented, be easily available to the readers who are not familiar with such tools. For this reason the results are given without proofs. After a brief recall of some basic tools, we come back to the circuits of Chap. 1 and rewrite their dynamics using new mathematical frameworks. Many of the tools which are presented in this chapter, will be used, or presented in an other way in Chap. 4.

2.1 Basics from Convex and Nonsmooth Analysis

In this section one recalls some definitions and properties that are associated with convex sets and functions, their subdifferentiation, and multifunctions (or set-valued functions). Classical and introductory references are Hiriart-Urruty and Lemaréchal (2001) and Rockafellar (1970) for convex analysis, Smirnov (2002) for multivalued functions, Facchinei and Pang (2003) and Murty (1988) for variational inequalities and complementarity problems.

2.1.1 Convex Sets and Functions

2.1.1.1 Definitions and Properties

Definition 2.1 (Convex sets) A subset C of $\mathbb{R}^n$ is said convex if $(1-\lambda)x+\lambda y \in C$ whenever $x \in C$ and $y \in C$ and $\lambda \in (0,1)$.

V. Acary et al., *Nonsmooth Modeling and Simulation for Switched Circuits,*
Lecture Notes in Electrical Engineering 69,
DOI 10.1007/978-90-481-9681-4_2, © Springer Science+Business Media B.V. 2011

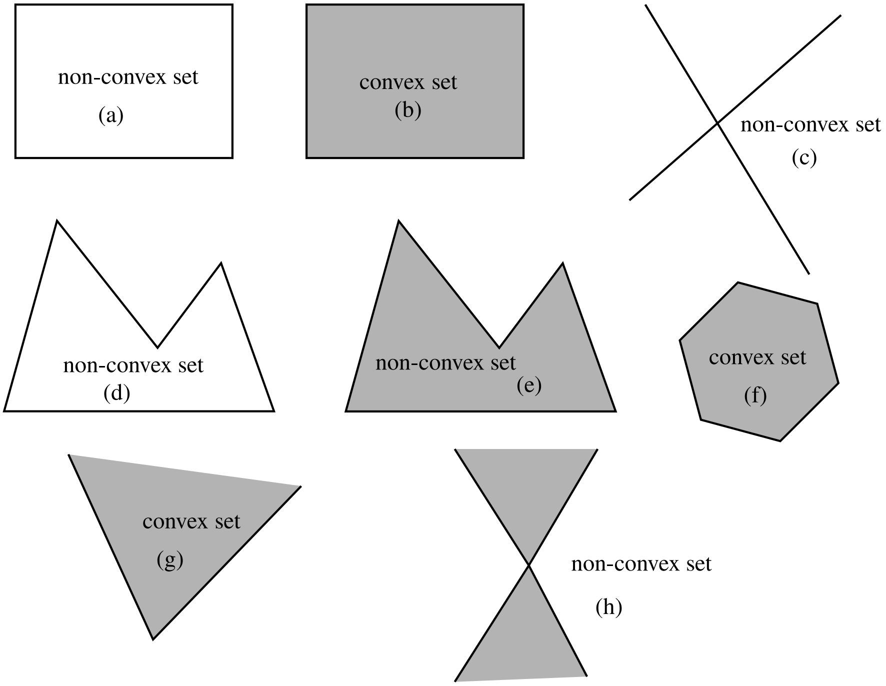


Fig. 2.1 Planar convex and non-convex sets

As a consequence C is convex if and only if it contains all the convex combinations of its elements. Examples of planar convex and non-convex sets are depicted in Fig. 2.1.

Definition 2.2 (Cones) A subset C of $\mathbb{R}^n$ is called a cone if it is closed under positive scalar multiplication, *i.e.* $\lambda x \in C$ when $x \in C$ and $\lambda > 0$.

Examples of convex and non-convex cones in three dimensions are depicted in Fig. 2.2. The sets in Fig. 2.1(c) and (h) are non-convex cones. The set in Fig. 2.1(g) is a convex cone. When a cone C is closed, then necessarily $0 \in C$. The set of solutions to $Ax \geqslant 0$ where A is a constant matrix, is a polyhedral convex cone.

Definition 2.3 (Polar cones) Let $C \subseteq \mathbb{R}^n$ be a non empty convex cone. The polar of C is the set

$$C^\circ = \{s \in \mathbb{R}^n \mid \langle s, x\rangle \leqslant 0 \text{ for all } x \in C\}. \tag{2.1}$$

Examples of cones and their polar cone are depicted in Fig. 2.3. Polarity may be seen as a generalization, in a unilateral way, of orthogonality. Hence, if C is a subspace then C° is its orthogonal subspace. The polar cone obtained from C depends on the scalar product that is used in the definition: changing the scalar

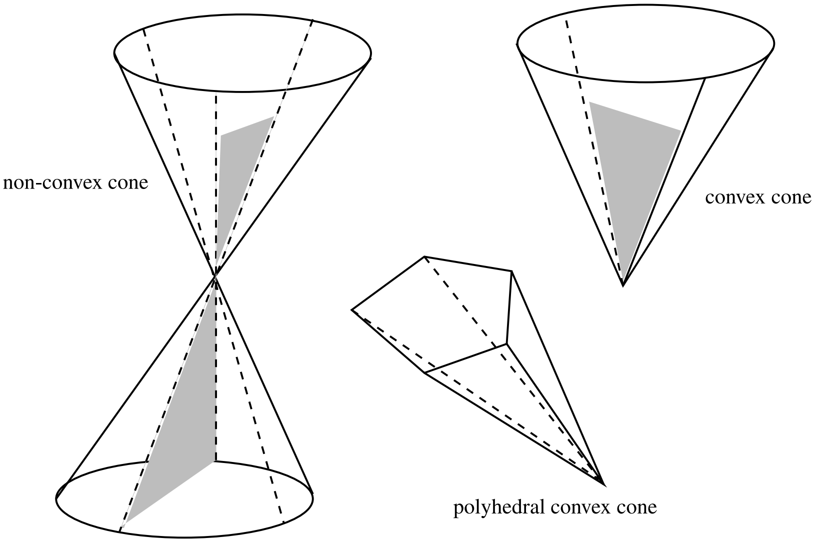
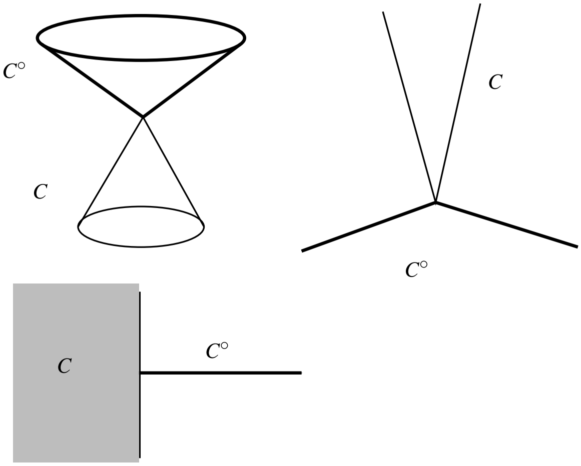


Fig. 2.2 Convex and non-convex cones

Fig. 2.3 Convex cones and their polar cones



product changes C°. When C is a non empty closed convex cone, then C° is also a non empty closed convex cone, and $C^{\circ\circ} = C$ (*i.e.* the polar of the polar is the original cone).

Many authors rather speak of *conjugate* or *dual* cones, which are defined as $C^* = \{s \in \mathbb{R}^n \mid \langle s, x \rangle \geqslant 0 \text{ for all } x \in C\}$. Therefore $C^\circ = -C^*$. The polar cone to $\mathbb{R}^n_+$ is $\mathbb{R}^n_-$, whereas its dual cone is simply itself.

Remark 2.4 Given a non empty set C, not necessarily convex, one may define also its dual cone as the set $C^* = \{s \in \mathbb{R}^n \mid s^T y \geqslant 0 \text{ for all } y \in C\}$. This is indeed a cone as can be checked.

An interesting result is the next one:

Proposition 2.5 *Let C_i, $1 \leqslant i \leqslant m$, be non empty convex cones of $\mathbb{R}^n$. Then $(\sum_{i=1}^m C_i)^\circ = C_1^\circ \cap C_2^\circ \cap \cdots \cap C_m^\circ$.*

Obviously this also holds for dual cones.

Definition 2.6 (Convex functions) Let C be a non empty convex set in $\mathbb{R}^n$. A function $f : C \to \mathbb{R}$ is said convex on C when, for all pairs $(x, y) \in C \times C$ and all $\lambda \in (0, 1)$, it holds that:

$$f(\lambda x + (1 - \lambda)y) \leqslant \lambda f(x) + (1 - \lambda) f(y).$$

If this holds with strict inequality then the function is said strictly convex. If $f(\cdot)$ is not identically $+\infty$ it is named a *proper* function.

The sum of two convex functions is again convex. The composition of a convex function $f : \mathbb{R}^n \to \mathbb{R}$ with a linear mapping $A : \mathbb{R}^m \to \mathbb{R}^n$, denoted as $(f \circ A)(\cdot) = f(A(\cdot))$, is again convex. The domain of a function $f(\cdot)$ is defined as $\text{dom}(f) = \{x \in \mathbb{R}^n | f(x) < +\infty\}$, so a proper function has $\text{dom}(f) \neq \emptyset$. Convex functions may have a bounded domain. For instance the indicator function of a convex set C, defined as $\psi_C(x) = 0$ if $x \in C$, $\psi_C(x) = +\infty$ if $x \notin C$, takes the value $+\infty$ everywhere outside the set C. Thus $\text{dom}(\psi_C) = C$. It is nevertheless a convex function, and $C \subseteq \mathbb{R}^n$ is a convex set if and only if $\psi_C(\cdot)$ is a convex function. Indicator functions have been introduced by J.J. Moreau in the context of unilaterally constrained mechanical systems. They may be interpreted as a nonsmooth potential function associated with the contact forces, when frictionless unilateral constraints are considered.

Differentiable convex functions, *i.e.* the functions $f(\cdot)$ which possess a gradient $\nabla f(x)$ at all $x \in \mathbb{R}^n$, enjoy the following properties.

Proposition 2.7 *Let $f : U \to \mathbb{R}$ be a function of class C^1, with $U \subset \mathbb{R}^n$ an open set, and let $C \subseteq U$ be a convex subset of U. Then $f(\cdot)$ is convex on C if and only if $f(y) \geqslant f(x) + \langle \nabla f(x), y - x\rangle$ for all x and y in C.*

We will see that this is generalized when $f(\cdot)$ fails to be C^1 but is only *subdifferentiable*. When the function is at least twice differentiable, it can be also characterized from its Hessian matrix.

Proposition 2.8 *Let $f : U \to \mathbb{R}$ be a function of class C^2, with $U \subset \mathbb{R}^n$ an open convex set. Then $f(\cdot)$ is convex on U if and only if its Hessian matrix $\nabla^2 f(x)$ is semi-positive definite for all $x \in U$,* i.e. *$\langle \nabla^2 f(x)y, y\rangle \geqslant 0$ for all $y \in \mathbb{R}^n$.*

However many convex functions are not differentiable everywhere (most of them, in fact). A first example is the above indicator function of a set C. The simplest example is the absolute value function $f : \mathbb{R} \to \mathbb{R}$, $x \mapsto |x|$, which is not differentiable in the usual sense at $x = 0$. We will see later that it is nevertheless *subdifferentiable* at $x = 0$: the usual derivative (the slope) is replaced by a *set of derivatives* (called the *subgradients*). The usual result that a convex function has a minimum at x if and only if its derivative is zero at x, extends to subdifferentiable convex functions.

Definition 2.9 (Conjugate functions) Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper convex function. The *conjugate* of $f(\cdot)$ is the function $f^*(\cdot)$ defined by:

$$\mathbb{R}^n \ni y \mapsto f^*(y) = \sup_{x \in \mathrm{dom}(f)} \{\langle y, x\rangle - f(x)\}. \tag{2.2}$$

The mapping $f \mapsto f^*$ is called the Legendre-Fenchel transform, or the conjugacy operation.

As we shall see below, the conjugacy operation is useful when one wants to invert the graph of a certain multifunction (see all definitions below) that may represent the characteristic of some electronic device. Representing the (current, voltage) characteristic or the (voltage, current) characteristic amounts then to invert a graph and this is done through the Legendre-Fenchel transform.

Theorem 2.10 (Fenchel-Moreau) *Assume that $f(\cdot)$ is convex, proper and lower semi-continuous. Then $f^{**}(\cdot) = f(\cdot)$.*

Applying twice the conjugacy operation yields the original function.

Example 2.11 Let us compute the conjugate function $g(y) = f^*(y)$ of the absolute value function $f(x) = |x|$. We get:

$$g(y) = \sup_{x \in \mathbb{R}} (\langle x, y\rangle - |x|). \tag{2.3}$$

If $x > 0$ then $g(y) = \sup_{x \in \mathbb{R}} x(y - 1)$. So if $y > 1$ one obtains $g(y) = +\infty$, and if $y \leqslant 1$ one obtains $g(y) = 0$. If $x < 0$ then $g(y) = \sup_{x \in \mathbb{R}} x(y + 1)$. So if $y \geqslant -1$ one obtains $g(y) = 0$, and if $y < -1$ one obtains $g(y) = +\infty$. If $x = 0$ clearly $g(y) = 0$. We deduce that $g(y) = \psi_{[-1,1]}(y)$, the indicator function of the interval $[-1, 1]$. By the Fenchel-Moreau theorem, it follows that $g^*(x) = f^{**}(x) = |x|$. More generally the conjugate of $f : \mathbb{R}^n \to \mathbb{R}$, $x \mapsto \|x\|$ is the indicator function of the unit ball of $\mathbb{R}^n$. The above calculations can be easily generalized by varying the slopes of the absolute value function. Take $f : \mathbb{R} \to \mathbb{R}$, $x \mapsto ax$ if $x \leqslant 0$, $x \mapsto bx$ if $x \geqslant 0$. Then $f^*(y) = \psi_{[a,b]}(y)$.

Example 2.12 Let C be a closed non empty convex cone, and C° its polar cone. Then the indicator function of C, $\psi_C(\cdot)$, is the conjugate to the indicator function of C°, *i.e.* $\psi_C^*(\cdot) = \psi_{C^\circ}(\cdot)$.

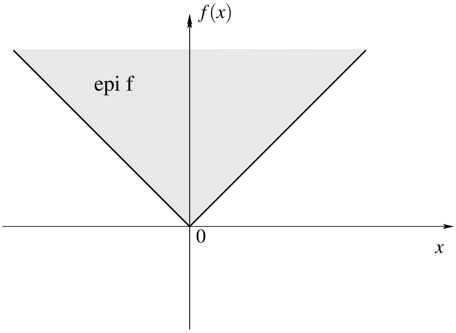


Fig. 2.4 Epigraph of the absolute value function

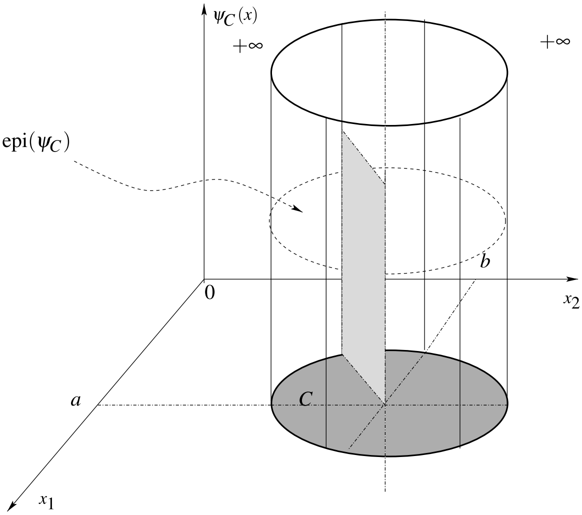


Fig. 2.5 Epigraph of the indicator of C

Let us now introduce a notion that is useful to characterize the convexity of a function, and which also permits to link convex functions and convex sets.

Definition 2.13 (Epigraph of a function) Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper function (not necessarily convex). The *epigraph* of $f(\cdot)$ is the non empty set:

$$\text{epi}(f) = \{(x, \eta) \in \mathbb{R}^n \times \mathbb{R} \mid \eta \geqslant f(x)\}.$$

Notice that η is taken in $\mathbb{R}$ so it does not take the infinite value. In particular a function is convex if and only if its epigraph is convex. This may even be taken as a definition of convex functions. The epigraph of the absolute value function is depicted in Fig. 2.4. This is a convex cone of the plane, defined as $\text{epi}(|x|) = \{(x, \eta) \in \mathbb{R} \times \mathbb{R} \mid \eta \geqslant |x|\} \subset \mathbb{R}^2$. Consider now the set $C = \{x \in \mathbb{R}^2 \mid (x_1 - a)^2 + (x_2 - b)^2 \leqslant r^2\}$ that is a closed disk with radius r centered at (a, b). The epigraph of its indicator function $\psi_C(\cdot)$ is depicted in Fig. 2.5: $\text{epi}(\psi_C) = \{(x, \eta) \in C \times \mathbb{R} \mid \eta \geqslant 0\}$. This is a half cylinder pointing outwards the plane (x_1, x_2).

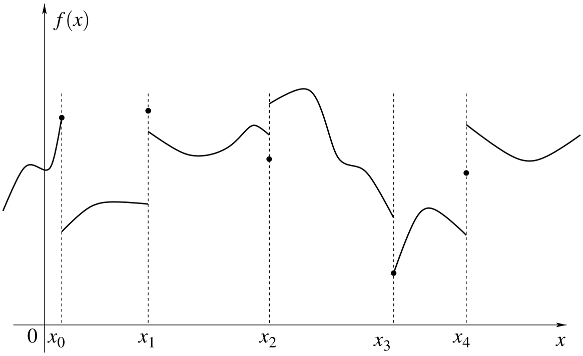


Fig. 2.6 Lower and upper semi-continuous functions

Remark 2.14 Convex functions can be identified with their epigraph. Convex sets can be identified with their indicator function. This permits to pass from functions to sets, *i.e.* from analysis to geometry.

Before introducing the next notion, let us recall that the notation lim inf means the lower limit. Given a subset $S \subseteq \mathbb{R}^n$, $l = \liminf_{y \to x} f(x)$ for $x \in \mathrm{cl}S$ means that:[1] for all $\epsilon > 0$, there exists a neighborhood $N(x)$ such that $f(y) \geqslant l - \epsilon$ for all $y \in N(x)$, and in any neighborhood $N(x)$, there is $y \in N(x)$ such that $f(x) \leqslant l + \epsilon$.

Definition 2.15 (Lower and upper semi-continuity) Let $f : S \subseteq \mathbb{R}^n \to \mathbb{R}$, and let $x \in S$. Then $f(\cdot)$ is *lower semi-continuous* at x if $f(x) \leqslant \liminf_{y \to x} f(y)$. It is *upper semi-continuous* at x if $f(x) \geqslant \limsup_{y \to x} f(y)$.

A function is both lower and upper semi-continuous at x if and only if it is continuous at x. There is a local version of lower and upper semi-continuity at a point x, which states that the property holds in a small ball centered at x. An example of a locally lower and upper semi-continuous function is depicted in Fig. 2.6. The function $f(\cdot)$ is locally lower semi-continuous at x_2 and x_3. It is locally upper semi-continuous at x_0 and x_1. It is neither lower nor upper semi-continuous at x_4. Lower semi-continuous functions have a closed epigraph. Lower semi-continuity is an important property for the existence of a minimum of a function.

Remark 2.16 For the time being we dealt only with single-valued functions, *i.e.* functions that assign to each $x \in \mathbb{R}^n$ a singleton $\{f(x)\}$. There exists a notion of upper semi-continuity for multivalued functions (see below for a definition). However it is not a generalization of the upper semi-continuity of single-valued functions, in the sense that a single-valued function that is upper semi-continuous in the sense of multivalued functions, is necessarily continuous. This is why J.-B. Hiriart-Urruty has proposed to name the multivalued upper semi-continuity the outer semi-continuity (Hiriart-Urruty and Lemaréchal 2001, §0.5), to avoid confusions.

[1] $\mathrm{cl}\, S$ is the closure of the set S.

A way to characterize the lower semi-continuity of a function $f(\cdot)$ is through its epigraph. Indeed $f(\cdot)$ is lower semi-continuous if and only if its epigraph epi(f) is closed. This can be checked in Fig. 2.6: locally the epigraph is open at x_0 and x_1, whereas it is closed at x_2 and x_3. The indicator function $\psi_C(\cdot)$ of a closed non empty set is lower semi-continuous. For instance the epigraph of the indicator function depicted in Fig. 2.5 is a closed half cylinder (an unbounded, but closed set).

Remark 2.17 As a matter of fact, convex functions that take bounded values on $\mathbb{R}^n$ (*i.e.* $\mathrm{dom}(f) = \mathbb{R}^n$) necessarily are continuous functions. They are even locally Lipschitz continuous at every point. This means that the semi-continuity is a notion that is automatically satisfied by bounded convex functions. The only convex function we shall meet for which this is not the case is the indicator of a convex set of $\mathbb{R}^n$, that is not continuous on $\mathbb{R}^n$ but is lower semi-continuous.

Definition 2.18 (Normal and tangent cones to a non empty convex set) Let $C \subseteq \mathbb{R}^n$ be a closed convex set. The (outward) *normal cone* to C at $x \in C$ is the set:

$$N_C(x) = \{s \in \mathbb{R}^n \mid \langle s, y - x\rangle \leqslant 0 \quad \text{for all } y \in C\}.$$

The *tangent cone* to C at $x \in C$ is the set:

$$T_C(x) = \Big\{ y \in \mathbb{R}^n \mid \exists\, (x_k)_{k \geqslant 0}, x_k \in C \text{ with } \lim_{k \to +\infty} x_k = x, \text{ and } \exists\, (\alpha_k)_{k\geqslant 0}, \alpha_k \geqslant 0, \\ \text{such that } \lim_{k\to+\infty} \alpha_k = 0 \text{ and } \lim_{k\to+\infty} x_k = \frac{x_k - x}{\alpha_k} = y \Big\}.$$

There are other, equivalent ways to define the tangent cone, like

$$T_C(x) = \mathrm{cl}\Big(\bigcup_{y \in C} \bigcup_{\lambda > 0} \lambda(y - x) \Big),$$

where cl$(\cdot)$ denotes the closure (the closure of a set $S \subseteq \mathbb{R}^n$ is the set plus its boundary; it is also the smallest closed set of $\mathbb{R}^n$ that contains S). It is important to remark that the normal cone is defined through a *variational* process: one varies y inside C to find the normal vectors s that form $N_C(x)$. The normal cone (see Fig. 2.7) is the *outward* normal cone, *i.e.* it points outside the set C. The definition of a tangent cone as given in Definition 2.18 is not very friendly. There is a much simpler way to characterize the tangent cone when C is convex, as the next proposition shows.

Proposition 2.19 *Let $C \subset \mathbb{R}^n$ be a closed non empty convex set and let $x \in C$. Then the tangent and normal cones are closed convex cones, and $N_C(x) = (T_C(x))^\circ$ and $T_C(x) = (N_C(x))^\circ$.*

Therefore starting from the definition of the normal cone, we may state at $x \in C$:

$$T_C(x) = \{d \in \mathbb{R}^n \mid \langle s, d\rangle \leqslant 0 \text{ for all } s \in N_C(x)\},$$

which is also a variational definition of the tangent cone. One finds that when $x \in \mathrm{Int}(C)$, then $N_C(x) = \{0\}$ and $T_C(x) = \mathbb{R}^n$.

Fig. 2.7 Normal cones

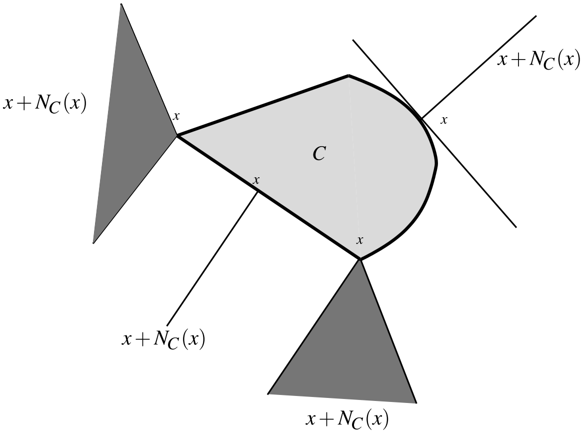
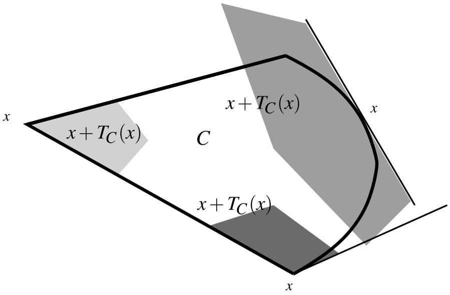


Fig. 2.8 Tangent cones



One sees in Fig. 2.8 that the tangent cones locally reproduce the "shape" of the set C. When C is polyhedral at x then $T_C(x) \approx C$. When C is differentiable at x then $T_C(x)$ is an inwards halfspace. It is also visible in the figures that the tangent and normal cones are polar cones one to each other. The fact that both the normal and tangent cones to C at x are the empty set when $x \notin C$ is a consequence of the definition of the indicator function of C, that takes infinite values in such a case.

Example 2.20 (Closed convex polyhedra) Let us assume that the set C is defined as $C = \{x \in \mathbb{R}^n \mid Ex + F \leqslant 0, E \in \mathbb{R}^{m\times n}, F \in \mathbb{R}^m\}$. In other words C is defined with m inequalities $E_i x + F_i \leqslant 0$ where the m vectors $E_i \in \mathbb{R}^{1\times n}$ are the rows of the matrix E and the F_is are the components of F. Let us define the set of the *active constraints* at $x \in C$ as

$$I(x) = \{i = 1, \dots, m \mid E_i x + F_i = 0\}$$

that is a set of indices. Then:

$$T_C(x) = \{d \in \mathbb{R}^n \mid \langle E_i, d\rangle \leqslant 0 \text{ for } i \in I(x)\}, \tag{2.4}$$

and

$$N_C(x) = \left\{ \sum_{i \in I(x)} \alpha_i E_i^T,\ \alpha_i \geqslant 0 \right\}. \tag{2.5}$$

Therefore the normal cone is generated by the outwards normal vectors to the facets that form the set C at x. When $x \notin C$ one ususally defines $T_C(x)$ and $N_C(x)$ both equal to $\emptyset$.

Remark 2.21 The fact that $N_C(\cdot)$ and $T_C(\cdot)$ are polar cones has a strong physical meaning. In mechanical systems subject to frictionless unilateral constraints, (normal) contact forces belong to $N_C(\cdot)$ whereas velocities belong to $T_C(\cdot)$. Thus the contact forces and the velocity form a pair of reciprocal variables (sometimes also called dual variables), whose product is a mechanical power. In electricity the voltage and the current are reciprocal variables since their product is an electrical power.

2.1.1.2 Subdifferentiation

Definition 2.22 (Subgradients, subdifferentials) A vector $\gamma \in \mathbb{R}^n$ is said to be a *subgradient* of a convex function $f(\cdot)$ at a point x if it satisfies:

$$f(y) - f(x) \geqslant \gamma^T (y - x) \tag{2.6}$$

for all $y \in \mathbb{R}^n$. The set of all subgradients of $f(\cdot)$ at x is the *subdifferential* of $f(\cdot)$ at x and is denoted $\partial f(x)$.

When $f(x)$ is finite, the inequality (2.6) says that the graph of the affine function $h(y) = f(x) + \gamma^T(y - x)$ is a non vertical supporting hyperplane to the convex epigraph of $f(\cdot)$ at $(x, f(x))$, see (2.8) below. If a function $f(\cdot)$ is differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$. The following holds:

Theorem 2.23 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. Then $f(\cdot)$ is minimized at x over $\mathbb{R}^n$ if and only if $0 \in \partial f(x)$.*

This is a generalization of the usual stationarity condition for differentiable functions.

Proposition 2.24 *Let $f(\cdot)$ be a lower semi-continuous, proper and convex function. Then $\partial f(\cdot)$ is a closed convex set, possibly empty. If $x \in \mathrm{Int}(\mathrm{dom}(f))$, then $\partial f(x) \neq \emptyset$. In particular, if $f : \mathbb{R}^n \to \mathbb{R}$ is convex, then for all $x \in \mathbb{R}^n$, $\partial f(x)$ is a non empty, convex and compact set of $\mathbb{R}^n$.*

Example 2.25 Let us start with the absolute value function. If $x \neq 0$, then it is differentiable and $\partial|x| = \{1\}$ if $x > 0$, $\partial|x| = \{-1\}$ if $x < 0$. At $x = 0$ one looks for reals γ such that $|y| \geqslant \gamma y$ for all reals y. If $y > 0$ one finds $\gamma \leqslant 1$. If $y < 0$ then one finds $\gamma \geqslant -1$. One concludes that $-1 \leqslant \gamma \leqslant 1$. Therefore $\partial|0| = [-1, 1]$. That $x = 0$ is a minimum is obvious.

Example 2.26 (Normal cone as the subdifferential of the indicator function) Let $C \subseteq \mathbb{R}^n$ be a non empty closed convex set, and such that $\mathrm{Int}(C)$ contains an n-dimensional ball of radius $r > 0$. Then the subgradients of the indicator function of C at x are the vectors γ satisfying $\psi_C(y) - \psi_C(x) \geqslant \gamma^T(y - x)$ for all $y \in \mathbb{R}^n$. Let $x \in \mathrm{Int}(C)$. We get $\psi_C(y) \geqslant \gamma^T(y - x)$. Let $y \in \mathrm{Int}(C)$, so that $0 \geqslant \gamma^T(y - x)$ for all $y \in \mathrm{Int}(C)$. In view of the assumptions on x and C there exists a ball of positive radius centered at x, contained in $\mathrm{Int}(C)$. We may choose y in C such that $y - x$ is anywhere inside this ball. It follows that necessarily $\gamma = 0$. Therefore $\partial\psi_C(x) = \{0\}$ when $x \in \mathrm{Int}(C)$. Let now $x \notin C$, so that $\psi_C(y) \geqslant +\infty + \gamma^T(y - x)$ for all $y \in \mathbb{R}^n$. Take for instance $y \in C$ so that we get $\gamma^T(x - y) \geqslant +\infty$. This is impossible and we conclude that $\partial\psi_C(x) = \emptyset$ when $x \notin C$. Let now $x \in \mathrm{Bd}(C)$, the boundary of the set C. We get $\psi_C(y) \geqslant \gamma^T(y - x)$ for all $y \in \mathbb{R}^n$. Take $y \in C$, then the subgradients have to satisfy $\gamma^T(y - x) \leqslant 0$ for all $y \in C$. Precisely, such vectors γ belong to the normal cone $N_C(x)$, see Definition 2.18. We conclude that provided one takes as a convention that $N_C(x) = \emptyset$ if $x \notin C$, then $\partial\psi_C(\cdot) = N_C(\cdot)$.

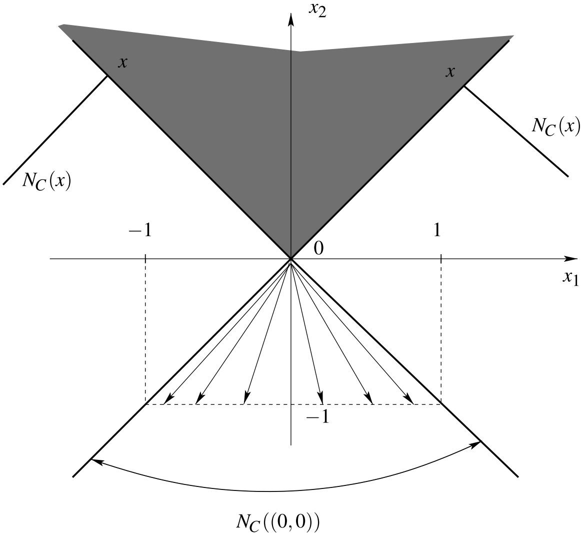
Example 2.27 (Normal cone to a finitely represented set) If C is finitely represented, *i.e.* $C = \{x \in \mathbb{R}^n \mid g(x) \leqslant 0\}$, with $g(\cdot)$ lower semi-continuous, proper, and convex such that $0 \notin \partial g(x)$, then:

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

The three different cases correspond respectively to x in the interior of C, x outside C, and x on the boundary of C. The notation $\mathbb{R}_+\partial g(x)$ is for $\{\lambda\eta \mid \lambda > 0$ and $\eta \in \partial g(x)\}$. One can say that on $\mathrm{Bd}(C)$ the normal cone is generated by the subgradients of the function $g(\cdot)$. Consider for instance the set of $\mathbb{R}^2$ defined as $C = \{(x_1, x_2) \mid x_2 \geqslant |x_1|\}$. Thus $g(x) = |x_1| - x_2$, and there is a corner at $x_1 = x_2 = 0$. One has $\partial g(0, 0) = \binom{[-1,1]}{-1}$. Therefore at the corner point $N_C((0, 0)) = \{\lambda\eta \mid \eta_1 \in [-1, 1], \eta_2 = -1, \lambda > 0\}$. We will see below that this can be interpreted as the normal cone to the epigraph of the absolute value function. This is depicted in Fig. 2.9.

Remark 2.28 This notion of a generalized derivative of a convex function that is not differentiable in the usual sense, is totally disjoint from the notion of generalized derivatives in the sense of Schwartz' distributions. A Schwartz' distribution T is a functional (*i.e.* a function of functions) which associates with test functions $\varphi(\cdot)$ taken in a special space of functions, a real (or complex) number denoted $\langle T, \varphi\rangle$. For instance, the generalized derivative of the absolute value function in the sense of Schwartz' distributions, is the function $f : \mathbb{R} \to \mathbb{R}$ with $f(x) = -1$ is $x < 0$, $f(x) = 1$ if $x > 0$, and $f(0)$ can be given any bounded value. The distribution is then defined as $\langle T, \varphi\rangle = \int_{\mathrm{dom}(\varphi)} f(t)\varphi(t)dt$. The so-called Heavyside function has a generalized derivative that is the Dirac measure at $t = 0$, however it is not a convex function and therefore does not possess a subdifferential in the sense of Definition 2.22.

Fig. 2.9 Normal cones to a finitely represented set



The usual differentiation rule for composed functions, the so-called chain rule, extends to subdifferentiation as follows.

Proposition 2.29 (Chain rule) *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a convex lower semi-continuous function, and $A : \mathbb{R}^m \to \mathbb{R}^n$ be an affine mapping.*[2] *Assume that a point $y_0 = Ax_0$ exists at which $f(\cdot)$ is finite and continuous. The subdifferential in the sense of convex analysis of the composite functional $f \circ A : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is given by*

$$\partial(f \circ A)(x) = A_0^T \partial f(Ax), \quad \forall\, x \in \mathbb{R}^n. \tag{2.7}$$

For the sum of convex functions the result is as follows.

Theorem 2.30 (Moreau-Rockafellar: subdifferentiate of a sum) *Let $f_i : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, $1 \leqslant i \leqslant 2$ be proper convex functions, and let $f(\cdot) = \sum_{i=1}^2 f_i(\cdot)$. Assume that the convex sets* $\mathrm{dom}(f_i)$, $1 \leqslant i \leqslant 2$, *have a point in common $\bar{x}$ and that $f_1(\cdot)$ is continuous at $\bar{x}$. Then*

$$\partial f(x) = \sum_{i=1}^{2} \partial f_i(x), \quad \textit{for all } x \in \mathrm{dom}(f_1) \cap \mathrm{dom}(f_2).$$

The result can obviously be extended to cope with the sums of more than two functions.

[2] *I.e.* $Ax = A_0 x + b$ with A_0 linear.

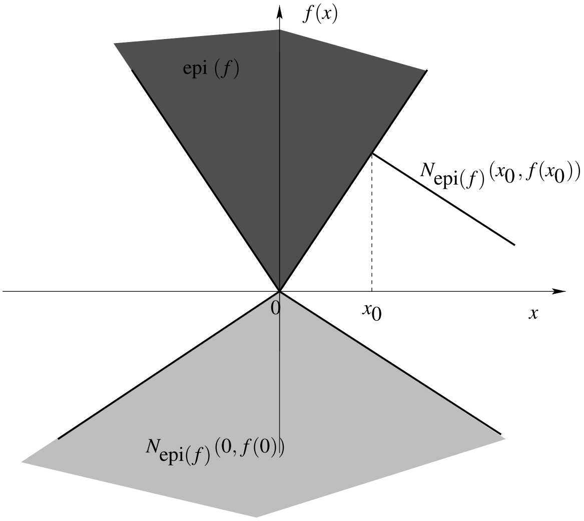

$f(x)$
epi (f)
$N_{\text{epi}(f)}(x_0, f(x_0))$
0
x_0
x
$N_{\text{epi}(f)}(0, f(0))$

Fig. 2.10 Normal cone to the epigraph

The Normal Cone to the Epigraph There is a relationship between the subgradients of a function and the normal cone to the epigraph of the function. Indeed:

$$N_{\text{epi}\, f}(x, f(x)) = \{(\lambda\gamma, -\lambda), \gamma \in \partial f(x) \text{ and } \lambda \geqslant 0\}. \tag{2.8}$$

The normal cone to the epigraph is therefore generated by the vectors $(\gamma, -1)$ where γ is a subgradient. Normal cones to the epigraph of the function $f(x) = ax$ if $x \leqslant 0$ and $f(x) = bx$ if $x \geqslant 0$ are depicted in Fig. 2.10.

Inversion of Graphs The graph of the subdifferential $\partial f(\cdot)$ is equal to the set

$$\text{gr}(\partial f) = \{(x, y) \in \mathbb{R}^n \times \mathbb{R}^n \mid y \in \partial f(x)\}.$$

The conjugacy operation on a convex lower semi-continuous proper function (*i.e.* not identically equal to $+\infty$) defines the inversion of the graph of its subdifferential. In fact, if $f(\cdot)$ is a closed proper convex function, $\partial f^*(\cdot)$ is the inverse of $\partial f(\cdot)$ in the sense of multivalued mappings. In other words:

$$x \in \partial f(y) \quad \text{if and only if } y \in \partial f^*(x). \tag{2.9}$$

This is illustrated in Fig. 2.11 for the absolute value function (see Examples 2.11, 2.25 and 2.26). In particular one has $N_{[-1,1]}(1) = \mathbb{R}_+$ and $N_{[-1,1]}(-1) = \mathbb{R}_-$. Inversion of graphs occurs when passing from $(i(t), v(t))$ to $(v(t), i(t))$ characteristics of electronic devices, see for instance the Zener diode voltage/current law in Fig. 1.8.

Link with Optimization let $f(\cdot)$ be a proper lower semi-continuous convex function. We consider the constrained optimisation problem:

$$\text{(COPT):} \quad \min_{x \in C} f(x) = \min_{x \in \mathbb{R}^n} (f + \psi_C)(x).$$

Clearly one has:

$$x \text{ is a solution of (COPT)} \quad \Leftrightarrow \quad 0 \in \partial(f + \psi_C)(x).$$

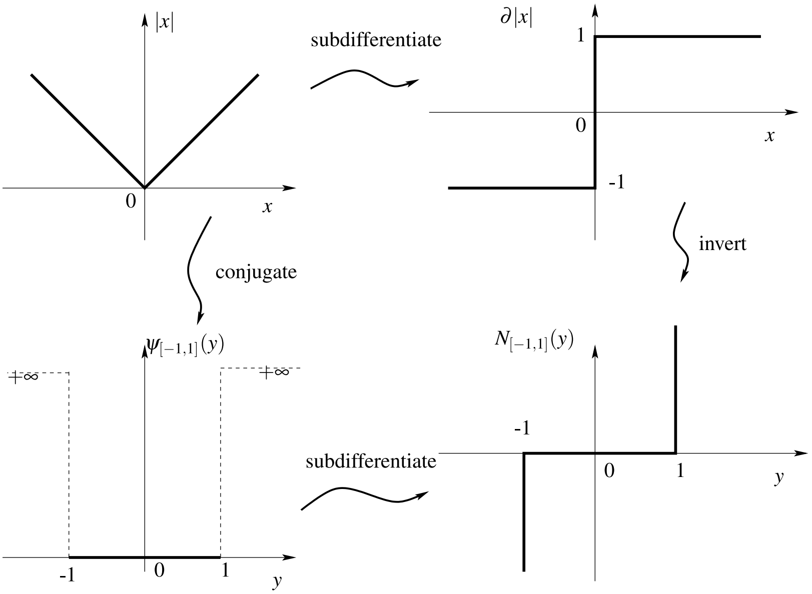


Fig. 2.11 Conjugating, subdifferentiating and inverting

Now if $f(\cdot)$ is continuous at a point in C we can rewrite the right-hand-side of the equivalence as (see Theorem 2.30):

$$0 \in \partial f(x) + \partial \psi_C(x) = \partial f(x) + N_C(x). \tag{2.10}$$

If $f(\cdot)$ is of class C^1 one obtains $-\nabla f(x) \in N_C(x)$ as a necessary and sufficient condition to be fullfilled by a solution.

Remark 2.31 Convex functions can be identified with their epigraph. Convex sets can be identified with their indicator function. This permits to pass from functions to sets, *i.e.* from analysis to geometry.

2.1.2 Multivalued Functions

The normal cone to a convex set $C \subseteq \mathbb{R}^n$ defines a multivalued mapping, since it assigns to each x in C a set $N_C(x) \subseteq \mathbb{R}^n$. Normal cones are an important example of set-valued mappings, or multifunctions. Another example taken from the previous section is the subdifferential $\partial f(\cdot)$ when $f(x) = |x|$. At $x = 0$ one has $\partial f(0) = [-1, 1]$. We conclude that the subdifferentials of convex functions $f(\cdot)$ usually are multifunctions $x \mapsto \partial f(x)$.

2.1.2.1 Definitions

Definition 2.32 (Multivalued function, domain, image, graph, inverse map) A multivalued function $F(\cdot)$ (or multi-function, or set-valued function, or set-valued map) from a normed space X to a normed space Y is a map that associates with any $x \in X$ a *set*: $F(x) \subset Y$. A multifunction is completely characterized by its *graph*, defined as

$$\text{gr}(F) = \{(x, y) \in X \times Y \mid y \in F(x)\}.$$

The *domain* of the multifunction $F(\cdot)$ is the set

$$\text{dom}(F) = \{x \in X \mid F(x) \neq \emptyset\}.$$

The *image* of the multivalued function $F(\cdot)$ is defined as

$$\text{im}(F) = \{y \in Y \mid \exists\, x \in X \text{ such that } y \in F(x)\}.$$

The *inverse map* $F^{-1} : Y \to X$ of $F(\cdot)$ is defined by:

$$F^{-1}(y) = \{x \in X \mid (x, y) \in \text{gr}(F)\}.$$

In most applications X and Y are subsets of or equal to $\mathbb{R}^n$ or $\mathbb{R}^m$, respectively, for some n and m. Notice that some authors adopt the convention $F : X \rightrightarrows Y$ to distinguish multivalued mappings, which we shall not do here. There are several different classes of multivalued maps. As pointed out in the introduction of this section, subdifferentials are multivalued functions. These are in fact the most common multifunctions we will encounter in this monograph. Other examples are:

- $F : \mathbb{R} \to \mathbb{R}$, $x \mapsto [-1, 1]$, which assigns to each x an interval, see Fig. 2.12(a).
- $F : \mathbb{R} \to \mathbb{R}$, $x \mapsto [-|x|, |x|]$, see Fig. 2.12(b).
- The inverse of many single-valued function is set-valued. For instance the function in Fig. 2.12(c) is single valued, and its inverse in Fig. 2.12(d) is set-valued since $F(0) = [-a, b]$ ($a > 0$, $b > 0$).

We shall not meet multifunctions of the type of Fig. 2.12(a) and (b) in this book.

2.1.2.2 Maximal Monotone Mappings

Definition 2.33 (Maximal monotone mapping) A multivalued mapping $F : S \subseteq \mathbb{R}^n \to \mathbb{R}^n$ is said to be *monotone* on S if for every pairs (x_1, y_1) and (x_2, y_2) in its graph one has:

$$\langle x_1 - x_2, y_1 - y_2 \rangle \geqslant 0. \tag{2.11}$$

It is *strictly* monotone on S if the inequality is strict > 0 for all $x \neq y$. It is $\xi-$monotone on S if there exists a constant $c > 0$ such that:

$$\langle x_1 - x_2, y_1 - y_2 \rangle \geqslant c\|x_1 - x_2\|^{\xi}. \tag{2.12}$$

If $\xi = 2$ is *strongly* monotone on S. It is *maximal* monotone if its graph is not properly contained in the graph of any other monotone mapping.

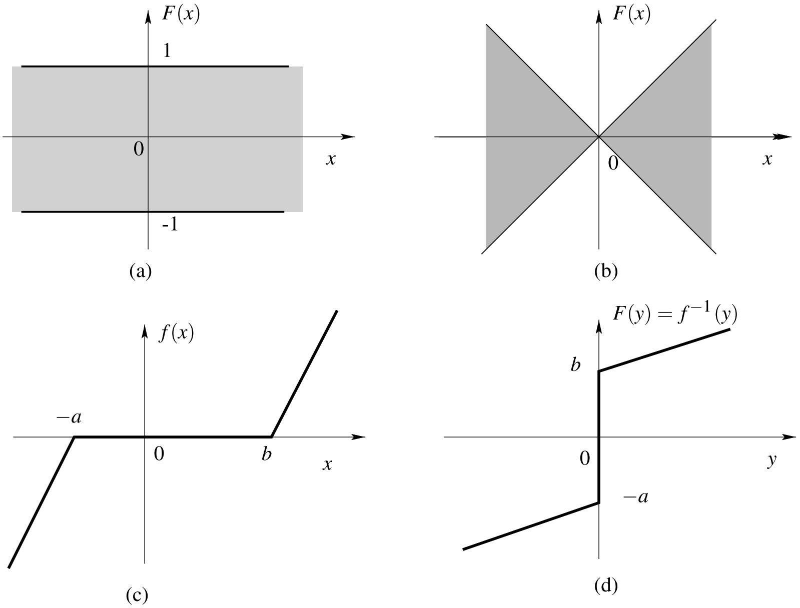


Fig. 2.12 Multivalued functions

The maximality is to be understood in terms of inclusions of graphs. If the mapping is maximal, then adding anything to its graph so as to obtain the graph of a new multivalued mapping, destroys the monotonicity (the extended mapping is no longer monotone). In other words, for every pair $(x, y) \in (\mathbb{R}^n \times \mathbb{R}^n) \setminus \mathrm{gr}(F)$ there exists $(x', y') \in \mathrm{gr}(F)$ such that $\langle x - x', y - y' \rangle < 0$. This is illustrated in Fig. 2.13. The mapping whose graph is in Fig. 2.13(a) is monotone, however it is not maximal. The one in Fig. 2.13(b) is maximal monotone. Intuitively, starting from a monotone mapping, maximality is obtained after "filling-in" the gaps (consequently continuous monotone mappings are maximal). In the planar case maximal monotone mappings have a non decreasing curve.

Operations that Preserve the Monotonicity, and Some Properties

- If $F : \mathbb{R}^n \to \mathbb{R}^n$ is monotone then its inverse mapping $F^{-1}(\cdot)$ is monotone (in the single valued case, a non decreasing function has a non decreasing inverse).
- If $F : \mathbb{R}^n \to \mathbb{R}^n$ is monotone then $\lambda F(\cdot)$ is monotone for any $\lambda > 0$.
- If $F_1 : \mathbb{R}^n \to \mathbb{R}^n$ and $F_2 : \mathbb{R}^n \to \mathbb{R}^n$ are monotone, then $(F_1 + F_2)(\cdot)$ is monotone.
- $F : \mathbb{R}^n \to \mathbb{R}^n$ is monotone, then for any matrix A and vector b, the mapping $T(x) = A^T F(Ax + b)$ is monotone.
- $F(\cdot)$ is maximal monotone if and only if $F^{-1}(\cdot)$ is maximal monotone.
- The graph of a maximal monotone mapping is closed.
- If $F(\cdot)$ is maximal monotone, then both $F(\cdot)$ and $F^{-1}(\cdot)$ are closed-convex-valued.

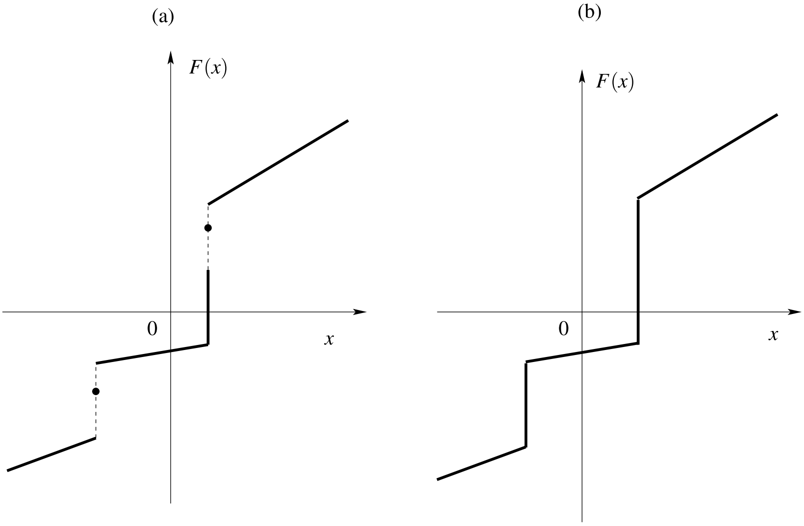


Fig. 2.13 Monotone mappings

Link with Subdifferentials of Convex Functions The following holds, which is the generalization that when a convex function $\mathbb{R} \to \mathbb{R}$ is differentiable, then its gradient is non decreasing.

Theorem 2.34 *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be convex and proper. Then the multivalued mapping $\partial f : \mathbb{R}^n \to \mathbb{R}^n$ is monotone. A proper lower semi-continuous function is convex if and only if $\partial f(\cdot)$ is maximal monotone.*

As a corollary, the normal cone to a non empty closed convex set of $\mathbb{R}^n$ is a maximal monotone mapping. Indeed the indicator function of such a set is proper, lower semi-continuous and convex. Let M be a positive semidefinite matrix (not necessarily symmetric). Then the mapping $x \mapsto Mx$ is maximal monotone. If M is positive definite then it is even strongly monotone.

2.1.2.3 Generalized Equations

A generalized equation is an equation of the form $0 \in F(x)$, where $F(\cdot)$ is a multivalued function. It is of great interest to study the conditions that assure the existence and the uniqueness of solutions to such equations, as a prerequisite to the development of efficient numerical algorithms to solve them (see for instance (2.10) that represents the necessary and sufficient conditions of a constrained optimisation problem). The notion of monotonicity has long been recognized as a crucial property that guarantees the well-posedness of generalized equations. The next result concerns generalized equations of the form:

$$0 \in F(x) + N_C(x), \tag{2.13}$$

where $C \subseteq \mathbb{R}^n$ and $F : C \to \mathbb{R}^n$ is a function. Implicitly it is understood that the solution satisfies $x \in C$, since otherwise $N_C(x) = \emptyset$. Let C be convex. This generalized equation therefore states that $-F(x) \in N_C(x)$, *i.e.* $-F(x)$ is a subgradient of the indicator function $\psi_C(\cdot)$ at the point x. We have already encountered such a generalized equation in (2.10).

Theorem 2.35 *Let C be closed convex and $F(\cdot)$ be continuous. Then*:

- *If $F(\cdot)$ is strictly monotone on C, the generalized equation in* (2.13) *has at most one solution.*
- *If $F(\cdot)$ is ξ-monotone on C for some $\xi > 1$, the generalized equation in* (2.13) *has a unique solution.*

Example 2.36 Let $F : \mathbb{R}^2 \to \mathbb{R}^2$, $x \mapsto \binom{\cos x}{\sin x}$, and $C = \{(x_1, x_2) \in \mathbb{R}^2 \mid x_1 \geqslant 0\}$. We know from (2.5) that $N_C(0) = \left\{\alpha\binom{-1}{0}, \alpha \geqslant 0\right\} = \mathbb{R}_-\mathbf{e}_1$ where $\mathbf{e}_1 = (1\ 0)^T$. We deduce that all x with $x_1 = 2k\pi$, $k \geqslant 0$, are solutions of the generalized equation $-F(x) \in N_C(x)$. Clearly $F(\cdot)$ is not monotone on C. Notice in passing that the solutions have to lie on the boundary of C, for otherwise one has $N_C(x) = \{(0\ 0)^T\}$ in the interior of C and it is impossible to have both components of $F(\cdot)$ which vanish at the same time.

Let us now state a result which related inclusions into normal cones and projections, for a particular value of the function $F(\cdot)$.

Proposition 2.37 *Let $M = M^T > 0$ be a $n \times n$ matrix, and $C \subseteq \mathbb{R}^n$ be a closed convex non empty set. Then*

$$\begin{gathered} M(x-y) \in -N_C(x) \\ \Updownarrow \\ x = \operatorname*{argmin}_{z \in C} \frac{1}{2}(z-y)^T M(z-y) \\ \Updownarrow \\ x = \mathrm{proj}_M(C; y), \end{gathered} \tag{2.14}$$

where proj_M *indicates that the projection is done in the metric defined by M.*

Notice one thing: we may rewrite the first inclusion as $Mx + N_C(x) \ni My$, *i.e.* $(M \cdot + N_C)(x) = My$. Let M be positive semidefinite. Then using basic arguments from nonsmooth analysis one may deduce that the operator $x \mapsto Mx + N_C(x)$ is maximal monotone, being the sum of two maximal monotone operators. Thus it has an inverse operator that is also maximal monotone and we may write $x = (M \cdot + N_C)^{-1}(My)$. In case M is definite positive symmetric we recover the projection operator.

2.2 Non Convex Sets

All the sets that we will meet in this book are convex sets, therefore we shall not need extensions of the foregoing definitions to the non convex case. Let us just mention in passing that such generalizations exist, and may be useful in other fields like contact mechanics where the sets one works with usually are *finitely represented*. That is, there exists functions $f_i : \mathbb{R}^n \to \mathbb{R}$, $1 \leqslant i \leqslant m$, such that $\mathbb{R}^n \ni C = \{x \in \mathbb{R}^n \mid f_i(x) \leqslant 0, 1 \leqslant i \leqslant m\}$. When the functions $f_i(\cdot)$ are linear, or affine functions of the form $f_i(x) = A_i x + a_i$ such that C is not empty, then C is a polyhedron, hence it is convex. When the functions $f_i(\cdot)$ are nonlinear, assuming the convexity of C is much too stringent and other notions have to be used.

2.3 Basics from Complementarity Theory

In Chap. 1 we have seen that complementarity is a notion that is often met in the nonsmooth modeling approach of electronic devices and mechanical systems with unilateral constraints. Complementarity theory is the branch of applied mathematics that deals with problems involving complementarity relations. There are many different such problems and we will present only few of them (see for instance Acary and Brogliato 2008 for an introduction, Facchinei and Pang 2003 and Cottle et al. 1992 for more complete presentations). Most importantly we shall insist on the links that exist between complementarity problems and convex analysis, normal cones to convex sets, generalized equations, and variational inequalities.

2.3.1 Definitions

Definition 2.38 (Linear Complementarity Problem (LCP)) Let $M \in \mathbb{R}^{n\times n}$ be a constant matrix, $q \in \mathbb{R}^n$ be a constant vector. A *linear complementarity problem* (LCP) is a problem of the form:

$$\begin{cases} z \geqslant 0, \\ w = Mz + q \geqslant 0, \\ w^T z = 0, \end{cases} \tag{2.15}$$

where z is the unknown of the LCP.

A more compact way to write the complementarity between two variables w and z is:

$$0 \leqslant w \perp z \geqslant 0. \tag{2.16}$$

This is adopted in the sequel. We will often name (2.16) the complementarity relations, or complementarity conditions between w and z. Strictly speaking, the *complementarity constraint* is the equality $w^T z = 0$. It is also worth noting that due to the non negativity conditions, $w^T z = 0$ is equivalent to its componentwise form $w_i z_i = 0$ for all $i \in \{1, \ldots, n\}$.

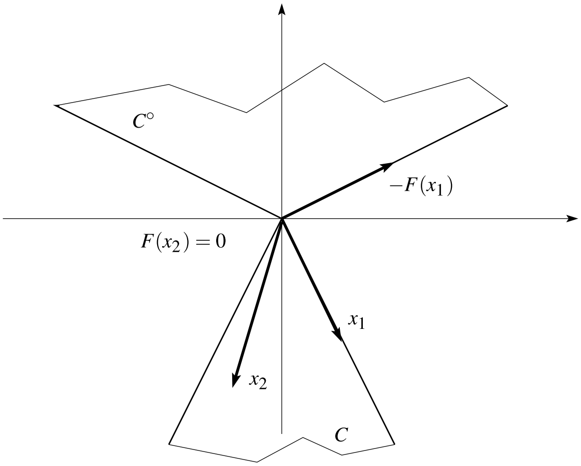


Fig. 2.14 Cone complementarity problem

Definition 2.39 (Nonlinear Complementarity Problem (NCP)) Let $F:\mathbb{R}^n \to \mathbb{R}^n$ be a nonlinear function. A *nonlinear complementarity problem* (NCP) is a problem of the form:

$$0 \leqslant x \perp F(x) \geqslant 0 \tag{2.17}$$

where x is the unknown of the NCP.

When $F(x)$ is affine then one obtains an LCP.

Definition 2.40 (Cone Complementarity Problem (CCP)) Let $C \subset \mathbb{R}^n$ be a cone, and $F: C \to \mathbb{R}^n$ a mapping. A Cone Complementarity Problem (CCP) is a problem of the form:

$$C \ni x \perp F(x) \in C^* \tag{2.18}$$

where x is the unknown of the CCP.

Obviously we may also write equivalently $C \ni x \perp -F(x) \in C^\circ$ using the polar cone. The LCP is a CCP with $F(\cdot)$ affine and $C = \mathbb{R}^n_+$. A CCP in the plane is depicted in Fig. 2.14. It is apparent that for $F(x)$ to be non zero, x has to lie on the boundary of C. When x is in the interior of C then $F(x) = (0\ 0)^T$, due to the orthogonality imposed between x and $F(x)$ and the fact that the boundaries of polar cones satisfy some orthogonality constraints. One therefore finds again a similar conclusion to the one drawn in Example 2.36. This suggests a close relation between the CCP and normal cones, see Sect. 2.3.3 for a confirmation of this observation.

Definition 2.41 (Mixed Linear Complementarity Problem (MLCP)) Given the matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{m\times m}$, $C \in \mathbb{R}^{n\times m}$, $D \in \mathbb{R}^{m\times n}$, and the vectors $a \in \mathbb{R}^n$, $b \in$

$\mathbb{R}^m$, the *mixed linear complementarity problem* denoted by MLCP(A, B, C, D, a, b) consists in finding two vectors $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^m$ such that

$$\begin{cases} Au + Cv + a = 0, \\ 0 \leqslant v \perp Du + Bv + b \geqslant 0. \end{cases} \tag{2.19}$$

The MLCP can be defined equivalently in the following form denoted by MLCP$(M, q, \mathscr{E}, \mathscr{I})$

$$\begin{cases} w = Mz + q, & \\ w_i = 0, & \forall i \in \mathscr{E}, \\ 0 \leqslant z_i \perp w_i \geqslant 0, & \forall i \in \mathscr{I}, \end{cases} \tag{2.20}$$

where $\mathscr{E}$ and $\mathscr{I}$ are finite sets of indices such that card$(\mathscr{E} \cup \mathscr{I}) = n$ and $\mathscr{E} \cap \mathscr{I} = \emptyset$.

The MLCP is a mixture between an LCP and a system of linear equations. In this book we shall see that MLCPs are common in nonsmooth electrical circuits, arising directly from their physical modeling and their time-discretization. To pass from (2.19) to (2.20), one may do as follows: define $z = \binom{u}{v}$, $M = \begin{pmatrix} A & C \\ D & B \end{pmatrix}$, $q = \binom{a}{b}$.

There is another way to define mixed complementarity problems as follows:

Definition 2.42 (Mixed Complementarity Problem (MCP)) Given a function $F : \mathbb{R}^q \to \mathbb{R}^q$ and lower and upper bounds $l, u \in \bar{\mathbb{R}}^q$, find $z \in \mathbb{R}^q$, $w, v \in \mathbb{R}^q_+$ such that

$$\begin{cases} F(z) = w - v, \\ l \leqslant z \leqslant u, \\ (z - l)^T w = 0, \\ (u - z)^T v = 0, \end{cases} \tag{2.21}$$

where $\bar{\mathbb{R}} = \mathbb{R} \cup \{+\infty, -\infty\}$.

Note that the problem (2.21) implies that

$$-F(z) \in N_{[l,u]}(z). \tag{2.22}$$

The relation (2.22) is equivalent to the MCP (2.21) if we assume that w is the positive part of $F(z)$, that is $w = F^+(z) = max(0, F(z))$ and v is the negative part of $F(z)$, that is $v = F^-(z) = max(0, -F(z))$. In case $F(z) = Mz + q$ one obtains a mixed linear complementarity problem.

2.3.2 Complementarity Problems: Existence and Uniqueness of Solutions

The fact that an LCP possesses at least one, several, or no solutions, heavily depends on the properties of the matrix M in (2.15). For instance, the LCP

$$0 \leqslant \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \perp \begin{pmatrix} 0 \\ -1 \end{pmatrix} + \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \geqslant 0$$

has an infinity of solutions of the form $x = (x_1\ 0)^T$, $x_1 \geqslant 1$. On the other hand, the scalar LCP

$$0 \leqslant x \perp -x + q \geqslant 0$$

has no solution if $q = -1$. Indeed the orthogonality implies $x(x + 1) = 0$, that is $x = 0$ or $x = -1$. The second solution is not acceptable, and $x = 0$ yields $-1 \geqslant 0$. If $q = 0$ there is a unique solution $x = 0$. If $q = 1$ there are two solutions: $x = 0$ and $x = 1$.

The fundamental result of complementarity theory is as follows:

Theorem 2.43 *The LCP $0 \leqslant x \perp Mx + q \geqslant 0$ has a unique solution for all q if and only if M is a P-matrix.*

This was proved by Samelson et al. (1958). The important point of this theorem is that the "if and only if" condition holds because one considers all possible vectors q. As the above little example shows, by varying q one may obtain LCPs whose matrix is not a P-matrix and which anyway do possess solutions, possibly a unique solution. A P-matrix is a matrix that has all its principal minors positive.[3] A positive definite matrix is a P-matrix. In turn a P-matrix that is symmetric, is positive definite. However many P-matrices are neither symmetric nor positive definite. For instance the matrices

$$\begin{pmatrix} 2 & 24 \\ 0 & 2 \end{pmatrix}, \quad \begin{pmatrix} 2 & 1 \\ 2 & 2 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 \\ 6 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & -1 & -3 \\ 1 & 1 & 1 \\ 1 & -3 & \alpha \end{pmatrix}$$

with $\alpha > 0$, are P-matrices. The determinants of the second and the third matrices are negative, so they are not positive definite. The following holds (Lootsma et al. 1999):

Lemma 2.44 *If $M \in \mathbb{R}^{n \times n}$ is a P-matrix, then M^{-1} is a P-matrix.*

Consequently the class of P-matrices plays a crucial role in complementarity problems. Other classes of matrices exist which assure the existence of solutions to LCPs. For instance *copositive* matrices and P_0-matrices. A matrix $M \in \mathbb{R}^{n \times n}$ is said copositive on a cone $C \subseteq \mathbb{R}^n$ if $x^T Mx \geqslant 0$ for all $x \in C$. It is *strictly copositive* on a cone $C \subseteq \mathbb{R}^n$ if $x^T Mx > 0$ for all $x \in C \setminus \{0\}$. It is *copositive plus* on a cone $C \subseteq \mathbb{R}^n$ if it is copositive on C and $\{x^T Mx = 0, x \in C\} \Rightarrow (M + M^T)x = 0$. When $C = \mathbb{R}^n_+$ then one simply says copositive. For instance $\left(\begin{smallmatrix} 1 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ is copositive on $\mathbb{R}^2_+$,

$$\begin{pmatrix} 2 & 2 & 1 & 2 \\ 3 & 3 & 2 & 3 \\ -2 & 1 & 5 & -2 \\ 1 & -2 & 1 & 2 \end{pmatrix}$$

is strictly copositive on $\mathbb{R}^4_+$.

[3] If A is an $m \times n$ matrix, I is a subset of $\{1, \ldots, m\}$ with k elements and J is a subset of $\{1, \ldots, n\}$ with k elements, then we write $[A]_{I,J}$ for the $k \times k$ minor of A that corresponds to the rows with index in I and the columns with index in J. If $I = J$, then $[A]_{I,J}$ is called a *principal minor*. They are sometimes called *subdeterminants*.

The study of copositive matrices is a hard topic, especially when copositivity on general convex sets is considered. One may simplify it in some cases. For instance if C is a closed convex polyhedral cone represented as $\{Gz \mid z \in \mathbb{R}^p_+\}$ where $G \in \mathbb{R}^{n\times p}$ has rank p, then copositivity of M on C is equivalent to the copositivity of $G^T MG$ on $\mathbb{R}^p_+$ (Hiriart-Urruty and Seeger 2010). There exists criteria to test the copositivity on positive orthant (cones of the form $\mathbb{R}^p_+$). Well-known results are the following ones:

Proposition 2.45 *Let $M = M^T \in \mathbb{R}^{2\times 2}$. Then M is copositive on $\mathbb{R}^2_+$ if and only if $a_{11} \geqslant 0$, $a_{22} \geqslant 0$, $a_{12} + \sqrt{a_{11}a_{22}} \geqslant 0$. Let $M = M^T \in \mathbb{R}^{3\times 3}$. Then M is copositive on $\mathbb{R}^3_+$ if and only if $a_{11} \geqslant 0$, $a_{22} \geqslant 0$, $a_{33} \geqslant 0$, $b_{12} \overset{\Delta}{=} a_{12} + \sqrt{a_{11}a_{22}} \geqslant 0$, $b_{13} \overset{\Delta}{=} a_{13} + \sqrt{a_{11}a_{33}} \geqslant 0$, $b_{23} \overset{\Delta}{=} a_{23} + \sqrt{a_{22}a_{33}} \geqslant 0$, and*

$$\sqrt{a_{11}a_{22}a_{33}} + a_{12}\sqrt{a_{33}} + a_{13}\sqrt{a_{22}} + a_{23}\sqrt{a_{11}} + \sqrt{2b_{12}b_{13}b_{23}} \geqslant 0.$$

See Hiriart-Urruty and Seeger (2010) for references and more results on copositive matrices, see also Goeleven and Brogliato (2004) for the first application in the field of Lyapunov stability of fixed points of evolution variational inequalities. The next proposition states some results on existence of solutions of complementarity problems with copositive matrices.

Proposition 2.46

(i) *Consider the LCP in (2.15). Suppose that M is copositive plus and that there exists an x^* satisfying $x^* \geqslant 0$ and $Mx^* + q \geqslant 0$. Then the LCP in (2.15) has a solution.*

(ii) *Consider the CCP in (2.18), with C a closed convex cone. Suppose that M is such that the homogeneous LCP $0 \leqslant x \perp Mx \geqslant 0$ has $x = 0$ as its unique solution. Then if M is copositive on C, the CCP in (2.18) has a non empty and bounded set of solutions.*

A matrix is P_0 if all its principal minors are non negative. For instance

$$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

are P_0. So a P_0-matrix is not necessarily positive semidefinite, however positive semidefinite matrices are P_0-matrices, and symmetric P_0-matrices are positive semidefinite. The following lemma holds (Lin and Wang 2002):

Lemma 2.47 *Let $M \in \mathbb{R}^{n\times n}$ be invertible. Then the following statements are equivalent:*

- *M is a P_0-matrix,*
- *M^T is a P_0-matrix,*
- *M^{-1} is a P_0-matrix.*

The P_0 property is not sufficient to guarantee the existence of solutions. Consider the LCP with $q = (-1\ 1)^T$, $M = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}$ that is a P_0-matrix. One may check by inspection that it has no solution. However the following is true.

Proposition 2.48 *Consider the LCP in* (2.15)*. Suppose that M is such that the homogeneous LCP $0 \leqslant x \perp Mx \geqslant 0$ has $x = 0$ as its unique solution. Then M is a P_0-matrix if and only if for all vectors q the LCP* (2.15) *has a connected solution set.*

This proposition does not state that the solution set is non empty, however. Therefore relaxing the P-property to the P_0-property destroys almost completely the powerful result of Theorem 2.43. It follows from Theorems 2.46 and 2.48 that the copositivity is much more useful than the P_0 property. We will, quite unfortunately, encounter P_0-matrices in nonsmooth electrical circuits!

Remark 2.49 A positive semidefinite matrix is copositive plus.

Other classes of matrices exist which guarantee under various conditions on q that the LCP (2.15) has solutions. We refer the reader to the above mentioned literature for more details on these classes.

Let us end this section by pointing out an important addendum to Theorem 2.43:

Proposition 2.50 *Let the matrix M be a P-matrix. Then the unique solution of the LCP in* (2.15) *is a piecewise-linear function of q, therefore Lipschitz continuous.*

This result is sometimes used to characterize the right-hand-side of some nonsmooth dynamical systems.

2.3.3 Links with Inclusions into Normal Cones

To see how things work, let us start with the complementarity conditions $0 \leqslant x \perp y \geqslant 0$ with x and y scalar numbers. Let us show that this is equivalent to the inclusion $-x \in N_C(y)$ with $C = \mathbb{R}_+$. Suppose x and y satisfy the inclusion. If $y > 0$ then $N_C(y) = \{0\}$ so that $x = 0$. If $y = 0$ then $N_C(y) = \mathbb{R}_-$ so $x \geqslant 0$. Now if $x > 0$ then $-x < 0$ and necessarily $y = 0$. Finally if $x = 0$ then y may be anywhere in $\mathbb{R}_+$. Consequently x and y satisfy $0 \leqslant x \perp y \geqslant 0$. Conversely let $0 \leqslant x \perp y \geqslant 0$. If $y > 0$ then $x = 0$. If $y = 0$ then $x \geqslant 0$ so that $-xz \leqslant xy$ for any $z \geqslant 0$. If $y > 0$ then $x = 0$ so that $xz = 0 \leqslant xy = 0$ for any $z \geqslant 0$. In any case the scalar $s \overset{\Delta}{=} -x$ satisfies $s(z - y) \leqslant 0$ for all $z \geqslant 0$, which precisely means that $s \in N_C(y)$, see Definition 2.18. We have shown that for $x \in \mathbb{R}$ and $y \in \mathbb{R}$

$$0 \leqslant x \perp y \geqslant 0 \quad \Leftrightarrow \quad -x \in N_C(y). \tag{2.23}$$

Obviously due to the symmetry of the problem we may replace the right-hand-side of (2.23) by $-y \in N_C(x)$. In fact the following is true, in a more general setting.

Proposition 2.51 *Let $C \subseteq \mathbb{R}^n$ be a non empty closed convex cone. Then*:

$$C \ni x \perp y \in C^\circ \quad \Leftrightarrow \quad y \in N_C(x). \tag{2.24}$$

We may also write CCPs with the dual cone C^* as $C \ni x \perp y \in C^* \Leftrightarrow -y \in N_C(x)$. The link with Fig. 2.14 is now clear. In this figure one has $-F(x_1) \in N_C(x_1)$ which is generated by the outwards normal vector to the right boundary of C. Also $N_C(x_2) = \{(0\,0)^T\}$ and one has $F(x_2) = (0\,0)^T$. If $y = -M(x-q)$ for some q and positive definite symmetric M then one may use Proposition 2.37 to calculate the solution x of the cone complementarity problem (2.24) as the projection of q in the metric defined by M on the cone C.

The link between the generalized equation (2.13) and the CCP is clear as well from Proposition 2.51. Finally let us see how to relate Propositions 2.37 and 2.50. Indeed one may easily deduce the following equivalences:

$$\begin{gathered} 0 \leqslant x \perp Mx + q \geqslant 0 \\ \Updownarrow \\ -Mx - q \in N_{\mathbb{R}^n_+}(x) \\ \Updownarrow \\ -x - M^{-1}q \in M^{-1}N_{\mathbb{R}^n_+}(x), \\ x = \mathrm{proj}_M(\mathbb{R}^n_+; -M^{-1}q) \end{gathered} \tag{2.25}$$

where the second equivalence is obtained under the assumption that $M = M^T > 0$. Since the projection operator is a single-valued Lipschitz continuous function, the result follows.

2.3.4 Links with Variational Inequalities

Let us start with a simple remark about the generalized equation (2.13) when C is convex. Using the definition of the normal cone in Definition 2.18, we may write equivalently:

$$\text{Find } x \in C \text{ such that:} \quad \langle F(x), y - x\rangle \geqslant 0 \quad \text{for all } y \in C \tag{2.26}$$

which is a variational formulation of the generalized equation. In fact (2.26) is a *variational inequality* (VI). In a more general setting, we have the following set of equivalences which extends Proposition 2.37. Let $\phi(\cdot)$ be a proper, convex lower semi-continuous function $\mathbb{R}^n \to \mathbb{R}$. Then for each $y \in \mathbb{R}^n$ there exists a unique $x \stackrel{\Delta}{=} P_\phi(y) \in \mathbb{R}^n$ such that

$$\langle x - y, v - x\rangle + \phi(v) - \phi(x) \geqslant 0, \quad \text{for all } v \in \mathbb{R}^n. \tag{2.27}$$

The mapping $P_\phi : \mathbb{R}^n \to \mathbb{R}^n$ is called the *proximation operator*. It is single-valued, non expansive and continuous. The next equivalences hold:

$$
\begin{gathered}
x \in \mathbb{R}^n: \quad \langle Mx+q, v-x\rangle + \phi(v) - \phi(x) \geqslant 0, \quad \text{for all } v \in \mathbb{R}^n \\
\Updownarrow \\
x \in \mathbb{R}^n: \quad x = P_\phi(x - (Mx+q)) \\
\Updownarrow \\
x \in \mathbb{R}^n: \quad Mx + q \in -\partial\phi(x).
\end{gathered}
\tag{2.28}
$$

The first formulation in (2.28) is called a VI of the second kind. Such variational inequalities are met in the study of static circuits (*i.e.* circuits with resistors and nonsmooth electronic devices) or in the study of the fixed points of dynamical circuits, see Addi et al. (2010). The link between (2.28) and (2.26) is done by setting $\phi(\cdot) = \psi_C(\cdot)$, the indicator function of C, and $F(x) = Mx + q$.

2.3.5 Links with Optimization

We have seen that there is a close link between inclusions into a normal cone (which are a special case of generalized equations) and optimization one side, and a close link between complementarity problems and inclusions into a normal cone on the other side. See (2.10) and Sect. 2.3.3 respectively. Consequently, there must exist a link between complementarity and optimization.

Let us consider the following optimization problem:

$$
\begin{aligned}
\text{Minimize} \quad & Q(x) = Cx + \frac{1}{2}x^T Dx \\
\text{subject to} \quad & Ax \geqslant b, \\
& x \geqslant 0,
\end{aligned}
\tag{2.29}
$$

where $D \in \mathbb{R}^{n\times n}$ is symmetric (if it is not, replace it by $D + D^T$ without modifying $Q(x)$). The so-called Karush-Kuhn-Tucker necessary conditions that have to be satisfied by any solution of (2.29) are:

$$
\begin{cases}
C^T + Dx - A^T y - u, \\
0 \leqslant y \perp Ax - b \geqslant 0, \\
0 \leqslant u \perp x \geqslant 0.
\end{cases}
\tag{2.30}
$$

Defining $\lambda = \binom{y}{u}$, $\tilde{A}^T = (A^T \; I_n)$, $\tilde{b} = \binom{b}{0}$, this may be rewritten more compactly as:

$$
\begin{cases}
C^T + Dx - \tilde{A}^T\lambda, \\
0 \leqslant \lambda \perp \tilde{A}x - \tilde{b} \geqslant 0.
\end{cases}
\tag{2.31}
$$

This is under the form of an MLCP, see (2.19). If the matrix D is invertible, one has $x = D^{-1}(-C^T + \tilde{A}^T\lambda)$ and we obtain:

$$
0 \leqslant \lambda \perp \tilde{A}D^{-1}(-C^T + \tilde{A}^T\lambda) - \tilde{b} \geqslant 0,
\tag{2.32}
$$

that is an LCP with matrix $M = \tilde{A}D^{-1}\tilde{A}^T$ and vector $q = -\tilde{A}D^{-1}C^T - \tilde{b}$. Conditions on A and D such that this LCP is well-posed may be studied.

2.4 Mathematical Formalisms

This section provides a quick overview of the definition and the well-posedness of various types of nonsmooth dynamical systems, and on the nature of their solutions (usually the solutions are at most $C^0[\mathbb{R}_+;\mathbb{R}^n]$, and they can contain jumps, or even Dirac measures or higher degree distributions). In view of the fact that complementarity problems, generalized equations, inclusions into normal cones, variational inequalities, possess strong links, it will not come as a surprise that their dynamical counterparts also are closely related. As we shall see later in this chapter and also in Chap. 4, the models of electrical circuits do not necessarily exactly fit within the mathematical formalisms below, in particular because in the simple circuits of Chap. 1, no algebraic equality appears. In more complex circuits the dynamical equations generation usually yields differential algebraic equations (DAE). Studying such “simplified” models is however a first mandatory step. For a more complete exposition of various nonsmooth models and formalisms we refer the reader to Part I of Acary and Brogliato (2008).

To start with, let us provide a general definition of what one calls a *differential inclusion*.

Definition 2.52 A differential inclusion may be defined by

$$\dot{x}(t) \in F(t, x(t)), \quad t \in [0, T],\ x(0) = x_0, \tag{2.33}$$

where $x : \mathbb{R} \to \mathbb{R}^n$ is a function of time t, $\dot{x} : \mathbb{R} \to \mathbb{R}^n$ is its time derivative, $F : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ is a set-valued map which associates to any point $x \in \mathbb{R}^n$ and time $t \in \mathbb{R}$ a set $F(t, x) \subset \mathbb{R}^n$, and $T > 0$.

In general the inclusion will be satisfied almost everywhere on $[0, T]$, because $x(\cdot)$ may not be differentiable for all $t \in [0, T]$. If $x(\cdot)$ is absolutely continuous then $\dot{x}(\cdot)$ is defined up to a set of Lebesgue measure zero on $[0, T]$. In fact it happens that there are several very different types of differential inclusions, depending on what the sets $F(x)$ look like.

Remark 2.53 One should not think that since the right-hand-side is multivalued then necessarily a differential inclusion has several solutions starting from a unique initial x_0. This depends a lot on the properties of $F(t, x)$, and many important classes of differential inclusions enjoy the property of uniqueness of solutions.

Definition 2.52 implicitly assumes that the solutions possess a certain regularity, for instance they are not discontinuous. When state jumps are present, one has to enlarge this definition to so-called *measure differential inclusions*. We shall not give a general definition of a measure differential inclusion (see Leine and van de Wouw 2008, Sect. 4.3 for this). In Sect. 2.4.1 important cases are presented. The literature on each of the class of nonsmooth dynamical systems presented below, is vast. Not all the references will be given, some classical or useful ones are provided, anyway.

2.4.1 Moreau's Sweeping Process, Measure Differential Inclusions

The sweeping process is a particular differential inclusion that has been introduced by Moreau (1971, 1972, 1973, 1977) in the context of unilateral mechanics. It has received considerable attention since then.

2.4.1.1 First Order Sweeping Process

The basic first order sweeping process as introduced by J.J. Moreau is a differential inclusion of the form

$$-\dot{x}(t) \in N_{C(t)}(x(t)), \quad \text{almost everywhere on } [0,T],\; x(0) = x_0 \in C(0), \tag{2.34}$$

where $C : [0,T] \to \mathbb{R}^n$ is a moving set. A function $x : [0,T] \to \mathbb{R}^n$ is a solution of (2.34) if:

- $x(t) \in C(t)$ for all $t \in [0,T]$,
- $x(\cdot)$ is differentiable at almost every point $t \in (0,T)$,
- $x(\cdot)$ satisfies the inclusion (2.34) for almost every $t \in (0,T)$.

An important extension is the *perturbed* sweeping process:

$$\begin{aligned} -\dot{x}(t) &\in N_{C(t)}(x(t) + f(t,x(t)), \\ &\text{almost everywhere on } [0,T],\; x(0) \in C(0). \end{aligned} \tag{2.35}$$

Remark 2.54 Why the name *sweeping process*? When $x(t) \in \text{Int}(C(t))$, where Int means the interior, then the normal cone $N_{C(t)}(x(t)) = \{0_n\}$, the zero vector of $\mathbb{R}^n$. The solution of (2.34) stays at rest, while the solution of (2.35) evolves according to an ordinary differential equation. When $x(t)$ lies on the boundary of $C(t)$, then the normal cone is not reduced to the zero vector, and the meaning of the inclusion is that there exists an element of $N_{C(t)}(x(t))$, call it $\gamma(t) \in \mathbb{R}^n$, such that the solution $x(\cdot)$ does not quit $C(\cdot)$ in a right neighborhood of t. If $C(\cdot)$ is moving then $x(\cdot)$ has the tendency to be swept by $C(\cdot)$. This is depicted in Fig. 2.15.

A basic existence and uniqueness of solutions result is the next one, that is simplified from Edmond and Thibault (2005, Theorem 1). The notions of absolutely continuous functions and sets may be found in Sect. A.1. Recall that $L^1([0,T],\mathbb{R})$ is the set of Lebesgue integrable functions such that $\int_a^b \|f(t)\|dt < +\infty$ for all $0 \leqslant a \leqslant b \leqslant T$.

Theorem 2.55 *Let $C(t)$ be for each t a non empty with non empty interior closed convex subset of $\mathbb{R}^n$, which varies in an absolutely continuous way. Suppose that*:

- *For every $\eta > 0$ there exists a non negative function $k_\eta(\cdot) \in L^1([0,T],\mathbb{R})$ such that for all $t \in [0,T]$ and for any $(x,y) \in B[0,\eta] \times B[0,\eta]$ one has*: $\|f(t,x) - f(t,y)\| \leqslant k_\eta(t)\;\|x - y\|$;

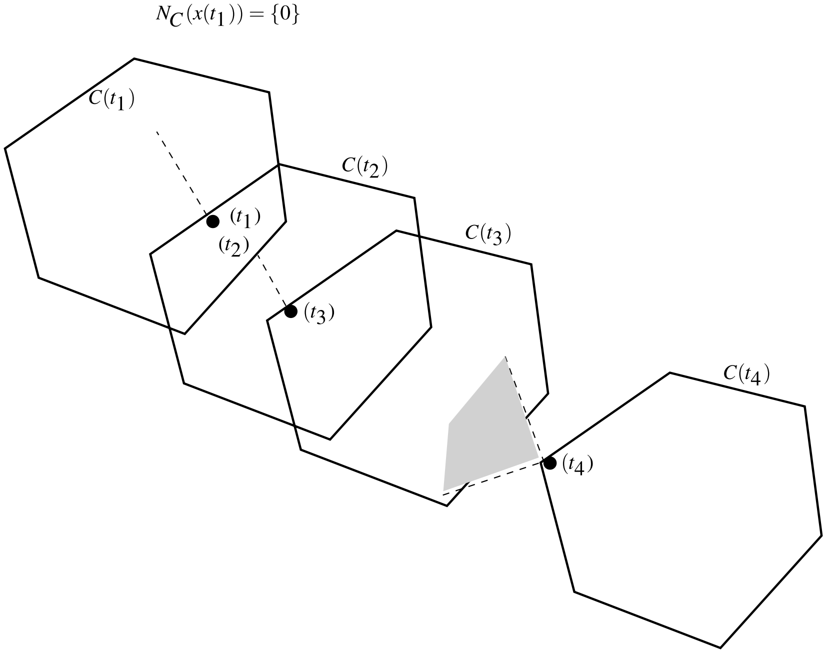


Fig. 2.15 A moving convex set $C(t)$ (the normal cones are depicted with *dashed lines*)

- *there exists a non negative function* $\beta(\cdot) \in L^1([0,T],\mathbb{R})$ *such that for all* $t \in [0,T]$ *and for any* $x \in \cup_{s\in[0,T]} C(s)$, *one has* $\|f(t,x)\| \leqslant \beta(t)(1+\|x\|)$.

Then for any $x_0 \in C(0)$ *the perturbed sweeping process in* (2.35) *has a unique absolutely continuous solution.*

Uniqueness is to be understood in the class of absolutely continuous functions. A quite similar result can be stated when $C(\cdot)$ is Lipschitz continuous in the Hausdorff distance. Then the solutions are Lipschitz continuous.[4] The next result is an existence result in the case where $C(t)$ may jump, and consequently the state $x(\cdot)$ may jump as well. One easily conceives that the inclusions in (2.34) and (2.35) have to be rewritten because at the times when $x(\cdot)$ jumps, its derivative is a Dirac measure. Then one has to resort to *measure differential inclusions* to treat in a proper way such systems. The relevant definitions can be found in Sects. A.3, A.4, A.5 and A.6. The next theorem is a simplified version of Edmond and Thibault (2006, Theorem 4.1).

Theorem 2.56 *Let* $C(t)$ *be for each* t *a non empty closed convex subset of* $\mathbb{R}^n$, *and let the set valued map* $C(\cdot)$ *be RCBV on* $[0,T]$.[5] *Suppose there exists some non*

[4] It is a fact that the solutions functional set is a copy of the multifunction $C(t)$ functional set.

[5] See Sect. A.4.

negative real β such that $\|f(t,x)\| \leqslant \beta(1+\|x\|)$ for all $(t,x) \in [0,T] \times \mathbb{R}^n$. Then for any $x_0 \in C(0)$ the perturbed sweeping process

$$-dx \in N_{C(t)}(x(t)) + f(t,x(t))d\lambda, \quad x(0) = x_0 \tag{2.36}$$

has at least one solution in the sense of Definition A.7.

The inclusion in (2.36) is a measure differential inclusion, see Sect. A.6 for an introduction to such evolution problems. λ is the Lebesgue measure (*i.e.* $d\lambda = dt$), dx is the differential measure associated with $x(\cdot)$. Roughly speaking, dx is the usual derivative outside the instants of jump, and it is a Dirac measure at the discontinuity times. This formalism may appear at first sight a mathematical fuss, however it is a rigorous way to represent such dynamical systems and naturally leads to powerful time-discretizations. In Edmond and Thibault (2006) it is considered a multivalued perturbation term in (2.36). In Moreau (1977) the perturbation is zero (this is the original version of the first order sweeping process) and uniqueness of solutions is proved. In Brogliato and Thibault (2010) the uniqueness of solutions is proved for both the absolutely continuous and the RCBV cases, when $f(t,x) = Ax + u(t)$. For an introduction to the sweeping process see Kunze and Monteiro Marquès (2000).

2.4.1.2 Second Order Sweeping Process

The second order sweeping process has been developed for Lagrangian mechanical systems subject to m unilateral constraints $f_i(q) \geqslant 0$, $1 \leqslant i \leqslant m$. However since some electrical circuits may be recast into the Lagrangian formalism (see Sect. 2.5.4), it is of interest to briefly recall it. The unilateral constraints define an *admissible domain* of the configuration space: $\Phi = \{q \in \mathbb{R}^n \mid f_i(q) \geqslant 0, 1 \leqslant i \leqslant m\}$, where q is the vector of generalized coordinates. In such systems the velocity may be discontinuous at the impact times, and the post-impact velocity is calculated as a function of the pre-impact one *via* a restitution law. Following similar steps as for the above measure differential inclusions, we may define the differential measure associated with the generalized acceleration, denoted as dv, where $v(\cdot)$ is almost everywhere equal to the generalized velocity $\dot{q}(\cdot)$. The original point is in the right-hand-side of the inclusion. If the constraints are perfect (no friction), then the contact reaction force R lies in the normal cone to Φ at q: $-R(t) \in N_\Phi(q(t))$. One would like, however, to go a step further: the measure differential inclusion should encapsulate the restitution law at impact times (more exactly, it should encapsulate *a particular* restitution law, since the choice of restitution laws is a modeling choice). J.J. Moreau has proposed to replace the inclusion $-R(t) \in N_\Phi(q(t))$ by the inclusion:

$$-R(t) \in N_{T_\Phi(q(t))}(w(t)) \tag{2.37}$$

which is the normal cone at $w(t) = \frac{v(t^+)+ev(t^-)}{1+e}$ to the tangent cone to Φ at $q(t)$, and $e \in [0,1]$ is a restitution coefficient. Let us now write the Lagrange measure differential inclusion:

$$-M(q(t))dv + F(q(t), v(t^+), t)dt \in N_{T_\Phi(q(t))}(w(t)) \tag{2.38}$$

where $F(q(t), v(t^+), t)$ accounts for the nonlinear and exogenous terms of the dynamics (Coriolis, centripetal forces, control inputs), and $M(q) = M^T(q)$ is positive definite. In order to analyze the differential inclusion in (2.38) we will use Proposition 2.37 and the material in Sects. A.5 and A.6. As we saw just above for the first order sweeping process with discontinuous state, at an impact time the velocity $v(\cdot)$ undergoes a discontinuity, and its differential measure is $dv = (v(t^+) - v(t^-))\delta_t + [\dot{v}(t)]dt + d\zeta_v$, see (A.3) in Sect. A.3. One has $dt(\{t\}) = 0$ and $d\zeta_v(\{t\}) = 0$ because these two measures are non atomic. From the interpretation of the inclusion of a measure in a convex cone we obtain:

$$-M(q(t))(v(t^+) - v(t^-)) \in N_{T_\Phi(q(t))}\left(\frac{v(t^+) + ev(t^-)}{1+e}\right). \tag{2.39}$$

Since the right-hand-side is a cone, we may multiply the left-hand-side by any non negative scalar and the inclusion remains true. Let us multiply it by $\frac{1}{1+e}$:

$$\begin{aligned} -M(q(t))&\left(\frac{v(t^+) - v(t^-) + ev(t^-) - ev(t^-)}{1+e}\right) \\ &\in N_{T_\Phi(q(t))}\left(\frac{v(t^+) + ev(t^-)}{1+e}\right). \end{aligned} \tag{2.40}$$

Using Proposition 2.37 we deduce that at an impact time t:

$$\frac{v(t^+) + ev(t^-)}{1+e} = \mathrm{proj}_{M(q(t))}(T_\Phi(q(t)); v(t^-)), \tag{2.41}$$

that is:

$$v(t^+) = -ev(t^-) + (1+e)\mathrm{proj}_{M(q(t))}(T_\Phi(q(t)); v(t^-)), \tag{2.42}$$

which is a generalized formulation of the well-known Newton's impact law between two frictionless rigid bodies. The advantage of Moreau's rule is that it provides in one shot the whole post-impact velocity vector. Also it is based on a geometrical analysis of the impact process which may serve as a basis for further investigations. It can be shown that Moreau's impact law is energetically consistent for $e \in [0, 1]$ (*i.e.* the kinetic energy decreases at impacts), and it guarantees that the post-impact velocity is admissible (*i.e.* it points inside Φ).

When $q(t) \in \mathrm{Int}(\Phi)$, then simple calculations show that $N_{T_\Phi(q(t))}(w(t)) = \{0_n\}$ since $T_\Phi(q(t)) = \mathbb{R}^n$. Thus the differential inclusion (2.38) is the smooth Lagrange dynamics. Notice that when $q(t)$ lies on the boundary of Φ, and if $v(t^-)$ belongs to the interior of $T_\Phi(q(t))$, then from (2.42) we get $\mathrm{proj}_{M(q(t))}(T_\Phi(q(t)); v(t^-)) = v(t^-)$ and $v(t^+) = v(t^-)$.

The well-posedness of the second order sweeping process has been studied in Monteiro Marques (1985, 1993), Mabrouk (1998), Dzonou et al. (2007), and Dzonou and Monteiro Marques (2007). The position $q(\cdot)$ is absolutely continuous, and the velocity $v(\cdot)$ is RCLBV. For non mathematical introductions to the Lagrangian sweeping process, see Acary and Brogliato (2008) and Brogliato (1999).

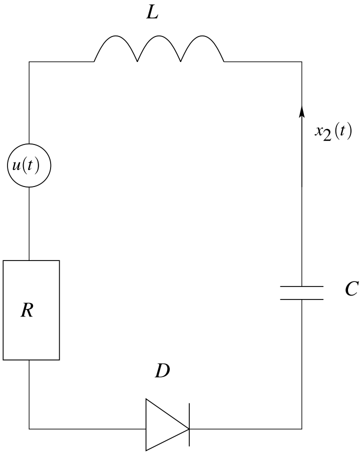


Fig. 2.16 An RLC circuit with a controlled voltage source

2.4.1.3 Higher Order Sweeping Process

The so-called higher-order sweeping process, defined and studied in Acary et al. (2008), is an extension of the above measure differential inclusion in cases where the solutions are not measures but distributions of larger degree. The interested reader may have a look at Acary et al. (2008) or at Acary and Brogliato (2008, Chaps. 5 and 11). Circuits with nonsmooth electronic devices may possess currents and/or voltages which are distributions (Dirac measure and its derivatives), provided the current and/or voltage sources are controlled by internal variables. It is known in circuits theory modeled by differential-algebraic equations (DAE) that such internally-controlled sources may increase the index of the system. This is directly linked to the relative degree of the complementarity variables. Clearly in the case of circuits made of dissipative elements, getting solutions that contain distributions of degree strictly larger than 2 (*i.e.* derivatives of Dirac measures) is possible only with controlled sources. Let us provide an example, with the RLCD circuit depicted in Fig. 2.16. Let us assume that the voltage $u(\cdot)$ is a dynamic feedback of the "output" the voltage across the diode, $\lambda(t)$:

$$\begin{cases} u = \lambda + Lx_3, \\ \dot{x}_3(t) = x_4(t), \\ \dot{x}_4(t) = \lambda(t). \end{cases} \tag{2.43}$$

Inserting this control input inside the circuit's dynamics, one obtains:

$$\begin{cases} \dot{x}_1(t) = x_2(t), \\ \dot{x}_2(t) = -\frac{R}{LC}x_2(t) - \frac{1}{RC}x_1(t) + x_3(t), \\ \dot{x}_3(t) = x_4(t), \\ \dot{x}_4(t) = \lambda(t), \\ 0 \leqslant \lambda(t) \perp w(t) = -x_2(t) \geqslant 0. \end{cases} \tag{2.44}$$

This dynamics is written under the form of a linear complementarity system (see (2.53) below). It is easily calculated that $D = CB = CAB = 0$ while $CA^2B = 1$, so that the relative degree between λ and w is equal to 3. The dynamical system as it is written in (2.44) is not complete, in the sense that one cannot perform its time-integration without adding supplementary modeling informations. In fact, it is missing in (2.44) a state re-initialization rule which enables one to compute a state jump when the admissible domain boundary $x_2 = 0$ is attained. In Acary et al. (2008) a complete framework is proposed that enables one to give a rigorous meaning to the dynamics in (2.44), together with a time-stepping method and some preliminary convergence results. Such a dynamical system is then embedded into a differential inclusion whose solutions are Schwartz' distributions, and which is an extension of (2.38). The state jumps are automatically taken into account in the formulation. Then the system can be integrated in time and the domain $\{x \in \mathbb{R}^4 \mid x_2 \leqslant 0\}$ is an invariant subset of the state space.

To summarize, Moreau's sweeping processes are particular differential inclusions into normal cones to moving sets. It was originally introduced in the field of Mechanics. Electrical circuits with nonsmooth electronic devices have recently been recast into sweeping processes, which facilitates their analysis.

2.4.2 *Dynamical Variational Inequalities*

Dynamical variational inequalities (DVI) are evolution problems of the form:

$$\begin{cases} x(t) \in \mathrm{dom}(\varphi) & \text{for all } t \geqslant 0, \\ \langle \dot{x}(t) + f(x(t),t), v - x(t) \rangle + \varphi(v) - \varphi(x(t)) \geqslant 0 & \text{for all } v \in \mathbb{R}^n, \end{cases} \tag{2.45}$$

for some convex, proper and lower semi-continuous function $\varphi : \mathbb{R}^n \to \mathbb{R}$. The DVI in (2.45) may be named a VI of the second kind. Let us choose $\varphi(\cdot) = \psi_C(\cdot)$ for some non empty, closed convex set $C \in \mathbb{R}^n$. Then we obtain:

$$\begin{cases} x(t) \in C & \text{for all } t \geqslant 0, \\ \langle \dot{x}(t) + f(x(t),t), v - x(t) \rangle \geqslant 0 & \text{for all } v \in C, \end{cases} \tag{2.46}$$

which is a DVI of the first kind. From (2.6) it easily follows that $-\dot{x}(t) - f(x(t),t)$ is a subgradient of $\varphi(\cdot)$ at $x(t)$. We may therefore rewrite (2.45) equivalently as:

$$\begin{cases} x(t) \in \mathrm{dom}(\varphi) \quad \text{for all } t \geqslant 0, \\ \dot{x}(t) + f(x(t),t) \in -\partial\varphi(x(t)), \end{cases} \tag{2.47}$$

which is a differential inclusion. If $\varphi(\cdot) = \psi_C(\cdot)$, the indicator function of the set C, then $\partial\varphi(x) = N_C(x)$, the normal cone to C at x. Then the DVI (2.45) is an inclusion into a normal cone. Suppose now that $\varphi(\cdot) = \psi_{C(t,x)}(\cdot)$, *i.e.* the set C may depend on t and x. Then we obtain:

$$\begin{cases} x(t) \in C(t, x(t)) & \text{for all } t \geqslant 0, \\ \langle \dot{x}(t) + f(x(t),t), v - x(t) \rangle \geqslant 0 & \text{for all } v \in C(t, x(t)), \end{cases} \tag{2.48}$$

which is a quasi DVI. Moreau's sweeping process is one particular type of a QDVI with $C(x) = T_{\Phi(q)}$ (the tangent cone to the admissible domain of the configuration space), see (2.37) and (2.38). A well-known result for the existence and uniqueness of solutions of DVIs is Kato's Theorem (Kato 1968). Let us present one extension of Kato's theorem. Let us introduce the following class of differential inclusions, where $x(t) \in \mathbb{R}^n$:

$$\begin{cases} \dot{x}(t) \in -A(x(t)) + f(t, x(t)), & \text{a.e. on } (0, T), \\ x(0) = x_0. \end{cases} \tag{2.49}$$

The following assumption is made:

Assumption 2.57 *The following items hold*:

(i) $A(\cdot)$ *is a multivalued maximal monotone operator from* $\mathbb{R}^n$ *into* $\mathbb{R}^n$, *with domain* $\mathrm{dom}(A)$, *i.e.*, *for all* $x \in \mathrm{dom}(A)$, $y \in \mathrm{dom}(A)$ *and all* $x' \in A(x)$, $y' \in A(y)$, *one has*

$$(x' - y')^T (x - y) \geqslant 0. \tag{2.50}$$

(ii) *There exists* $L \geqslant 0$ *such that for all* $t \in [0, T]$, *for all* $x_1, x_2 \in \mathbb{R}^n$, *one has* $\|f(t, x_1) - f(t, x_2)\| \leqslant L\|x_1 - x_2\|$.

(iii) *There exists a function* $\Phi(\cdot)$ *such that for all* $R \geqslant 0$:

$$\Phi(R) = \sup\left\{ \left\| \frac{\partial f}{\partial t}(\cdot, v) \right\|_{\mathscr{L}^2((0,T);\mathbb{R}^n)} \,\middle|\, \| v \|_{\mathscr{L}^2((0,T);\mathbb{R}^n)} \leqslant R \right\} < +\infty.$$

The following is proved in Bastien and Schatzman (2002).

Proposition 2.58 *Let Assumption* 2.57 *hold*, *and let* $x_0 \in \mathrm{dom}(A)$. *Then the differential inclusion* (2.49) *has a unique solution* $x : (0, T) \to \mathbb{R}^n$ *that is Lipschitz continuous with essentially bounded derivatives.*

It suffices to recall that the subdifferential of a convex proper lower semicontinuous function $\varphi(\cdot)$ defines a maximal monotone mapping (see Theorem 2.34), to conclude about the well-posedness of the DVI in (2.45) using Proposition 2.58.

2.4.3 Complementarity Dynamical Systems

Just as there are many kinds of complementarity problems, there are many kinds of complementarity systems, *i.e.* systems that couple an ordinary differential equation to a set of complementarity conditions between two slack variables. The circuits whose dynamics are in (1.3), (1.16), and (1.38) are particular complementarity systems.

2.4.3.1 Some Classes of Complementarity Systems

Let us give a very general complementarity formalism as follows:

$$\begin{cases} G(\dot{x}(t), x(t), t, \lambda) = 0, \\ \mathbf{C}^* \ni \lambda \perp w(t) \in \mathbf{C}, \\ F(x(t), t, \lambda, w(t)) = 0, \end{cases} \tag{2.51}$$

where $\mathbf{C} \subseteq \mathbb{R}^m$ is a closed convex cone, $\mathbf{C}^*$ is its dual cone, $\lambda \in \mathbb{R}^m$ may be interpreted as a Lagrange multiplier, $x(t) \in \mathbb{R}^n$, $F(\cdot)$ and $G(\cdot)$ are some functions. The variables λ and w form a pair of slack variables. Such a formalism is by far too general to be analyzed efficiently (and to be subsequently simulated efficiently!). One has to split the class of dynamical systems in (2.51) into more structured subclasses. Some examples are given now.

Definition 2.59 (Dynamical Complementarity Systems) A dynamical complementarity system (DCS) in an explicit form is defined by:

$$\begin{cases} \dot{x}(t) = f(x(t), t, \lambda(t)), \\ w(t) = h(x(t), \lambda(t)), \\ 0 \leqslant w(t) \perp \lambda(t) \geqslant 0. \end{cases} \tag{2.52}$$

If the smooth dynamics and the input/output function are linear, we speak of linear complementarity systems.

Definition 2.60 (Linear Complementarity Systems) A linear complementarity system (LCS) is defined by:

$$\begin{cases} \dot{x}(t) = Ax(t) + B\lambda(t), \\ w(t) = Cx(t) + D\lambda(t), \\ 0 \leqslant w(t) \perp \lambda(t) \geqslant 0. \end{cases} \tag{2.53}$$

When the functions $F(\cdot)$ and $G(\cdot)$ are linear and the cone $\mathbf{C}$ is a non negative orthant one gets:

Definition 2.61 (Mixed Linear Complementarity Systems) A mixed linear complementarity system (MLCS) is defined by:

$$\begin{cases} E\dot{x}(t) = Ax(t) + B\lambda(t) + F, \\ Mw(t) = Cx(t) + D\lambda(t) + G, \\ 0 \leqslant w(t) \perp \lambda(t) \geqslant 0. \end{cases} \tag{2.54}$$

If both the matrices E and M are square full rank and $E = F = 0$, we are back to an LCS as in (2.53). See for instance Example 7 in Brogliato (2003) for a system that fits within MLCS. One may also call such systems *descriptor variable complementarity systems*. As shown in Brogliato (2003) many systems with piecewise-linear characteristics may be recast into (2.54).

Remark 2.62 It is not clear whether or not the variable x in (2.54) should be called the *state* of the MLCS. Indeed if E is not full rank then some of the components of

x do not vary and satisfy only an algebraic constraint. As such they cannot be called a state variable. Some examples are given in Chap. 7, Sects. 7.2 and 7.3.

Definition 2.63 (Nonlinear Complementarity Systems) A nonlinear complementarity system (NLCS) is defined by:

$$\begin{cases} \dot{x}(t) = f(x(t), t) + g(x(t))\lambda(t), \\ w(t) = h(x(t), \lambda(t)), \\ 0 \leqslant w(t) \perp \lambda(t) \geqslant 0. \end{cases} \tag{2.55}$$

If $g(x) = -\nabla h(x)$, one obtains so-called gradient type complementarity systems which are defined as follows:

Definition 2.64 (Gradient Complementarity System) A gradient complementarity system (GCS) is defined by:

$$\begin{cases} \dot{x}(t) + f(x(t)) = \nabla g(x(t))\lambda(t), \\ w(t) = g(x(t)), \\ 0 \leqslant w(t) \perp \lambda(t) \geqslant 0. \end{cases} \tag{2.56}$$

The above complementarity systems are autonomous, without explicit dependence on time. Obviously one may define non autonomous CS, with exogenous inputs. For instance the non autonomous LCS dynamics is:

$$\begin{cases} \dot{x}(t) = Ax(t) + B\lambda(t) + Eu(t), \\ w(t) = Cx(t) + D\lambda(t) + Fu(t), \\ 0 \leqslant w(t) \perp \lambda(t) \geqslant 0. \end{cases} \tag{2.57}$$

More details on the definitions and the mathematical properties of CS can be found in Camlibel et al. (2002b), Camlibel (2001), van der Schaft and Schumacher (1998), Shen and Pang (2007), Heemels and Brogliato (2003), Brogliato (2003), and Brogliato and Thibault (2010). Roughly speaking, a lot depends on the *relative degree* between the two complementarity variables w and λ. The relative degree is the number of times one needs to differentiate the "output" w along the dynamics in order to recover the "input" λ. As an example let us consider the following scalar LCS:

$$\begin{cases} \dot{x}(t) = x(t) + \lambda, \\ 0 \leqslant \lambda \perp w(t) = x(t) \geqslant 0. \end{cases} \tag{2.58}$$

Then $\dot{w}(t) = \dot{x}(t) = \lambda(t)$ so that the relative degree is $r = 1$. If now $w(t) = x(t) + \lambda(t)$ then $r = 0$. Most of the results on existence and uniqueness of solutions to complementarity systems hold for relative degrees 0 or 1, in which case only measures appear in the dynamics. When $r \geqslant 2$ distributional solutions have to be considered, see Acary et al. (2008) where such LCS are embedded into the higher order sweeping process. The well-posedness of (2.57) has been shown in Camlibel et al. (2002b) when (A, B, C, D) defines a dissipative system (see Brogliato et al. 2007 for a definition). Local existence and uniqueness results are presented in van der Schaft and Schumacher (1998) for (2.55). Global existence and uniqueness of RCLBV (with state jumps) and absolutely continuous solutions is shown for LCS

(2.57) and NLCS (2.55) in Brogliato and Thibault (2010), under an "input-output" constraint. In Acary et al. (2008) LCS with high relative degree have been embedded into the so-called higher order sweeping process, that is a differential inclusion whose solutions are distributions.[6]

In the field of electrical circuits, one shall often encounter systems of the type (2.54) with a singular matrix E. From the material of Chap. 1 and the analysis of the dynamics of the circuits in Fig. 1.10, one easily guesses that the dynamics in (2.51) through (2.57) are not complete: a state reinitialization rule is missing (this was already pointed out in (1.61)). See Sect. 2.4.3.2 for more details on jump rules in a complementarity setting.

2.4.3.2 State Jump Laws

State jump rules are a well-known and widely studied topic in nonsmooth mechanics, where they correspond to velocity discontinuities created by impacts between rigid bodies. The realm of impact dynamics in nonsmooth mechanics is vast, and it has its counterpart in nonsmooth circuits. It is apparent from most of the examples which are analyzed in Chap. 1, that we may in a first instance write the dynamics of the presented circuits as:

$$\begin{cases} \dot{x}(t) = Ax(t) + B\lambda(t) + Eu(t), \\ w(t) = Cx(t) + D\lambda(t) + Fu(t), \\ 0 \leqslant w(t) \perp \lambda(t) \geqslant 0 \end{cases} \tag{2.59}$$

for some matrices A, B, C, D, E and F of appropriate dimensions. The state is $x(t)$, the external excitation is $u(t)$ (it may be voltage sources or current sources). Let us analyze intuitively the necessity for state jumps (see also the analysis we made for the circuit in (1.16)). Suppose for instance that $D = 0$, and that $w(t) = 0$ for some t. Assume that at t, $u(\cdot)$ jumps from a value $u(t^-)$ such that $Cx(t^-) + Fu(t^-) = 0$ to a value $u(t^+)$ such that $Cx(t^-) + Fu(t^+) < 0$. In order to respect the model dynamics the state has to jump to a value such that $Cx(t^+) + Fu(t^+) > 0$. If such a right-limit does not exist, we may conclude that the model is not well-posed and should be changed. The necessity for state jumps may also arise in some circuits with ideal switches, from topology changes. When the switch is ON, the dynamics is a certain differential-algebraic equation (DAE). When the switch is OFF, it becomes another DAE. However the value of the state just before the switch, may not be admissible initial data for the DAE just after the switch. It is well-known that a DAE with inconsistent initial data, has a solution that may be a distribution (Dirac and derivatives of Dirac).

State jumps have been introduced in Sect. 1.1.5 for the circuit in Fig. 1.10. There the numerical method in (1.17) suggested the jump law in (1.20) (equivalently (1.26) and (1.27)). In particular the form (1.27) is a quadratic program, hence an attractive

[6] This seems to be the very first instance of a distribution differential inclusion, with a complete analysis and a numerical scheme.

formulation from a numerical point of view. Before stating the state jump law presented in Frasca et al. (2007, 2008) and Heemels et al. (2003) (below the formulation is different from the one in these papers, and rather follows from convex analysis arguments as in Brogliato and Thibault (2010, Remark 2), we need some preparatory material. The quadruple (A, B, C, D) is said to be passive if the linear matrix inequality:

$$\begin{pmatrix} -A^T P + PA & -PB + C^T \\ -B^T P + C & D + D^T \end{pmatrix} \geqslant 0 \quad \text{and} \quad P = P^T > 0 \tag{2.60}$$

has a solution P. The quadratic function $V(x) = \frac{1}{2}x^T Px$ is then a so-called storage function of the system $\dot{x}(t) = Ax(t) + B\lambda(t)$, $w(t) = Cx(t) + D\lambda(t)$, with supply rate $\omega(\lambda, w) = \lambda^T w$. The linear matrix inequality in (2.60) is then equivalent to the dissipation inequality:

$$V(x(t)) - V(x(0)) \leqslant \int_0^t \omega(w(s), \lambda(s))ds \quad \text{for any } t \geqslant 0. \tag{2.61}$$

Let us define the set $K = \{z \in \mathbb{R}^n \mid Cz + Fu(t^+) \in Q_D\}$, with $Q_D = \{z \in \mathbb{R}^m \mid z \geqslant 0, Dz \geqslant 0, z^T Dz = 0\}$. Q_D^* and K^* are their dual cones. If $D = 0$ then $Q_D = \mathbb{R}^m_+ = Q_D^*$.

Proposition 2.65 *Let us consider the LCS in* (2.59), *and suppose that* (A, B, C, D) *is passive with storage function* $V(x) = \frac{1}{2}x^T Px$, $P = P^T > 0$. *Suppose a jump occurs in* $x(\cdot)$ *at time* t, *so that* $x(t^+) = x(t^-) + Bp_t$ *where* $\lambda = p_t\delta_t$. *Suppose that* F *and* C *are such that* $Fu(t) \in Q_D^* + \mathrm{Im}(C)$. *For any* $x(t^-)$ *there is a unique solution to*:

$$x(t^+) = \underset{x \in K}{\operatorname{argmin}} \frac{1}{2}(x - x(t^-))^T P(x - x(t^-)) \tag{2.62}$$

that is equivalent to:

$$P(x(t^+) - x(t^-)) \in -N_K(x(t^+)) \tag{2.63}$$

and to

$$K \ni x(t^+) \perp P(x(t^+) - x(t^-)) \in K^*. \tag{2.64}$$

Then the post-jump state $x(t^+)$ *is consistent with the complementarity system's dynamics on the right of* t.

The equivalences are a consequence of Propositions 2.37 and 2.51. The condition $Fu(t) \in Q_D^* + \mathrm{Im}(C)$ is a sort of constraint qualification condition, which guarantees that the LCP $0 \leqslant \lambda \perp Cx + Fu + D\lambda \geqslant 0$ has a solution (see Sect. 5.2.2 for a similar condition, stated in a different context). Notice that we have implicitly assumed that λ is a measure, which indeed is the case. Recall also that the LCS in (2.59) can be interpreted, by splitting y into its components satisfying $w_i(t^+) > 0$ and those satisfying $w_j(t^+) = 0$, as a DAE. Such a DAE corresponds to what one may call a mode of the system. Consistency of $x(t^+)$ means consistency with respect to this DAE. In other words, the state jump rule does not only have a physical motivation but also guarantees that the system is coherent once $x(\cdot)$ has jumped to a new value, in the sense that there is a unique mode of the LCS such that the resulting DAE has $x(t^+)$ as its consistent initial state.

We note that $B^T PBp_t = B^T P(x(t^+) - x(t^-))$. If $B \in \mathbb{R}^{n\times m}$ has full rank m (which in particular implies that $m < n$) then the multiplier magnitude at t is given uniquely by $p_t = (B^T PB)^{-1}B^T P(x(t^+) - x(t^-))$.

Remark 2.66 This way of modeling and formulating state jump rules for electrical circuits with nonsmooth electronic devices, is inspired from J.J. Moreau's framework of unilateral mechanics, see Sect. 2.4.1 and *e.g.* Brogliato (1999, pp. 199–200). Notice that (2.63) means that $x(t^+)$ is the projection of $x(t^-)$ onto K in the metric defined by the matrix P. Compare with (2.42) with $e = 0$. In Frasca et al. (2008) the state jumps in electrical circuits are given a physical meaning in terms of charge/flux conservation. It is noteworthy that Proposition 2.65 does not apply to the controlled circuit (2.44) which has to be embedded into the higher order sweeping process.

Let D have full rank m. Then $Q_D = \{0\}$, $Q_D^* = \mathbb{R}^m$, $K = \mathbb{R}^n$ and $K^* = \{0\}$. Therefore from Proposition 2.65 $x(t^+) = x(t^-)$: there is no state jumps, and the trajectories are continuous functions of time. This is quite consistent with the observation that when D is a P-matrix, then the complementarity conditions of the LCS define an LCP that has a unique solution λ^* whatever $u(t)$ and $x(t)$. Moreover this λ^* is a Lipschitz function of u and x. Consequently the LCS in (2.57) is an ordinary differential equation with a Lipschitz continuous right-hand-side, and with $C^1(\mathbb{R}^+; \mathbb{R}^n)$ solutions.

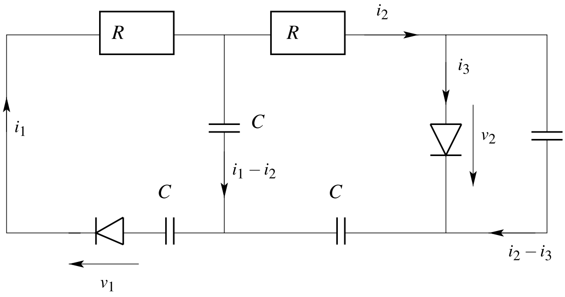
When $D = 0$, then one has $Q_D = \mathbb{R}^m_+$, $Q_D^* = \{0\}$, $K = \{z \in \mathbb{R}^n \mid Cz + Fu(t^+) \geqslant 0\}$. Then a state jump may occur depending on the value of $u(t^+)$ (see Sect. 5.2 for further comments on state jumps).

> Complementarity dynamical systems constitute a large class of nonsmooth systems. Existence and uniqueness of global solutions have been shown in particular cases only. Simple electrical circuits with nonsmooth electronic devices like ideal diodes are modeled with linear complementarity systems. They undergo state jumps which may be justified from physical energetical arguments, similarly to restitution laws of mechanics.

2.4.3.3 Examples

Let us end this section on complementarity dynamical systems by providing further illustrating examples (several examples have already been presented in the foregoing chapter). Let us consider the electrical circuit in Fig. 2.17 that is composed of two resistors R with voltage/current law $u(t) = Ri(t)$, four capacitors

Fig. 2.17 Electrical circuit with capacitors, resistors and ideal diodes



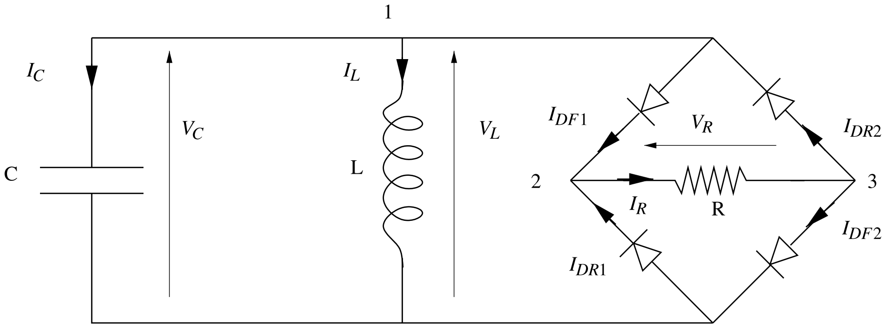


Fig. 2.18 A 4-diode bridge wave rectifier

C with voltage/current law $C\dot{u}(t) = i(t)$, and two ideal diodes with characteristics $0 \leqslant v_1(t) \perp i_1(t) \geqslant 0$ and $0 \leqslant v_2(t) \perp i_3(t) \geqslant 0$ respectively. The state variables are $x_1(t) = \int_0^t i_1(t)dt$, $x_2(t) = \int_0^t i_2(t)dt$, $x_3(t) = v_2(t)$, and $\lambda_1(t) = -i_3(t)$, $\lambda_2(t) = v_1(t)$.

The dynamics of this circuit is given by:

$$\begin{pmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \\ \dot{x}_3(t) \end{pmatrix} = \begin{pmatrix} \frac{-2}{RC} & \frac{1}{C} & 0 \\ \frac{1}{C} & \frac{-2}{RC} & 1 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{pmatrix} + \begin{pmatrix} 0 & \frac{1}{R} \\ 0 & 0 \\ \frac{1}{C} & 0 \end{pmatrix} \lambda(t),$$
$$0 \leqslant \lambda(t) \perp w(t) = \begin{pmatrix} 0 & 0 & 1 \\ \frac{-2}{RC} & \frac{1}{RC} & 0 \end{pmatrix} \begin{pmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & \frac{1}{R} \end{pmatrix} \lambda(t) \geqslant 0. \tag{2.65}$$

The matrices A, B, C and D in (2.53) are easily identified. It is noteworthy that the feedthrough matrix D is positive semi-definite only.

Let us consider the four-diode bridge wave rectifier in Fig. 2.18, with a capacitor $\mathbf{C} > 0$, an inductor $\mathbf{L} > 0$, a resistor $\mathbf{R} > 0$. Its dynamics is given by:

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} 0 & -\frac{1}{\mathbf{C}} \\ \frac{1}{\mathbf{L}} & 0 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 0 & 0 & -\frac{1}{\mathbf{C}} & \frac{1}{\mathbf{C}} \\ 0 & 0 & 0 & 0 \end{bmatrix} \lambda(t), \tag{2.66}$$
$$0 \leqslant w(t) \perp \lambda(t) \geqslant 0,$$

where $x_1 = v_L$, $x_2 = i_L$, $\lambda = (-v_{DR1}\ \ -v_{DF2}\ i_{DF1}\ i_{DR2})^T$, $y = (i_{DR1}\ i_{DF2}\ \ -v_{DF1}\ \ -v_{DR2})^T$ and

$$w = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ -1 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} \frac{1}{\mathbf{R}} & \frac{1}{\mathbf{R}} & -1 & 0 \\ \frac{1}{\mathbf{R}} & \frac{1}{\mathbf{R}} & 0 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \lambda. \tag{2.67}$$

Notice that in this example the dimension of the state vector is 2 while the dimension of the LCP variables is 4 (in a Systems and Control language, the “input” has a larger dimension than the state). The matrix D is a full rank, positive semi-definite matrix. As a second example of a diode bridge, let us consider the circuit obtained from the circuit of Fig. 2.18 by dropping the capacitor and the inductance outside the bridge, and adding a capacitor **C** in parallel with the resistor inside the bridge. The state x is the voltage across the capacitor. We assume that each diode has a current/voltage law of the form $V_k \in -\partial\varphi_k(i_k)$, $k = 1, 2, 3, 4$, for some convex, proper lower semi-continuous functions $\varphi_k(\cdot)$. The material of Sect. 2.3.3 together with Example 2.26 should help the reader to find that if $\varphi_k(\cdot) = \psi_K(\cdot)$ for some convex set K, then the diode k possesses a complementarity formulation of its current/voltage law. The dynamics of this circuit is given by:

$$\begin{cases} \dot{x}(t) = -\frac{1}{\mathbf{RC}} x(t) + \begin{pmatrix} \frac{1}{\mathbf{C}} & 0 & \frac{1}{\mathbf{C}} & 0 \end{pmatrix} \lambda(t), \\ w(t) = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} x(t) + \begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 1 & -1 \\ 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \lambda(t), \end{cases} \tag{2.68}$$

with $w_1 = V_{DR1}$, $w_2 = i_{DF2}$, $w_3 = V_{DF1}$, $w_4 = V_{DR2}$, and $\lambda = (i_{DR1}\ V_{DF2}\ i_{DF1}\ i_{DR2})^T$. The matrix D has rank 2, it is positive semi-definite since it is skew symmetric.

These three examples show that electrical circuits may yield LCS as in (2.53) with matrices D that may be positive semi-definite with full rank, skew symmetric, or positive semi-definite with low rank. The fact that the D matrix, which is the system’s LCP matrix, may be non symmetric, is a strong feature of electrical circuits with ideal diodes.

In Chap. 7 we will study other examples that yield MLCS as in (2.54).

2.4.4 Filippov’s Inclusions

Filippov’s inclusions are closely linked to so-called variable structure systems, or switching systems. The study of such systems started in the fifties in the former

USSR, and is still a very active field of research in control theory, because of the efficiency of sliding mode controllers (Yu and Kaynak 2009; Utkin et al. 2009). Let us start from a switching system of the form:

$$\dot{x}(t) = A_i x(t) + a_i(t) \quad \text{if } x(t) \in \chi_i,\ i \in \mathscr{I}_1,\ x(0) = x_0 \in \mathbb{R}^n \tag{2.69}$$

for constant matrices A_i and time-functions $a_i(t)$, and a partitioning of $\mathbb{R}^n$ in polyhedral sets χ_i is defined:

(i) the sets χ_i are finitely represented as $\chi_i = \{x \in \mathbb{R}^n \mid C_i x + D_i \geqslant 0\}$, $C_i \in \mathbb{R}^{m_i \times n}$, $D_i \in \mathbb{R}^{m_i \times 1}$,
(ii) $\bigcup_{i=1}^{m} \chi_i = \mathbb{R}^n$,
(iii) for all $i \neq j$, $(\chi_i \setminus \partial\chi_i) \cap (\chi_j \setminus \partial\chi_j) = \emptyset$,
(iv) the sets χ_i have an nonempty interior.

Conditions (iii) and (iv) imply that $\chi_i \cap \mathrm{Int}(\chi_j) = \emptyset$ for all $i \neq j$. We denote the set of indices of the partition as $\mathscr{I}_1$, i.e. the set of polyhedra is $\{\chi_i\}_{i \in \mathscr{I}_1}$. Obviously $\mathscr{I}_1$ may be finite, or infinite. The properties (ii) and (iii) mean that the polyhedral sets χ_i cover $\mathbb{R}^n$, and their interiors are disjoint: only their boundary may be common with the boundary of other sets. The dynamics in (2.69) defines a *polyhedral switching affine system*. We may write compactly the system (2.69) as $\dot{x}(t) = f(x(t), t)$ for some function $f(\cdot, \cdot)$ that is constructed from the vector fields $f_i(x, t) = A_i x + a_i(t)$. It is clear that unless some conditions are imposed on the boundaries $\partial\chi_i$, the vector field $f(\cdot, \cdot)$ is discontinuous on $\partial\chi_i$. The simplest example is when $f_i(x) = a_i$, $f_j(x) = a_j$, $i \neq j$, and $a_i \neq a_j$. Then three situations may occur when a solution reaches a boundary between two cells χ_i and χ_j: (i) the trajectory crosses the switching surface $\partial\chi_i$ (that coincides with $\partial\chi_j$ at the considered point in the state space), (ii) the trajectory remains on the boundary and then evolves on it (this is called a *sliding motion*, (iii) there are several possible future trajectories: one that stays on the boundary, and others that leave it (this is called a *spontaneous jump* in the solution derivative).

2.4.4.1 Simple Examples

The simplest cases that enable one to clearly see this are the scalar switching systems:

$$\dot{x}(t) = g(t) + \begin{cases} 1 & \text{if } x < 0, \\ -1 & \text{if } x > 0, \end{cases} \tag{2.70}$$

$$\dot{x}(t) = g(t) + \begin{cases} -1 & \text{if } x < 0, \\ 1 & \text{if } x > 0, \end{cases} \tag{2.71}$$

$$\dot{x}(t) = g(t) + \begin{cases} 1 & \text{if } x < 0, \\ 1 & \text{if } x > 0, \end{cases} \tag{2.72}$$

with $x(0) \in \mathbb{R}$ and $|g(t)| \leqslant \frac{1}{2}$ for all $t \geqslant 0$, where $g(\cdot)$ is a continuous function of time (for instance $g(t) = \frac{1}{2}\sin(t)$). In (2.70)–(2.72) we intentionally ignored the value of the discontinuous vector field $f(x, t)$ at $x = 0$. It is easy to see that:

- In case (2.70) all trajectories with $x(0) \neq 0$ converge (in finite time) to the "surface" $x = 0$.
- In case (2.71) all trajectories starting with $x(0) < 0$ diverge to $-\infty$, all trajectories with $x(0) > 0$ diverge to $+\infty$.
- In case (2.72) all trajectories starting with $x(0) < 0$ reach $x = 0$ in finite time; all trajectories starting with $x(0) > 0$ diverge to $+\infty$.

In all three cases we are not yet able to determine what happens on the "surface" $x = 0$. The solution proposed by Filippov is to embed these systems into a class of differential inclusions, whose right-hand-side is the closed convex hull of the vector fields at a discontinuity, disregarding the value (if any) of the vector fields on surfaces of zero measure in the state space. This gives for the three above cases:

$$\dot{x}(t) \in \{g(t)\} + \begin{cases} 1 & \text{if } x < 0, \\ -1 & \text{if } x > 0, \\ [-1, 1] & \text{if } x = 0, \end{cases} \tag{2.73}$$

$$\dot{x}(t) \in \{g(t)\} + \begin{cases} -1 & \text{if } x < 0, \\ 1 & \text{if } x > 0, \\ [-1, 1] & \text{if } x = 0, \end{cases} \tag{2.74}$$

$$\dot{x}(t) = 1 + g(t) \tag{2.75}$$

with $x(0) \in \mathbb{R}$. Some comments arise:

- Since Filippov ignores values on sets of measure zero, one can in particular assign any value to the vector field on $x = 0$ in (2.70), (2.71) or (2.72): this does not change the right-hand-sides of the differential inclusions in (2.73), (2.74) or (2.75);
- Let us write (2.70)–(2.72) as $\dot{x}(t) = g(t) + h(x(t))$. Suppose we assign the value $h(0) = a$ to the vector field in the above three systems in (2.70), (2.71) and (2.72). Then:
  - the three systems have a fixed point at $x = 0$ if and only if $g(t) = -a$ for all t;
  - if $x(0) = 0$, then (2.70) has a solution on $\mathbb{R}^+$ if and only if $g(t) = -a$; this solution is $x(t) \equiv 0$. Otherwise the system can not be given a solution, because if at some t one has $x(t) = 0$, then $\dot{x}(t) \neq 0$ so that the trajectory has to leave the origin. However the vector field outside $x = 0$ tends to immediately push again the solution to $x = 0$: a contradiction. We conclude that the trajectories that start with $x(0) \neq 0$ exist until they reach $x = 0$, and not after;
  - if $x(0) = 0$, then (2.71) has a unique global in time solution that diverges asymptotically either to $+\infty$ or $-\infty$ depending on the sign of $g(0) + a$; (2.72) also has a unique solution that diverges to $+\infty$.

Consider now the three Filippov's systems in (2.73), (2.74) and (2.75). Then:

- $x = 0$ is a fixed point of (2.73) and (2.74). However (2.75) has no fixed point except if $g(t) \equiv -1$;
- the trajectories of (2.73) with $x(0) \neq 0$ reach $x = 0$ in a finite time t^*, and then stay on the "surface" $x = 0$; this is due to the fact that on the switching surface $x = 0$, there is always one element of the multivalued part of the right-hand-side,

i.e. $[-1, 1]$, that is able to compensate for $g(t)$ and to guarantee that $\dot{x}(t) = 0$ for all $t > t^*$; the origin $x = 0$ is an *attractive surface* called a *sliding surface* (the name surface is here not quite appropriate, but will be in higher dimensional systems).
- the differential inclusion in (2.74) has at least three solutions starting from $x(0) = 0$: $x(t) \equiv 0$, $x(t) = t + \int_0^t g(s)ds$ and $x(t) = -t + \int_0^t g(s)ds$. A spontaneous jump exists. Actually for all $T > 0$ the functions $x(t) = 0$ for $t \in [0, T]$, and $x(t) = t - T + \int_T^t g(s)ds$ for $t \geqslant T$ or $x(t) = -t - T + \int_T^t g(s)ds$ for $t \geqslant T$ are solutions.

The conclusion to be drawn from these simple examples is that embedding switching systems into Filippov's inclusions, may drastically modify their dynamics. This is a *modeling* step whose choice has to be carefully made from physical considerations.

2.4.4.2 Filippov's Sets

The general definition of a Filippov's set, starting from a general bounded vector field $f(x)$ (with possible points of discontinuity) is as follows:

$$F(x) = \bigcap_{\epsilon>0} \bigcap_{\mu(N)=0} \overline{\text{conv}}\, f((x + \epsilon B_n) \setminus N) \tag{2.76}$$

where B_n is the unit ball of $\mathbb{R}^n$, μ is the Lebesgue measure and $\overline{\text{conv}}(v_1, v_2, \ldots, v_n)$ denotes the closed convex hull of the vectors $v_1, v_2, \ldots, v_n$. Let us provide some insight on (2.76):

- by construction $F(x)$ is always non empty, closed and convex for each x;[7]
- let $x \in \mathbb{R}^n$. One considers the convex hull of all the values of $f(z)$, with $z \in x + \epsilon B_n$ and $\epsilon \to 0$. If $f(\cdot)$ is continuous at x then there is only one such values that is nothing else but $f(x)$, and $F(x) = \{f(x)\}$. If $f(\cdot)$ is discontinuous at x then all the different values that it takes in a neighborhood of x are taken into account;
- the definition of the set in (2.76) disregards what happens on subspaces of measure zero in $\mathbb{R}^n$, denoted as N in (2.76). In $\mathbb{R}^3$, it ignores the "isolated" values the vector field $f(x)$ may take on planes, lines, points. For instance in (2.70) one may assign any value of the right-hand-side at $x = 0$, without changing its Filippov's set in (2.73). Similarly for the other two systems;
- as alluded to above, embedding switching systems into Filippov's inclusions is a particular choice; other notions exist, see Cortés (2008) for an introduction.
- in practice the computation of a solution in the sense of Filippov may not always be easy, because it may boil down to calculate the intersection between a hypersurface and a polyhedral set. This is particularly true when switching attractive surfaces with co-dimension larger than 2 exist.

[7] The boundedness of $f(x)$ is essential here.

Starting from (2.76), the Filippov's differential inclusion is:

$$\dot{x}(t) \in F(x(t)), \quad x(0) = x_0 \in \mathbb{R}^n. \tag{2.77}$$

When particularized to the switching systems in (2.69), one obtains for $x \in \Sigma \overset{\Delta}{=} \chi_{i_1} \cap \chi_{i_2} \cap \cdots \cap \chi_{i_k}$ with $i_1 \neq i_2 \neq \cdots \neq i_k$:

$$F(x) = \overline{\mathrm{conv}}(A_{i_1}x + a_{i_1}, A_{i_2}x + a_{i_2}, \dots, A_{i_k}x + a_{i_k}) \tag{2.78}$$

and one disregards the possible values on Σ which is of codimension $k > 0$ and therefore of measure zero in $\mathbb{R}^n$. The set $F(x)$ in (2.78) is a polyhedral set of $\mathbb{R}^n$: a segment if $k = 2$, a triangle if $k = 3$, *etc*.

2.4.4.3 Existence of Absolutely Continuous Solutions

It happens that a differential inclusion whose right-hand-side is a Filippov's set, always possesses at least one solution that is absolutely continuous. Before stating the result let us provide a definition.

Definition 2.67 (Outer semi-continuous differential inclusions) A differential inclusion is said to be outer semi-continuous if the set-valued map $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ satisfies the following conditions:

1. it is closed and convex for all $x \in \mathbb{R}^n$;
2. it is outer semi-continuous, *i.e.* for every open set M containing $F(x), x \in \mathbb{R}$, there exists a neighborhood Ω of x such that $F(\Omega) \subset M$.

Filippov's sets satisfy such requirements when the discontinuous vector field $f(\cdot)$ is bounded, and the next Lemma applies to Filippov's differential inclusions.

Lemma 2.68 *Let $F(x)$ satisfy the conditions of Definition 2.67, and in addition $\|F(x)\| \leqslant c(1 + \|x\|)$ for some $c > 0$ and all $x \in \mathbb{R}^n$. Then there is an absolutely continuous solution to the differential inclusion $\dot{x}(t) \in F(x(t))$ on $\mathbb{R}^+$, for every $x_0 \in \mathbb{R}^n$.*

This result extends to time-varying inclusions $F(t, x)$ (Theorem 5.1 in Deimling 1992). The notation $\|F(x)\| \leqslant c(1 + \|x\|)$ means that for all $\xi \in F(x)$ one has $\|\xi\| \leqslant c(1 + \|x\|)$: this is a linear growth condition. In view of (2.78) a solution has to satisfy the differential equation

$$\dot{x}(t) = \sum_{j=1}^{k} \alpha_{i_j}(A_{i_j}x(t) + a_{i_j}), \tag{2.79}$$

for some $\alpha_{i_j} \in (0, 1)$ with $\sum_{j=1}^{k} \alpha_{i_j} = 1$.

2.4.4.4 Uniqueness of Solutions

The uniqueness of solutions is a more tricky issue than the existence one, as in general it is not guaranteed by the Filippov's set. Example (2.74) shows that even in very simple cases uniqueness may fail. In order to obtain the uniqueness property one has to impose more on the set-valued mapping $F(\cdot)$. The maximal monotone property can be used to guarantee the uniqueness of the solutions, see Proposition 2.58. It is easy to check that the system in (2.73) fits within the framework of Proposition 2.58, whereas the system in (2.74) does not.

When the switching surface is of codimension 1 (said otherwise: there is only one differentiable switching surface), then the following criterion that is due to Filippov (1964, 1988), assures the uniqueness of solutions.

Proposition 2.69 *Let us consider the polyhedral switching system in* (2.69) *with two cells* χ_1 *and* χ_2 *with a common boundary* $\partial\chi_1 = \partial\chi_2$ *denoted as* Σ. *Let us denote* $f : \mathbb{R}^n \to \mathbb{R}^n$ *its discontinuous piecewise-linear vector field. If, for each* $x \in \Sigma$, *either* $f_{\chi_1}(x) = A_1x + a_1$ *points into* χ_2, *or* $f_{\chi_2}(x) = A_2x + a_2$ *points into* χ_1, *then there exists a unique Filippov's solution for any* $x(0) \in \mathbb{R}^n$.

The proposition says that if the switching surface Σ is attractive, or if it is crossing, then the differential inclusion constructed with the Filippov's set (2.76) enjoys the uniqueness of solutions property, within the set of absolutely continuous functions. When Σ is attractive then the solution slides along it (a sliding motion), in the other case it justs crosses Σ.

Notice that if the convex combination in (2.79) is unique so is the solution. The point is that when the discontinuity surface is of codimension larger than 2, the conditions of Proposition 2.69 are no longer sufficient to guarantee the uniqueness of such a convex combination.

Example 2.70 This example is taken from Johansson (2003). We consider the following piecewise-linear system:

$$\begin{cases} \dot{x}_1(t) = x_2(t) - \mathrm{sgn}(x_1(t)), \\ \dot{x}_2(t) = x_3(t) - \mathrm{sgn}(x_2(t)), \\ \dot{x}_3(t) = -2x_1(t) - 4x_2(t) - 4x_3(t) - x_3(t)\,\mathrm{sgn}(x_2(t))\,\mathrm{sgn}(x_1(t)+1), \end{cases} \tag{2.80}$$

where $\mathrm{sgn}(\cdot)$ is here just the discontinuous single valued sign function. This switching system has four cells χ_i. The surfaces $\Sigma_1 = \{x \in \mathbb{R}^3 \mid x_1 = 0, |x_2| \leqslant 1\}$, $\Sigma_2 = \{x \in \mathbb{R}^3 \mid x_2 = 0, |x_3| \leqslant 1\}$, and the line $\Sigma_{12} = \{x \in \mathbb{R}^3 \mid x_1 = 0, x_2 = 0, |x_3| \leqslant 1\}$ are attractive. Therefore the Filippov solutions slide on these surfaces once they attain them. Both Σ_1 and Σ_2 are of codimension 1 so that Proposition 2.69 applies. However Σ_{12} is of codimension 2. It can be checked that the following two sets of coefficients:

$$\alpha_1 = \frac{1+\mathrm{sgn}(x_3)}{4}, \qquad \alpha_2 = \frac{1+x_3}{2}, \qquad \alpha_3 = -\frac{x_3}{2} + \alpha_1, \qquad \alpha_4 = \frac{1}{2} - \alpha_1$$

and

$$\beta_1 = \alpha_2, \qquad \beta_2 = \alpha_1, \qquad \beta_3 = \alpha_4, \qquad \beta_4 = \alpha_3,$$

both define differential equations as in (2.79) whose solution is a solution of the Filippov's inclusion for (2.80).

There exists a more general property than monotonicity which guarantees the uniqueness of solutions: the one-sided-Lipschitz-continuity. This property, that is useful to show uniqueness of solutions, was introduced for stiff ordinary differential equations by Dekker and Verwer (1984) and Butcher (1987), and for differential inclusions in Kastner-Maresch (1990–1991) and Dontchev and Lempio (1992). It was already used by Filippov to prove the uniqueness of solutions for ordinary differential equations with discontinuous right-hand-side Filippov (1964). Let us provide a definition that may be found in Dontchev and Farkhi (1998).

Definition 2.71 The set valued map $F : \mathbb{R}^n \to 2^{\mathbb{R}^n} \setminus \emptyset$ where $F(t,x)$ is compact for all $x \in \mathbb{R}^n$ and all $t \geqslant 0$, is called *one-sided Lipschitz continuous* (OSLC) if there is an integrable function $L : \mathbb{R}^+ \to \mathbb{R}$ such that for every $x_1, x_2 \in \mathbb{R}^n$, for every $y_1 \in F(t,x_1)$, there exists $y_2 \in F(t,x_2)$ such that

$$\langle x_1 - x_2, y_1 - y_2 \rangle \leqslant L(t)\|x_1 - x_2\|^2.$$

It is called *uniformly one-sided Lipschitz continuous* (UOSLC) if this holds for all $y_2 \in F(t,x_2)$.

It is noteworthy that $L(\cdot)$ may be constant, time-varying, positive, negative, or zero. We recall that here $\langle \cdot, \cdot \rangle$ simply means the inner product in $\mathbb{R}^n$, but the OSLC condition may also be formulated for other inner products.

Example 2.72 All set-valued mappings that may be written as $F(t,x) = f(t,x) - \varphi(x)$, with $\varphi : \mathbb{R}^n \to \mathbb{R}^n$ are multivalued monotone mappings, and $f(t,x)$ is Lipschitz continuous, are UOSLC. The OSLC constant L is equal to $\max(0, \lambda)$, where λ is the Lipschitz constant of the function $f(\cdot,\cdot)$.

Example 2.73 Consider $F(x) = \mathrm{sgn}(x)$, the set-valued sign function. For all x_1, x_2, and $y_1 \in F(x_1)$, $y_2 \in F(x_2)$, one has $\langle x_1 - x_2, y_1 - y_2 \rangle \geqslant 0$. Therefore the multifunction $-F(\cdot)$ satisfies $\langle x_1 - x_2, -y_1 + y_2 \rangle \leqslant 0$ and is UOSLC with constant $L = 0$ (this is consistent with Example 2.72 with $\varphi(x) = \partial |x|$). However $F(\cdot)$ is not OSLC, hence not UOSLC. Indeed take $x_1 > 0$, $x_2 < 0$, so that $y_1 = 1$, $y_2 = -1$. We get $(x_1 - x_2)(y_1 - y_2) = 2(x_1 - x_2) > 0$. OSLC implies that $2(x_1 - x_2) \leqslant L(x_1 - x_2)^2$ for some L. A negative L is impossible, and a nonnegative L yields $L \geqslant \frac{2}{x_1 - x_2}$. As $x_1 - x_2$ approaches 0, L diverges to infinity.

As shown in Cortés (2008), the one-sided-Lipschitz-continuity cannot be satisfied by discontinuous vector fields as in (2.69), with $L > 0$. However a maximal monotone mapping $F(\cdot)$ necessarily has its opposite $-F(\cdot)$ that is UOSLC with $L = 0$. The next result holds.

Lemma 2.74 *Let $F(\cdot,\cdot)$ be UOSLC with constant L, and let $x_1 : [t_0, +\infty) \to \mathbb{R}^n$, $x_2 : [t_0, +\infty) \to \mathbb{R}^n$ be two absolutely continuous solutions of the DI: $\dot{x}(t) \in$*

$F(t, x(t))$, i.e. $\dot{x}_1(t) \in F(t, x_1(t))$ and $\dot{x}_2(t) \in F(t, x_2(t))$ almost everywhere on $[t_0, +\infty)$. Then

$$\|x_1(t) - x_2(t)\| \leqslant \exp(L(t - t_0))\, \|x_1(t_0) - x_2(t_0)\| \tag{2.81}$$

for all $t \geqslant t_0$. In particular, the differential inclusion: $\dot{x}(t) \in F(t, x(t))$ enjoys the uniqueness of solutions property.

When particularized to maximal monotone mappings one has to consider inclusions of the form $\dot{x}(t) \in -F(t, x(t))$ (see (2.49) and Proposition 2.58).

2.4.4.5 Detection of the Sliding Modes

Let us consider the switching system in (2.69). It is of interest to propose a criterion for the detection of the attractive surfaces. First of all notice that a sliding mode may occur if the (discontinuous) vector field points towards the switching surface on both sides of it: this is called a first-order (or regular) sliding mode. But it may also occur if it is tangent to the switching surface on both sides of it, while its time derivatives still both point towards the switching surface: this is called a second-order sliding mode. And so on for higher order sliding modes.

To start with let us assume that the boundary Σ_{ij} between the two cells χ_i and χ_j, is included into the subspace $\{x \in \mathbb{R}^n \mid c_{ij}^T x + d_{ij} = 0\}$. Suppose also that the polyhedron χ_i is such that $c_{ij}^T x + d_{ij} \geqslant 0$ for all $x \in \chi_i$ (and consequently $c_{ij}^T x + d_{ij} \leqslant 0$ for all $x \in \chi_j$). Then the set:

$$S_{ij} = \{x \in \Sigma_{ij} \mid c_{ij}^T(A_i x + a_i) < 0 \text{ and } c_{ij}^T(A_j x + a_j) > 0\} \tag{2.82}$$

is a first-order (or regular) sliding set for the switching system (2.69) on Σ_{ij}. If at some point $x \in \Sigma_{ij}$ one has $c_{ij}^T(A_i x + a_i) = c_{ij}^T(A_j x + a_j) = 0$ while $\frac{d^2}{dt^2} c_{ij}^T x(t) = c_{ij}^T < 0$ in χ_i and $\frac{d^2}{dt^2} c_{ij}^T x(t) > 0$ in χ_j (in other words $c_{ij}^T(A_i^2 x + A_i a_i) < 0$ and $c_{ij}^T(A_j^2 x + A_j a_j) > 0$) then a second-order sliding mode occurs. It is possible to construct a linear programme to calculate the points inside a regular sliding set as follows (Johansson 2003):

$$
\begin{aligned}
&(x^*, \epsilon^*) = \operatorname{argmin} \epsilon \\
&\text{subject to:} \quad \begin{pmatrix} C_i \\ C_j \\ -c_{ij}^T A_i \\ c_{ij}^T A_j \end{pmatrix} x + \begin{pmatrix} D_i \\ D_j \\ -c_{ij}^T a_i \\ c_{ij}^T a_j \end{pmatrix} \geqslant \begin{pmatrix} 0 \\ 0 \\ \epsilon \\ \epsilon \end{pmatrix}.
\end{aligned}
$$

If $\epsilon^* > 0$ then the switching system has a non empty regular sliding set on S_{ij}.

2.4.5 Maximal Monotone Inclusions, Unilateral Differential Inclusions

Maximal monotone differential inclusions are essentially differential inclusions as in (2.49). They are not "standard" differential inclusions, because their right-hand-side may not be a compact subset of $\mathbb{R}^n$. The most typical example is when the right-hand-side is a normal cone to a convex non empty set. In view of the material of Sect. 2.4.2 we will not investigate more such differential inclusions.

Remark 2.75 (Relay systems) A popular class of discontinuous systems in the Systems and Control research community, is made of so-called *relay systems*. Their well-posedness has been investigated in several papers, see *e.g.* Lootsma et al. (1999), Lin and Wang (2002) and Acary and Brogliato (2010). Relay systems are as follows:

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t), \\ y(t) = Cx(t) + Du(t), \\ u(t) \in -\operatorname{Sgn}(y(t)), \end{cases} \tag{2.83}$$

where $\operatorname{Sgn}(y) = (\operatorname{sgn}(y_1)\operatorname{sgn}(y_2)\cdots\operatorname{sgn}(y_m))^T$, $\operatorname{sgn}(\cdot)$ is the sign multifunction, $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$, $y(t) \in \mathbb{R}^m$. Such discontinuous systems may belong to the class of Filippov's differential inclusions, or maximal monotone differential inclusions, and can also be rewritten into a complementarity systems formalism. Some subclasses of relay systems are Filippov's inclusions (see for instance the simple example (2.73) and replace $g(t)$ by a linear term $Ax(t)$), and other subclasses are of the maximal monotone type with a right-hand-side that is not necessarily a Filippov's set (see Acary and Brogliato 2010). This last result may come as a surprising fact because the right-hand-side of relay systems contains the multivalued sign function, that is a common ingredient in simple Filippov (and sliding mode) systems. It is however easily checked that the system:

$$\dot{x}(t) \in -C^T \operatorname{Sgn}(Cx(t)) \tag{2.84}$$

with $C = \left(\begin{smallmatrix} 1 & -1 \\ 1 & 1 \end{smallmatrix}\right)$, $\operatorname{Sgn}(z) = (\operatorname{sgn}(z_1, \dots, \operatorname{sgn}(z_n))^T$ for any vector $z \in \mathbb{R}^n$, has a maximal monotone right-hand-side $x \mapsto C^T \operatorname{Sgn}(Cx)$. However the set $C^T \operatorname{Sgn}(Cx)$ may strictly contain the Filippov's set of the associated discontinuous vector field at $x = 0$, that is the closed convex hull of the vectors $(2, 0)^T$, $(0, 2)^T$, $(0, -2)^T$, $(-2, 0)^T$. This indicates that the Filippov framework for embedding switching systems may not always be the most suitable framework.

In Lootsma et al. (1999) and Lin and Wang (2002) the uniqueness of continuous, piecewise-analytic solutions is proved, relying on complementarity arguments. In Acary and Brogliato (2010) relay systems are recast into differential inclusions (2.49) and the well-posedness is shown *via* Proposition 2.58.

2.4.6 Equivalences Between the Formalisms

We have seen in Sects. 2.3.3, 2.3.4 and 2.3.5 the close link between generalized equations, complementarity problems, and variational inequalities. Quite naturally similar relations exist between their dynamical counterparts. In Sect. 2.4.2 the link between dynamical variational inequalities and differential inclusions into normal cones is established, see (2.47). To start with let us consider the DVI in (2.45), with $\varphi(\cdot) = \psi_{\mathbf{C}}(\cdot)$ (the indicator function of $\mathbf{C}$) for some non empty closed convex set $\mathbf{C}$. Then the DVI is equivalent to (2.46) and using (2.24) it is easy to obtain that it is also equivalent to the complementarity system:

$$\begin{cases} \dot{x}(t) = -f(x(t),t) + \lambda(t), \\ \mathbf{C} \ni x(t) \perp \lambda(t) \in \mathbf{C}^*. \end{cases} \tag{2.85}$$

As another example we may consider the differential inclusion in (2.73). As seen in Sect. 2.4.4 this is a Filippov differential inclusion. This is also a differential inclusion of the type (2.49) whose set-valued right-hand-side is a maximal monotone operator $x \mapsto \operatorname{sgn}(x)$, where $\operatorname{sgn}(\cdot)$ is the multivalued sign function. We also have the following for two reals y and z:

$$\begin{aligned} y \in \operatorname{sgn}(z) \quad &\Leftrightarrow \quad y = \frac{\lambda_1 - \lambda_2}{2}, \\ \lambda_1 + \lambda_2 = 2, \quad &\begin{cases} 0 \leqslant \lambda_1 \perp -z + |z| \geqslant 0, \\ 0 \leqslant \lambda_2 \perp z + |z| \geqslant 0. \end{cases} \end{aligned} \tag{2.86}$$

Indeed let $z > 0$, then $\lambda_2 = 0$ and $\lambda_1 \geqslant 0$ so that $\lambda_1 = 2$ and $y = 1$. Let $z < 0$, then $\lambda_1 = 0$ and $\lambda_2 \geqslant 0$ so that $\lambda_2 = 2$ and $y = -1$. Let $z = 0$, then $\lambda_1 \geqslant 0$ and $\lambda_2 \geqslant 0$. Since $\lambda_1 = 2 - \lambda_2$ we get $y = 1 - \lambda_2$ so $y \leqslant 1$. Similarly $\lambda_2 = 2 - \lambda_1$ and $y = \lambda_1 - 1$ so $y \geqslant -1$. Finally when $z = 0$ we obtain that $y \in [-1, 1]$. The complementarity conditions in (2.86) do represent the multivalued sign function. One may therefore rewrite in an equivalent way the differential inclusion (2.73) as:

$$\begin{cases} \dot{x}(t) = -\frac{\lambda_1 - \lambda_2}{2}, \\ \lambda_1 + \lambda_2 = 2, \\ 0 \leqslant \lambda_1 \perp -x(t) + |x(t)| \geqslant 0, \\ 0 \leqslant \lambda_2 \perp x(t) + |x(t)| \geqslant 0, \end{cases} \tag{2.87}$$

which is a complementarity system that may be recast into (2.51). Still there exists another formalism (Camlibel 2001):

$$\begin{cases} \dot{x}(t) = 1 - 2\lambda_1, \\ 0 \leqslant \begin{pmatrix} -x \\ 1 \end{pmatrix} + \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} \perp \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} \geqslant 0. \end{cases} \tag{2.88}$$

This complementarity system belongs to the class in (2.54) with E and M identity matrices of appropriate dimensions, $F = 1$, $G = \binom{0}{1}$. It may also be recast into (2.57) choosing $u(t) \equiv 1$. Let us continue with another mathematical formalism for (2.73). We know from Example 2.25 that the subdifferential of the absolute value function $x \in \mathbb{R} \mapsto |x|$, is the sign multifunction. We can therefore use (2.47) and its equivalent form in (2.45) to rewrite equivalently (2.73) as the dynamical variational inequality:

$$\begin{cases} x(t) \in \mathbb{R} & \text{for all } t \geqslant 0, \\ \langle \dot{x}(t), v - x(t) \rangle - |v| + |x(t)| \geqslant 0 & \text{for all } v \in \mathbb{R}. \end{cases} \tag{2.89}$$

Let us provide the detailed proof of the fact that the multivalued relay function may be rewritten as a variational inequality. The variational inequality formalism of $y \in -\text{sgn}(x)$ is: $x \in \mathbb{R}$ and

$$\langle y, v - x \rangle + |v| - |x| \geqslant 0 \quad \text{for all } v \in \mathbb{R}.$$

Indeed:

- $x = 0$: we get $\langle y, v \rangle + |v| \geqslant 0$ for all v, i.e. $y \in [-1, +1]$,
- $x > 0$: we get $\langle y, v - x \rangle + |v| - x \geqslant 0$ for all v. Take $v = 0$: $\langle y, -x \rangle - x \geqslant 0$ i.e. $x(y + 1) = 0$ which implies $y = -1$.
- $x < 0$: we get $\langle y, v - x \rangle + |v| - x \geqslant 0$ for all v. Take $v = 0$: $\langle y, -x \rangle + x \geqslant 0$ i.e. $x(y - 1) = 0$ which implies $y = 1$.

The sign multifunction, also called the relay multifunction, is maximal monotone, it is a Filippov's set, and it can be represented through various complementarity relations or with variational inequalities of the second kind.

This however does not contradict the comments in Remark 2.75 that circuits with relay functions may not always be Filippov's inclusions, because a lot depends then on the matrices A, B, C, D. Let us finally notice that using Examples 2.11, 2.25, 2.26, and finally (2.9), one infers that the following holds:

$$y \in \text{sgn}(x) \quad \Leftrightarrow \quad x \in N_{[-1,1]}(y). \tag{2.90}$$

Let us now consider the LCS in (2.59). Let us assume that $D = 0$, and that there exists a matrix $P = P^T > 0$ such that

$$PB = C^T. \tag{2.91}$$

This may be a consequence of the LMI in (2.60) (see Sect. A5 in Brogliato et al. 2007). Let us make the state space variable change $z = Rx$, where R is the symmetric positive definite square root of P. We further define the following two sets:

$$K(t) := \{x \in \mathbb{R}^n \mid Cx + Fu(t) \geqslant 0\} \tag{2.92}$$

and

$$S(t) := R(K(t)) = \{Rx \mid x \in K(t)\}, \tag{2.93}$$

which are convex polyhedral for each fixed t. In Brogliato and Thibault (2010) it is shown that, when the input signal $u(\cdot)$ is absolutely continuous and under certain conditions, the LCS in (2.59) is equivalent to a perturbed sweeping process and to a dynamical variational inequality. When $u(\cdot)$ is locally BV, things are a bit more tricky in the sense that the perturbed sweeping process formalism has to be recast into measure differential inclusions, and encapsulates the LCS one. The following constraint qualification is supposed to hold:

$$\mathrm{Rge}(C) - \mathbb{R}^m_+ = \mathbb{R}^m, \tag{2.94}$$

where Rge is the range. This is quite similar to the constraint qualification in Proposition 2.65 when $D = 0$. The equality in (2.94) means that for all $x \in \mathbb{R}^m$, there exists $y \in \mathrm{Rge}\,(C)$ and $z \in \mathbb{R}^m_+$ such that $z - y = x$. Obviously it holds whenever the linear mapping associated with C is onto, *i.e.* the matrix C has rank m, but also in many other cases. Then we have the following result when solutions are absolutely continuous: the LCS in (2.59) is equivalent to the differential inclusion

$$-\dot{z}(t) + RAR^{-1}z(t) + REu(t) \in N_{S(t)}(z(t)), \tag{2.95}$$

which is a perturbed sweeping process, that is in turn equivalent to the DVI

$$\begin{aligned} &\langle \dot{z}(t) - RAR^{-1}z(t) - REu(t), v - z(t)\rangle \geqslant 0 \\ &\text{for all } v \in S(t),\ z(t) \in S(t) \text{ for all } t \geqslant 0. \end{aligned} \tag{2.96}$$

The passage from the complementarity system to the perturbed sweeping process uses the fact that thanks to (2.91) one can formally rewrite the complementarity system into a gradient form in the z coordinates.

The equivalences between various formalisms are understood as follows: given an initial condition $x(0) = R^{-1}z(0)$, then both systems possess the same unique solution over $\mathbb{R}^+$. The rigorous proof may be found in Brogliato and Thibault (2010), where it is also shown that the state jump laws in Proposition 2.65 readily follow from basic convex analysis arguments. When the state is prone to discontinuities then the measure differential inclusion formalism has to be used, similarly to (2.36) where the solution is to be understood as in Definition A.7. The state variable change $z = Rx$ relying on the input/output property $PB = C^T$ has been introduced in Brogliato (2004), where the equivalence between passive LCS and inclusions into normal cones is established. Equivalences between gradient complementarity systems in (2.64), projected dynamical systems, dynamical variational inequalities and inclusions into normal cones are shown in Brogliato et al. (2006). Such studies are rooted in Cornet (1983) and Henry (1973).

Complementarity dynamical systems, dynamical variational inequalities, differential inclusions into normal cones, belong to the same family of nonsmooth evolution problems. The dynamics of electrical circuits with nonsmooth electronic devices such as ideal diodes, can be recast into such mathematical formalisms.

2.5 The Dynamics of the Simple Circuits

Let us now return to Sect. 1.1 of Chap. 1 and use the material of this chapter to rewrite the dynamics of the simple circuits. The objective of this section is to show

how one may take advantage of the mathematical tools which have been introduced in the foregoing sections, to analyze and better understand the dynamics of nonsmooth electrical circuits.

2.5.1 The Ideal Diode Voltage/Current Law

Let us consider the ideal diode of Fig. 1.1 whose complementarity formalism is in Fig. 1.2(b). Using (2.23) we may rewrite its voltage/current law as

$$-(v(t)+a) \in N_{\mathbb{R}^+}(i(t)+b) \quad \Leftrightarrow \quad -(i(t)+b) \in N_{\mathbb{R}^+}(v(t)+a). \tag{2.97}$$

A variational inequality formalism is also possible using (2.26) and (2.13): find $v(t) \geqslant -a$ such that:

$$\langle v(t)+a, y-i(t)-b\rangle \geqslant 0 \quad \text{for all } y \geqslant 0. \tag{2.98}$$

2.5.2 The Piecewise-Linear Diode Voltage/Current Law

We now consider the diode of Fig. 1.2(d). Let us see how the MCP formulation in (2.21) may be used to represent its characteristic. First of all its voltage/current law is expressed in a complementarity formalism as:

$$0 \leqslant -v(t) + R_{\mathit{off}}\, i(t) \perp -v(t) \geqslant 0, \tag{2.99}$$

which we may again rewrite as

$$-v(t) + R_{\mathit{off}}\, i(t) \in N_{\mathbb{R}^+}(-v(t)). \tag{2.100}$$

Using (2.25) we infer that

$$-v(t) = \mathrm{proj}\big(\mathbb{R}_+;\, R_{\mathit{off}}\, i(t)\big). \tag{2.101}$$

Let us now turn our attention to (2.21). We may choose $w = F^+(z) = F(z) = v(t) - R_{\mathit{off}}\, i(t) = w \geqslant 0$,[8] with $v = 0$ in (2.21), $z = -v(t)$, $l = 0$ and $u = +\infty$. Thus we rewrite equivalently the voltage/current law as:

$$\begin{cases} v(t) - R_{\mathit{off}}\, i(t) \geqslant 0, \\ 0 \leqslant -v(t) \leqslant +\infty, \\ v(t)(v(t) - R_{\mathit{off}}\, i(t)) = 0. \end{cases} \tag{2.102}$$

2.5.3 A Mixed Nonlinear/Unilateral Diode

The various diode models in Fig. 1.2 may be enlarged towards mixed models that contain some unilateral effects, and nonlinear smooth behaviour. Consider for instance the voltage/current law whose graph is in Fig. 2.19. The function $v \mapsto g(v)$

[8] $F^+(z) = \max(0, F(z))$.

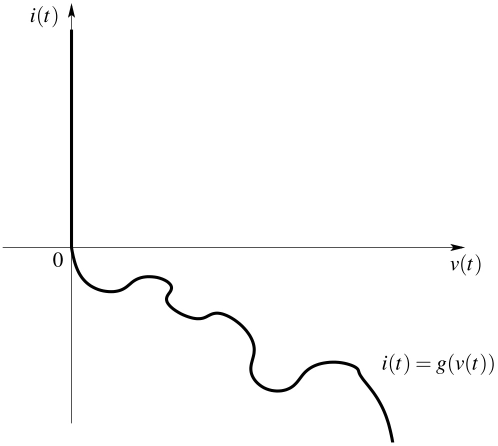


Fig. 2.19 A diode with a mixed nonlinear/unilateral behaviour

satisfies $g(0) = 0$ and $g(v) < 0$ for all $v > 0$. The voltage/current law may take various equivalent forms:

$$0 \leqslant -g(v(t)) + i(t) \perp v(t) \geqslant 0 \quad \Leftrightarrow \quad -g(v(t)) + i(t) \in -N_{\mathbb{R}^+}(v(t)). \tag{2.103}$$

One checks that $v(t) > 0$ implies that $i(t) = g(v(t))$, while $v(t) = 0$ implies $i(t) \geqslant 0$. Let us now consider the circuit of Fig. 2.20, with a voltage source $u(t)$. The state variables are $x_1(\cdot)$ the capacitor charge, and $x_2(\cdot)$ the current $i(t)$ through the circuit. The convention of Fig. 1.1 is chosen. One obtains:

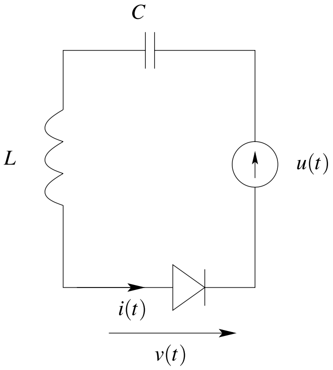
$$\begin{cases} \dot{x}(t) = \begin{pmatrix} 0 & 1 \\ \frac{-1}{LC} & 0 \end{pmatrix} x(t) + \begin{pmatrix} 0 \\ \frac{1}{L} \end{pmatrix} (v(t) + u(t)), \\ 0 \leqslant w(t) = -g(v(t)) + x_2(t) \perp v(t) \geqslant 0, \end{cases} \tag{2.104}$$

which is an NLCS as in (2.55), letting $\lambda(t) = v(t)$. The complementarity conditions in (2.104) define an NLCP (or NCP). If $x_2(t) < 0$ and $x_2(t)$ is in the image of $g(\cdot)$, then $-g(v(t)) = -x_2(t) > 0$ for some $v(t) > 0$. Uniqueness holds if the function $g(\cdot)$ is monotone (strictly decreasing). If $x_2(t) > 0$ then $v(t) = 0$ is a solution of the NCP. In Sect. 2.3.2 we gave results for LCP only. Well-posedness results for NCPs as in (2.17) exist, see for instance Facchinei and Pang (2003), Propositions 2.2.12 and 3.5.10.

2.5.4 From Smooth to Nonsmooth Electrical Powers

The introduction of the indicator function and of its subdifferential, allows one to embed the ideal diode into a rigorous mathematical framework that is useful for the analysis of circuits which contain such devices. As depicted in Fig. 2.21 it also permits in a quite convenient way to define the electrical power that is associated with such a nonsmooth multivalued electrical device. This is quite related with so-called Moreau's superpotential functions. So-called electrical superpotentials have

Fig. 2.20 A circuit with a mixed diode



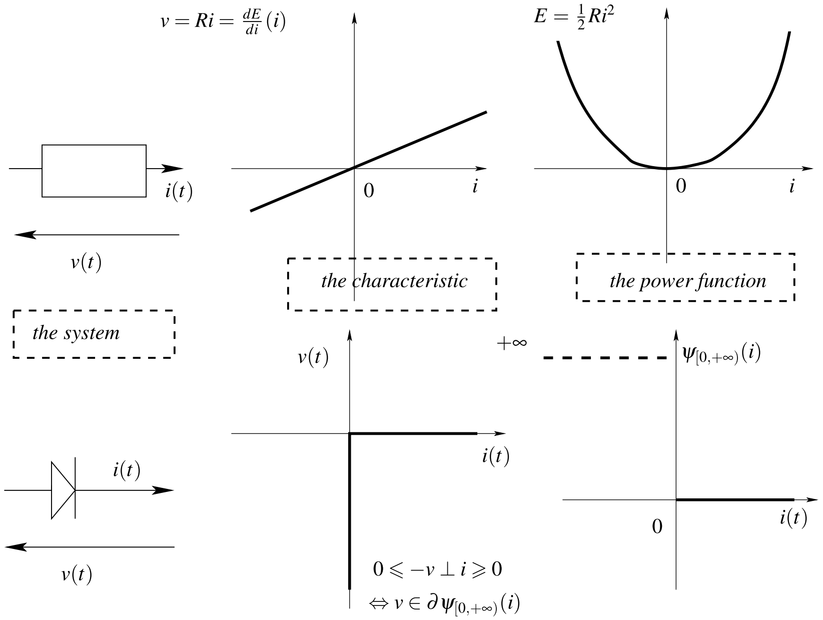


Fig. 2.21 From smooth to nonsmooth powers

been introduced in Addi et al. (2007, 2010) and Goeleven (2008). Let us consider a proper convex lower semi-continuous function $\varphi : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$. Suppose that an electrical device has the ampere-volt characteristic that is represented by $v \in \partial\varphi(i)$. Then $\varphi(\cdot)$ is called an electrical superpotential. Superpotentials have been introduced in mechanics by Moreau (1968). Consider Fig. 1.3 and let us reverse the coordinates so as to obtain the characteristic of $v(t)$ as a function of $i(t)$. The superpotential of the ideal diode is easily found to be $\varphi(i) = \psi_{\mathbb{R}^+}(i)$, the indicator function of $\mathbb{R}^+$. Thus $v \in \partial\psi_{\mathbb{R}^+}(i)$ so that $v = 0$ if $i > 0$ while $v \leqslant 0$ if $i = 0$. The (i, v) characteristic is maximal monotone. One may draw the parallel between the ampere-volt characteristic of a constant positive resistor, $u = Ri$, whose power func-

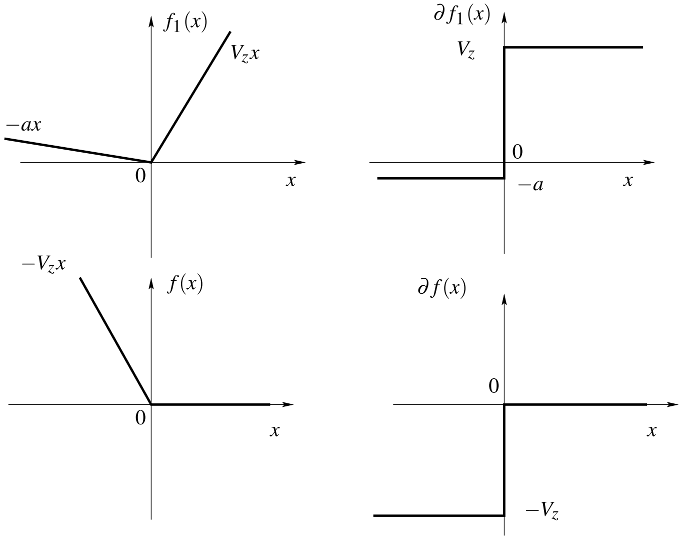


Fig. 2.22 Subdifferentials

tion is $E = \frac{1}{2}Ri^2$, and the ampere-volt characteristic of the ideal diode $v \in \partial\psi_{\mathbb{R}^+}(i)$ whose power multifunction is $\psi_{\mathbb{R}^+}(i)$. The same applies to the Zener diode, with a different superpotential, see (1.7) and Figs. 2.11 or 2.22.

Remark 2.76 As a convention superpotentials define a maximal monotone mapping. This means that the current/voltage mapping has to be chosen in accordance. The conventions of Fig. 1.1 and (2.97) are not suitable.

2.5.5 The RLD Circuit in (1.16)

From (2.24) one deduces that $0 \leqslant w(t) = x(t) - i(t) \perp v(t) \geqslant 0$ is equivalent to $v(t) \in -N_{\mathbb{R}_+}(x(t)-i(t))$. From the fact that $N_{\mathbb{R}_+}(x(t)-i(t)) = \partial\psi_{\mathbb{R}_+}(x(t)-i(t))$ and that $\psi_{\mathbb{R}_+}(x(t)-i(t)) = \psi_{[i(t),+\infty)}(x(t))$ we find that (1.16) may be equivalently rewritten as:

$$-\dot{x}(t) - \frac{R}{L}x(t) \in N_{[i(t),+\infty)}(x(t)). \tag{2.105}$$

When $i : \mathbb{R} \to \mathbb{R}$ is not a constant function, this is a first order perturbed sweeping process. When $i(\cdot)$ is an absolutely continuous function, it follows from Theorem 2.55 that $x(\cdot)$ is also absolutely continuous. If $i(\cdot)$ is of local bounded variations and right continuous, then it may jump and at the times of discontinuities in $i(\cdot)$, $x(\cdot)$ may jump as well. In this situation Theorem 2.56 applies. Suppose for instance that at time t one has $x(t^-) = i(t^-)$ and that $i(t^+) > i(t^-)$. Then if $x(t^-) = x(t^+)$ it follows that $x(t^+) < i(t^+)$: this is not possible since it implies that

$N_{[i(t^+),+\infty)}(x(t^+)) = \emptyset$. There fore a jump has to occur in $x(\cdot)$ at t to keep the state inside the set $[i(t), +\infty)$.

Assume that $x(\cdot)$ jumps at $t = t_1$. From Theorem 2.56 we know that it is of local bounded variation and right continuous provided $i(\cdot)$ is. The differential inclusion in (2.105) has to be interpreted as a measure differential inclusion, *i.e.*:

$$-dx - \frac{R}{L}x(t^+)dt \in N_{[i(t^+),+\infty)}(x(t^+)), \tag{2.106}$$

where dx is the differential measure associated with $x(\cdot)$. Thus (2.106) represents the inclusion of measures into a normal cone to a convex set. Recall that due to the way we constructed this inclusion, the elements of the normal cone are the $-v(t)$ and that they also are measures. More precisely, it follows from the first line of (1.16) that if $x(\cdot)$ jumps at t, then necessarily v is a Dirac measure with atom equal to t (something like δ_t). At $t = t_1$ which is an atom of the differential measure dx one obtains:

$$-x(t_1^+) + x(t_1^-) \in N_{[i(t_1^+),+\infty)}(x(t_1^+)), \tag{2.107}$$

since $dt(\{t_1\} = 0$. We now may use (2.14) to infer that:

$$x(t_1^+) = \text{proj}([i(t_1^+), +\infty); x(t_1^-)). \tag{2.108}$$

It may be verified by inspection that (2.108) is equivalent to (1.26). Remember that we deduced (1.26) from the backward-Euler discretization algorithm of (1.16). This result suggests that the backward-Euler method in (1.17) is the right time-discretization of the measure differential inclusion (2.106). The advantage of using (2.106) is that it provides the whole dynamics in one shot. And it provides a rigorous explanation of the state jump rule.

Remark 2.77 All quantities are evaluated at their right limits in (2.106). Intuitively, this is because one wants to represent the dynamics in a *prospective* way. More mathematically, this permits to integrate the system on the whole real axis even in the presence of jumps in $i(\cdot)$.

Let us investigate the time-discretization of (2.106), with time step $h > 0$. We propose in a systematic way to approximate dx by $\frac{x_{k+1}-x_k}{h}$ on $[t_k, t_{k+1})$, and to approximate the right–limits by the discrete variable at t_{k+1}. Then one obtains from (2.106):

$$-x_{k+1} + x_k - h\frac{R}{L}x_k \in N_{[i_{k+1},+\infty)}(x_{k+1}) = N_{\mathbb{R}_+}(x_{k+1} - i_{k+1}). \tag{2.109}$$

Using (2.24) this is equivalent to:

$$\mathbb{R}_- \ni -x_{k+1} + x_k - h\frac{R}{L}x_k \perp x_{k+1} - i_{k+1} \in \mathbb{R}, \tag{2.110}$$

because $\mathbb{R}_-$ is the polar cone to $\mathbb{R}_+$. Transforming again we get:

$$0 \leqslant x_{k+1} - x_k + h\frac{R}{L}x_k \perp x_{k+1} - i_{k+1} \geqslant 0. \tag{2.111}$$

Recalling that the elements of $N_{\mathbb{R}_+}(x_{k+1} - i_{k+1})$ are equal to $-\sigma_{k+1} = -hv_{k+1}$ (see Sect. 1.1.5), one finds that (2.111) is equal to the complementarity conditions of (1.17). It suffices to replace x_{k+1} by its value in the first line of (1.17) to recover an LCP with unknown hv_{k+1}.

Therefore the time-discretization of the measure differential inclusion (2.106) yields the backward Euler scheme in (1.17). From the BV version of Theorem 2.88 the approximated piecewise-linear solution $x^N(\cdot)$ converges to the right continuous of local bounded variations solution of (2.106), including state jumps. The measure differential inclusion formalism is very well suited for the derivation of time-stepping schemes. Moreover it shows that the scheme has to be implicit. Indeed it is easy to see that writing $N_{[i_{k+1},+\infty)}(x_k)$ in the right-hand-side of (2.109) yields a discrete-time system which cannot be advanced to step $k+1$. The implicit way is the only way.

Remark 2.78 In the left-hand-side of (2.109) we can replace $\frac{R}{L}x_k$ by $\frac{R}{L}x_{k+1}$ to get a fully implicit scheme, then we get the same algorithm if the same operation is performed in (1.17).

Remark 2.79 Let us rewrite (2.105) as:

$$\begin{cases} \dot{x}(t) + \frac{R}{L}x(t) = -v(t), \\ v(t) \in N_{[i(t),+\infty)}(x(t)). \end{cases} \tag{2.112}$$

The representation as a Lur'e system in Fig. 1.18 is clear. From (2.109) the same can be done with respect to Fig. 1.19.

The measure differential inclusion in (2.106) is *the* correct formalism for the circuit of Fig. 1.10 when the current source delivers a current $i(t)$ that jumps. It allows one to encompass all stages of motion (continuous and discontinuous portions of the state trajectories) and to get a suitable discretization in one shot.

2.5.6 The RCD Circuit in (1.3)

Using (2.25) the second line of (1.3) can be rewritten as $-\frac{v(t)}{R} - \frac{u(t)}{R} + \frac{1}{RC}z(t) \in N_{\mathbb{R}_+}(v(t))$. This is equivalent to $v(t) = \mathrm{proj}(\mathbb{R}_+; -u(t) + \frac{1}{C}z(t))$. Inserting this into the first line of (1.3) one obtains:

$$\dot{z}(t) = -\frac{u(t)}{R} + \frac{1}{RC}z(t) + \frac{1}{R}\,\mathrm{proj}\left(\mathbb{R}_+; -u(t) + \frac{1}{C}z(t)\right). \tag{2.113}$$

Since the projection operator is single valued Lipschitz continuous, (2.113) is nothing else but an ordinary differential equation with Lipschitz continuous right-hand-side.

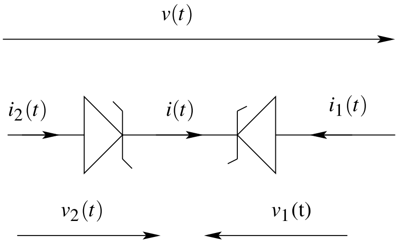


Fig. 2.23 Two Zener diodes mounted in series

2.5.7 The RLZD Circuit in (1.7)

The inclusion that represents the Zener diode voltage/current law is $v(t) \in \mathscr{F}_{z_i}(-i(t))$ in (1.7). From Fig. 1.6(a) it follows that the graph of this voltage/current law is maximal monotone. From Theorem 2.34 we may write it as the subdifferential of some convex lower semi-continuous proper function. This function is given by $f_1(-i) = \begin{cases} ai & \text{if } i \geqslant 0 \\ -V_z i & \text{if } i \leqslant 0 \end{cases}$ (see Fig. 2.22). Thus $\mathscr{F}_{z_i}(-i(t)) = \partial f_1(-i)$. We infer that the dynamics of this circuit is given by the differential inclusion:

$$\dot{x}(t) + \frac{R}{L}x(t) - \frac{u(t)}{L} \in \frac{1}{L}\partial f_1(-x(t)). \tag{2.114}$$

The right-hand-side of (2.114) takes closed convex values, and the multivalued mapping $y \mapsto \frac{1}{L}\partial f_1(y)$ is maximal monotone. Therefore this differential inclusion may be recast either into Filippov's inclusions, or in maximal monotone inclusions (see Sects. 2.4.4 and 2.4.5).

Similar developments hold for the voltage/current law in Fig. 1.6(b). Both characteristics can also be represented in a complementarity formalism.

2.5.8 Coulomb's Friction and Zener Diodes

Let us consider the Zener diode characteristic in Fig. 1.6 with $a = 0$. If two diodes are mounted in opposite series as in Fig. 2.23, then the voltage/current law is given by:

$$v(t) \in V_z\ \partial|z(t)|, \quad z(t) = -i(t), \tag{2.115}$$

where each diode has the voltage/current law $v_j(t) \in \mathscr{F}_{z_1}(-i_j(t))$ of Fig. 1.6 on the left, with $a = 0$. One obtains (2.115) by performing the operations as depicted in Fig. 2.24. This may be proved using Moreau-Rockafellar's Theorem 2.30. One has $v_2 \in \partial f_2(-i)$ with $f_2(-i) = \begin{cases} 0 & \text{if } -i < 0 \\ V_z(-i) & \text{if } -i > 0 \end{cases}$, $-v_1 \in \partial f_1(-i)$ with $f_1(-i) = \begin{cases} -V_z(-i) & \text{if } -i < 0 \\ 0 & \text{if } -i > 0 \end{cases}$. By Theorem 2.30 one has $v_2 - v_1 \in \partial f_2(-i) + \partial f_1(-i) = \partial(f_1 + f_2)(-i)$. And $\partial f_2(-i) + \partial f_1(-i) = \begin{cases} -V_z(-i) & \text{if } i < 0 \\ V_z(-i) & \text{if } -i > 0 \end{cases}$, whose subdifferential is multivalued at $i = 0$.

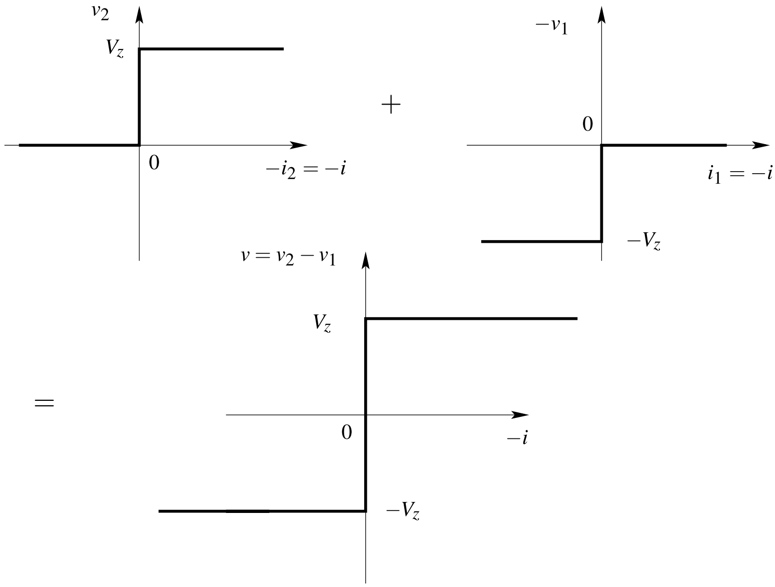


Fig. 2.24 The sum of the two Zener voltage/current laws

Reversing the sense of the diodes does not change the voltage/current law of the two-diode system, as may be checked. Consider now the circuit of Fig. 2.25, where each Zener box contains two Zener diodes mounted in series as in Fig. 2.23. Its dynamics is given by:

$$\begin{cases} L\frac{di_1}{dt}(t) + \frac{1}{\mathrm{C}}\int_0^t (i_1(s) - i_2(s))ds = v_1(t), \\ L\frac{di_2}{dt}(t) + \frac{1}{\mathrm{C}}\int_0^t (i_2(s) - i_1(s))ds = v_2(t), \\ v_1(t) \in V_z\, \partial|z_1(t)|, \quad z_1(t) = -i_1(t), \\ v_2(t) \in V_z\, \partial|z_2(t)|, \quad z_2(t) = -i_2(t). \end{cases} \tag{2.116}$$

Denoting $x_1(t) = \int_0^t i_1(s)ds$ and $x_2(t) = \int_0^t i_2(s)ds$ we can rewrite (2.116) as:

$$\begin{cases} \ddot{x}_1(t) + \frac{1}{L\mathrm{C}}(x_1(t) - x_2(t)) \in -\frac{V_z}{L}\,\mathrm{sgn}(x_1(t)), \\ \ddot{x}_2(t) + \frac{1}{L\mathrm{C}}(x_2(t) - x_1(t)) \in -\frac{V_z}{L}\,\mathrm{sgn}(x_2(t)), \\ x_1(0) = x_{10}, x_2(0) = x_{20}, \dot{x}_1(0) = \dot{x}_{10}, \dot{x}_2(0) = \dot{x}_{20}, \end{cases} \tag{2.117}$$

where we used $\partial|x| = \mathrm{sgn}(x)$ for all reals x, and Proposition 2.29 with $A = -1$. The circuit in Fig. 2.25 has therefore exactly the same dynamics as a two degree-of-freedom mechanical system made of two balls subjected to Coulomb's friction at the two contact points, related by a constant spring and moving on a line (see Sect. 3.11 in Acary and Brogliato 2008). The quantity V_z plays the role of the friction coefficient, L plays the role of the mass, $\frac{1}{\mathrm{C}}$ is the stiffness of the spring. As shown in Pratt et al. (2008) such a system can undergo, with a specific choice of the initial data, an infinity of events (stick-slip transitions in Mechanics) when a specific external

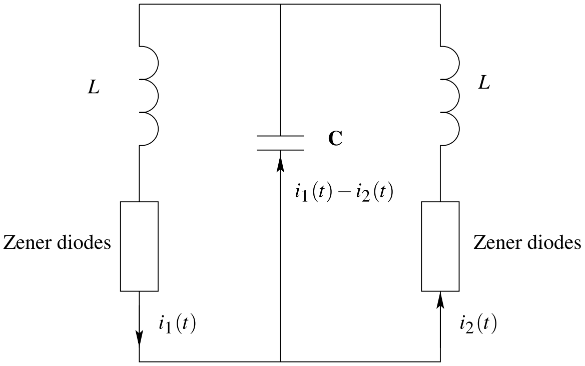


Fig. 2.25 Circuit with Zener diodes

excitation is applied to it. Obviously the dynamics in (2.117) can be recast into the framework of Fig. 1.18. It is also a Filippov's differential inclusion and Lemma 2.68 applies.

Remark 2.80 In Glocker (2005), Moeller and Glocker (2007) it is shown that the DC-DC buck converter can be written as a Lagrangian system, whose mass matrix consists of a diagonal matrix with either inductances or capacitances as its entries (this depends on the choice of the state variables). This is related to the choice of the state variables as the capacitors charges and the currents. We recover from another example that such a choice of the state variables yields a Lagrangian system whose mass matrix is made of the inductances. Indeed we can rewrite (2.117) as:

$$M\ddot{x}(t) + Kx(t) \in -B\,\mathrm{Sgn}(Cx(t)) \tag{2.118}$$

with $x^T = (x_1\ x_2)$, $M = \begin{pmatrix} L & 0 \\ 0 & L \end{pmatrix}$, $K = \begin{pmatrix} \frac{1}{C} & -\frac{1}{C} \\ -\frac{1}{C} & \frac{1}{C} \end{pmatrix}$, $\mathrm{Sgn}(Cx) = (\mathrm{sgn}(x_1)\,\mathrm{sgn}(x_2))^T$, $B = \begin{pmatrix} V_z & 0 \\ 0 & V_z \end{pmatrix}$, $C = I_2$ the identity matrix. One remarks that the condition (2.91) is trivially satisfied with $P = B^{-1}$. The multivalued mapping $x \mapsto B\,\mathrm{Sgn}(Cx)$ is maximal monotone. The system is already under the canonical form in (2.49) and Proposition 2.58 applies.

2.5.9 The RCZD Circuit in (1.11)

The voltage/current law $v(t) \in \mathscr{F}_z(i(t))$ in (1.11) may be rewritten using the subdifferential of the convex lower semi-continuous proper function $f(i) = \begin{cases} -V_z i & \text{if } i \leqslant 0 \\ 0 & \text{if } i \geqslant 0 \end{cases}$ (see Fig. 2.22). We obtain that $\mathscr{F}_z(i) = \partial f(i)$. From (1.11) we deduce that

$$v(t) \in \partial f\left(-\frac{1}{RC}x(t) + \frac{u(t)}{R} - \frac{v(t)}{R}\right), \tag{2.119}$$

that is a generalized equation. Since $f(\cdot)$ is convex proper lower semi-continuous, this is equivalent to:

$$0 \in \frac{1}{RC}x(t) - \frac{u(t)}{R} + \frac{v(t)}{R} + \partial f^*(v(t)) = N_{[-V_z,0]}(v(t)), \tag{2.120}$$

where we made use of (2.9) to pass from (2.119) to (2.120) (see also Fig. 2.11). The last equality should be obvious from Fig. 1.8(a) and from Fig. 2.11. Since $R > 0$ it follows that the mapping $v \mapsto \frac{v}{R}$ is strongly monotone. From Theorem 2.35 one infers that the generalized equation (2.120) has a unique solution. In Chap. 1 we studied this generalized equation in a graphical way, see Fig. 1.9.

Now let us rewrite (2.120) as:

$$\frac{1}{RC}x(t) - \frac{u(t)}{R} + \frac{v(t)}{R} \in -N_{[-V_z,0]}(v(t)). \tag{2.121}$$

Using Proposition 2.37 we deduce that:

$$v(t) = \mathrm{proj}\left([-V_z,0]; -\frac{1}{RC}x(t) + \frac{u(t)}{R}\right). \tag{2.122}$$

Inserting (2.122) into (1.11) one finds that the dynamics of this circuit is an ordinary differential equation with Lipschitz continuous right-hand-side.

2.5.10 The Circuit in (1.41)

2.5.10.1 Embedding into Differential Inclusions

First of all it follows from (2.23) (or from (2.25)) that the linear complementarity system in (1.41) can be rewritten as the differential inclusion:

$$\begin{cases} \dot{x}_1(t) = x_2(t) - \frac{1}{RC}x_1(t), \\ \dot{x}_2(t) \in -\frac{1}{LC}x_1(t) - \partial\psi_{\mathbb{R}^-}(x_2(t)), \end{cases} \tag{2.123}$$

where $\psi_{\mathbb{R}^+}(\cdot)$ is the indicator function of $\mathbb{R}^+$, and we used several tools from convex analysis: the equivalence (2.23) and Proposition 2.29. This allows us to transform the complementarity $0 \leqslant v(t) \perp -x_2(t) \geqslant 0$ into $-v(t) \in \partial\psi_{\mathbb{R}^+}(-x_2(t))$. Letting $f(x_2) \overset{\Delta}{=} \psi_{\mathbb{R}^+}(-x_2)$ we get $\partial f(x_2) = -\partial\psi_{\mathbb{R}^+}(-x_2)$ and since $f(x_2) = \psi_{\mathbb{R}^-}(x_2)$ we obtain that $v(t) \in \partial\psi_{\mathbb{R}^-}(x_2(t))$. Thus for obvious definitions of the matrices A, B and C[9] we may rewrite the system (2.123) as:

$$\dot{x}(t) - Ax(t) \in -BN_{\mathbb{R}^-}(Cx(t)), \tag{2.124}$$

with the state vector $x^T = (x_1\ x_2)$. For such a circuit it may be checked that the "input-output" relation (2.91) is satisfied trivially because $B = C^T$. Therefore using again Proposition 2.29 we infer that there exists a proper convex lower semi-continuous function $g(\cdot)$ such that $\partial g(x) = BN_{\mathbb{R}^-}(Cx(t))$. Using Theorem 2.34

[9] The matrix C in (2.124) is not to be confused with the capacitor value in (2.123).

it follows that the multivalued operator $x \mapsto \partial g(x)$ is maximal monotone. Using again Proposition 2.29 we infer that $g(x) = N_K(x)$ where $K = \{x \in \mathbb{R}^2 \mid Cx \leqslant 0\}$ is a convex set. Therefore we can rewrite (2.124) as:

$$\dot{x}(t) - Ax(t) \in -N_K(x(t)) \tag{2.125}$$

which fits within (2.49) so that Proposition 2.58 applies. Notice that the condition $x_0 \in \mathrm{dom}(A)$ of Proposition 2.58 translates into $x_2(0) \leqslant 0$ for our circuit. If $x_2(0^-) > 0$ then a jump has to be applied initially to the state, according to Proposition 2.65. In such a case the right mathematical formalism for (2.123) is that of a measure differential inclusion:

$$dx - Ax(t)dt \in -N_K(x(t)) \tag{2.126}$$

and the solution has to be understood in the sense of Definition A.7. In particular at an atom t of the differential measure dx one obtains $x(t^+) - x(t^-) \in -N_K(x(t^+))$ and it follows from (2.14) that $x(t^+) = \mathrm{proj}(K; x(t^-))$. Notice that we wrote $x(t^+)$ in the normal cone argument, because the solution is right-continuous, see Definition A.7. Therefore within the framework of measure differential inclusions one has $x(t) = x(t^+)$.

2.5.10.2 Linear Complementarity Problems

Let us now consider this system from another point of view. Let us assume that on some time interval $[t_1, t_2]$, $t_1 < t_2$, one has $x_2(t) = 0$ for all $t \in [t_1, t_2]$. Let us first construct an LCP which allows us to compute $\dot{x}_2(t)$ at any time t inside $[t_1, t_2]$ (in fact we are interested mainly by what happens on the right of $t = t_2$ since we suppose that $x_2(\cdot)$ is identically zero on the whole interval). From (1.41) it follows that $v(t) = -L\dot{x}_2(t) - \frac{1}{C}x_1(t)$. Since $x_2(t) = 0$ and the state is continuous, it follows that the complementarity $0 \leqslant v(t) \perp -x_2(t) \geqslant 0$ implies:

$$0 \leqslant v(t) \perp -\dot{x}_2(t) \geqslant 0. \tag{2.127}$$

Indeed if $-\dot{x}_2(t) < 0$ it follows from Proposition 7.1.1 in Glocker (2001) (see also Proposition C.8 in Acary and Brogliato 2008) that $x_2(\tau) > 0$ in a right neighborhood of t, which is forbidden. Moreover if $-\dot{x}_2(t) > 0$ then by the same proposition it follows that $x_2(\tau) < 0$ in a right neighborhood of t, and therefore $v(\tau) = 0$ in this neighborhood. Consequently the complementarity between v and $\dot{x}_2$ holds as well.

Starting from (2.127) it easily follows:

$$0 \leqslant -L\dot{x}_2(t) - \frac{1}{C}x_1(t) \perp -\dot{x}_2(t) \geqslant 0, \tag{2.128}$$

which is an LCP with unknown $-\dot{x}_2(t)$. From Theorem 2.43 this LCP has a unique solution, which can be found by simple inspection:

(i) if $x_1(t) < 0$ then $-\dot{x}_2(t) = 0$: the trajectory stays on the boundary;
(ii) if $x_1(t) > 0$ then $-\dot{x}_2(t) = \frac{1}{LC}x_1(t) > 0$: the trajectory leaves the boundary;
(iii) if $x_1(t) = 0$ then $-\dot{x}_2(t) = 0$: this is a degenerate case.

Now notice that we may instead work with the multiplier $v(t)$ and rewrite the LCP (2.128) as:

$$0 \leqslant \frac{1}{C}x_1(t) + \frac{1}{L}v(t) \perp v(t) \geqslant 0. \tag{2.129}$$

Then we have:

(i) if $x_1(t) < 0$ then $v(t) = -\frac{1}{LC}x_1(t) > 0$ and from the dynamics $-\dot{x}_2(t) = 0$: the trajectory stays on the boundary;
(ii) if $x_1(t) > 0$ then $v(t) = 0$ and from the dynamics $-\dot{x}_2(t) = \frac{1}{LC}x_1(t) > 0$: the trajectory leaves the boundary;
(iii) if $x_1(t) = 0$ then $v(t) = 0$ and from the dynamics $-\dot{x}_2(t) = 0$: this is a degenerate case.

One may therefore work with either LCP in (2.128) or in (2.129) and reach the same conclusions.

2.5.10.3 Some Comments

For such a simple system both the differential inclusion and the complementarity formalisms may be used to design a backward Euler numerical scheme, as done in Chap. 1 for several circuits, and in Sects. 2.6.1 and 2.6.2 in a more general setting. The obtained set of discrete-time equations boils down to solving an LCP at each time step. If the trajectory is in a contact mode as in Sect. 2.5.10.2, the LCP solver takes care of possible "switching" between the contact and the non-contact modes. The material in Sect. 2.5.10.2 is useful when one wants to use an event-driven numerical method. From the knowledge of the state vector x_k at some discrete time t_k, and under the condition that $x_2(t_k) = 0$, one then constructs an LCP as in (2.128) or (2.129) to advance the method. The LCP that results from the time-stepping backward Euler method in (2.142) and the event-driven LCP obtained from (2.129), are obviously not equal one to each other.

2.5.11 The Switched Circuit in (1.52)

Concerning a piecewise–linear system as in (1.52), one has to know whether the vector field is continuous or discontinuous on the switching surface that is defined here by the boundary bdχ that separates χ_1 and χ_2. At the points x such that $x \in \partial\chi$ and $A_1x \neq A_2x$, something has to be done. One solution is to embed the right-hand-side into Filippov's sets (see Sect. 2.4.4), so as to obtain a Filippov's differential inclusion.

If one considers the piecewise–linear system in (1.51) where the triggering signal $u_c(t)$ is purely exogenous, the picture is different. The system is then a non-autonomous system (due to the exogenous switches). One may assume that $u_c(t)$ is such that the switching instants satisfy $t_{k+1} > t_k + \delta$ for some $\delta > 0$. An ambiguity still remains in (1.51) because the right-hand-side is not specified when $u_c = 0$. One may choose to write the right-hand-side as a convex combination of $A_1 x(t)$ and $A_2 x(t)$ if t corresponds to a switching instant.

2.5.12 Well-Posedness of the OSNSP in (1.45)

The OSNSP in (1.45) possesses a unique solution x_{k+1} at each step k, for any data. To prove this, let us transform the system (1.42), that is written compactly as

$$\begin{cases} \dot{x}(t) = Ax(t) - Bv(t) + Eu(t), \\ v(t) \in \mathscr{F}(w(t)), \\ w(t) = Cx(t). \end{cases} \tag{2.130}$$

A key property of the pair (B, C) is that there exists a 3×3 matrix $P = P^T >$ such that:

$$PB_1 = C_1^T, \qquad PB_2 = C_2^T, \tag{2.131}$$

where B_1 and B_2 are the two columns of B, C_1 and C_2 are the two rows of C. The matrix P is given by

$$P = \begin{pmatrix} \frac{1}{C} & 0 & 0 \\ 0 & L_1 & 0 \\ 0 & 0 & L_2 \end{pmatrix}.$$

Let us consider the symmetric positive definite square root of P, *i.e.* $R = R^T >$ and $R^2 = P$. Let us perform the state vector change $z = Rx$. The system in (2.130) can be rewritten as:

$$\begin{cases} \dot{z}(t) = RAR^{-1}z(t) - RBv(t) + REu(t), \\ v_1(t) \in \mathscr{F}_1(C_1 x(t)), \qquad v_2(t) \in \mathscr{F}_2(C_2 x(t)), \end{cases} \tag{2.132}$$

with obvious definitions of $\mathscr{F}_1(\cdot)$ and $\mathscr{F}_2(\cdot)$ from (1.42). A key property of the multivalued functions $\mathscr{F}_i(\cdot)$ is that there exist proper convex lower semi-continuous functions $\varphi_i(\cdot)$ such that $\mathscr{F}_i(\cdot) = \partial\varphi_i(\cdot)$. These functions are given by $\varphi_1(x) = \begin{cases} -V_z x & \text{if } x < 0 \\ 0 & \text{if } x > 0 \end{cases}$, and $\varphi_2(x) = \psi_K(x)$ with $K = \mathbb{R}^+$. We may rewrite (2.132) as:

$$\begin{cases} \dot{z}(t) = RAR^{-1}z(t) - RB_1 v_1(t) - RB_2 v_2(t) + REu(t), \\ v_1(t) \in \partial\varphi_1(C_1 R^{-1} z(t)), \qquad v_2(t) \in \partial\varphi_2(C_2 R^{-1} z(t)). \end{cases} \tag{2.133}$$

Using (2.131) the terms $RB_1 v_1$ and $RB_2 v_2$ may be rewritten as $R^{-1}C_1^T v_1$ and $R^{-1}C_2^T v_2$, respectively. Using the inclusions in (2.133) one obtains the two terms $R^{-1}C_1^T \partial\varphi_1(C_1 R^{-1} z)$ and $R^{-1}C_2^T \partial\varphi_2(C_2 R^{-1} z)$. Now we may use Proposition 2.29 to deduce that $R^{-1}C_1^T \partial\varphi_1(C_1 R^{-1} z) = \partial(\varphi_1 \circ C_1 R^{-1})(z)$ and $R^{-1}C_2^T \partial\varphi_2(C_2 R^{-1} z) = \partial(\varphi_2 \circ C_2 R^{-1})(z)$. Let us denote $\varphi_1 \circ C_1 R^{-1}(\cdot) = \phi_1(\cdot)$ and $\varphi_2 \circ C_2 R^{-1}(\cdot) = \phi_2(\cdot)$, and $\Phi(\cdot) = \phi_1(\cdot) + \phi_2(\cdot)$. A key property is that since the functions $\phi_1(\cdot)$ and $\phi_2(\cdot)$ are proper convex lower semi-continuous, then the multivalued mapping $\partial\Phi(\cdot) = \partial\phi_1(\cdot) + \partial\phi_2(\cdot)$ is maximal monotone (see Theorem 2.30 and the properties in Sect. 2.1.2.2). Introducing this in the first line of (2.133) we obtain:

$$-\dot{z}(t) + RAR^{-1}z(t) + REu(t) \in \partial\Phi(z(t)), \tag{2.134}$$

that is equivalent to (2.130) in the sense that if $x(\cdot)$ is a solution of (2.130) then $z = Rx$ is a solution of (2.134), and *vice-versa*. Let us now proceed with the implicit Euler discretization of the transformed differential inclusion (2.134). We obtain:

$$-z_{k+1} + z_k + hRAR^{-1}z_{k+1} + hREu_{k+1} \in h\partial\Phi(z_{k+1}), \tag{2.135}$$

which we rewrite as

$$0 \in (I_3 - hRAR^{-1})z_{k+1} + z_k + hREu_{k+1} + h\partial\Phi(z_{k+1}). \tag{2.136}$$

It is noteworthy that the generalized equation (2.136) is strictly equivalent to the generalized equation (1.45). However it is now in a more suitable form $0 \in F(z_{k+1}) = Mz_{k+1} + q_k + h\partial\Phi(z_{k+1})$, where M is positive definite for sufficiently small $h > 0$ and $h\partial\Phi(\cdot)$ is maximal monotone. It follows that the multivalued mapping $F(\cdot)$ is strongly monotone, and from Theorem 2.35 the generalized equation $0 \in F(z_{k+1})$ has a unique solution. We have thus proved the following:

Lemma 2.81 *Let $h > 0$ be sufficiently small so that $(I_3 - hRAR^{-1})$ is positive definite. The OSNSP in (1.45) has a unique solution for any data x_k and u_{k+1}.*

The arguments that we used to prove Lemma 2.81 generalize those which we used to study the OSNSP in (1.18) and (1.15). As an illustration let us consider the OSNSP in (1.18). Using the equivalence in (2.23) it may be rewritten as

$$0 \in hv_{k+1} + q_k + N_K(v_{k+1}), \tag{2.137}$$

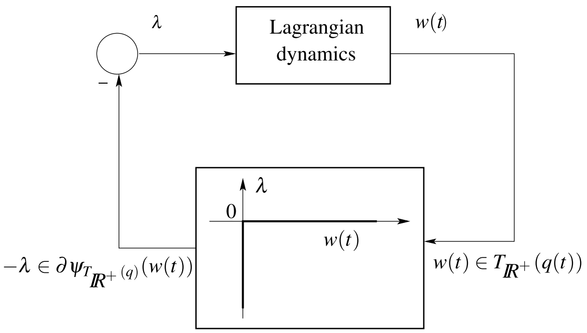
with $q_k = (1 - h\frac{R}{L})x_k - i_{k+1}$ and $K = \mathbb{R}^+$. Since the normal cone to a convex non empty set defines a maximal monotone mapping and since $h > 0$, the proof follows.

2.5.13 The Bouncing Ball

Let us come back on the dynamics in (1.58). This may be recast into the Lagrangian sweeping process (2.38):

$$\begin{cases} mdv + mgdt + u(t)dt = \lambda, \\ -\lambda \in N_{T_{\mathbb{R}^+}(q(t))}(w(t)). \end{cases} \tag{2.138}$$

The interpretation as the negative interconnection of two blocks is then clear. The first block of Fig. 2.26 is the Lagrangian dynamics with input λ and output $w(t)$. The second block is the nonsmooth part due to the unilateral constraint and the impact law. It is fed by $w(t)$, and its output is $-\lambda$. The analogy between Figs. 2.26 and 1.18 is clear. Another example showing the analogy between Mechanics (with Coulomb friction) and circuits (with Zener diodes) is worked in Sect. 2.5.4.



λ
Lagrangian dynamics
$w(t)$
$-$
λ
0
$w(t)$
$-\lambda \in \partial\psi_{T_{\mathbb{R}^+}(q)}(w(t))$
$w(t) \in T_{\mathbb{R}^+}(q(t))$

Fig. 2.26 Bouncing-ball feedback interconnection with the corner law

Notice that (2.138) is a measure differential inclusion, so that λ and dv are differential measures as defined in Appendix A.5. The solution has to be understood as in Definition A.7. The first line in (2.138) is therefore an equality of measures which we may write as $d\mu = \lambda$. At an impact time t one has $d\mu(\{t\}) = dv(\{t\}) = v(t^+) - v(t^-)$ and $dt(\{t\}) = 0$ (see Sect. A.5). The measure λ thus has a density p with respect to the Dirac measure δ_t and we obtain $p = m(v(t^+) - v(t^-))$. Going on as in (2.40) through (2.42) one recovers the restitution law in (1.58).

The negative feedback interconnection of Fig. 2.26 shows that the bouncing ball may be interpreted as a Lur'e system: the Lagrangian dynamics defines a dissipative subsystem, and the feedback path is a maximal monotone operator. The advantage of Moreau's sweeping process is that it allows one to represent the nonsmooth dynamics in one shot, without requiring any "hybrid-like" point of view. The stability Brogliato (2004) and the time-discretization method (Acary and Brogliato 2008) follow from it.

Remark 2.82 Compare (2.105), or (2.106), with (2.138). In (2.106) the multivalued part is a normal cone to a time varying set. In (2.138) the multivalued part is a normal cone to a state-dependent set (a tangent cone). So if $i(\cdot)$ is a constant in (2.106) the multivalued part of the inclusion that represents the electrical circuit is just a normal cone to a constant convex set. In the case of the bouncing ball the set remains state-dependent even if $u(t) = 0$.

2.6 Time-Discretization Schemes

In Chap. 1 the backward Euler method has been introduced on the simple examples which are studied. An insight on how the sliding trajectories that evolve on attractive switching surfaces are simulated, is given in Fig. 1.7. This can be generalized to more complex systems, as shown in Acary and Brogliato (2010). The numerical schemes that will be used in the next chapters of this book are some extensions of the backward Euler method. In this part let us focus on the implicit (or backward) Euler scheme only. Since the objective of this book is more about "practical" numerics than pure numerical analysis, only few results of convergence will be given in this section. The first result concerns the maximal monotone differential inclusions in (2.49), the second result is for linear complementarity systems (LCS) as in (2.53), and the third result concerns Moreau's sweeping process in (2.36).

2.6.1 Maximal Monotone Differential Inclusions

Let $T > 0$. The differential inclusion (2.49) is time-discretized on $[0, T]$ with a backward Euler scheme as follows:

$$\begin{cases} \frac{x_{k+1}-x_k}{h} + A(x_{k+1}) \ni f(t_k, x_k), & \text{for all } k \in \{0, \dots, N-1\}, \\ x_0 = x(0), \end{cases} \tag{2.139}$$

where $h = \frac{T}{N}$. The fully implicit method uses $f(t_{k+1}, x_{k+1})$ instead of $f(t_k, x_k)$. The convergence and order results stated in Proposition 2.83 below have been derived for the semi-implicit scheme (2.139) in Bastien and Schatzman (2002). So the analysis in this section is based on such a discretization. However this is only a particular case of a more general θ-method which is used in practical implementations. The next result is proved in Bastien and Schatzman (2002).

Proposition 2.83 *Under Assumption 2.57,*[10] *there exists η such that for all $h > 0$ one has*

$$\textit{For all } t \in [0, T], \quad \|x(t) - x^N(t)\| \leqslant \eta\sqrt{h}. \tag{2.140}$$

Moreover $\lim_{h\to 0^+} \max_{t\in[0,T]} \|x(t) - x^N(t)\|^2 + \int_0^t \|x(s) - x^N(s)\|^2 ds = 0.$

Thus the numerical scheme has at least order $\frac{1}{2}$, and convergence holds.

2.6.2 Linear Complementarity Systems

Let us consider the LCS in (2.53). Its backward Euler discretization is:

$$\begin{cases} x_{k+1} = x_k + hAx_{k+1} + hB\lambda_{k+1}, \\ w_{k+1} = Cx_{k+1} + D\lambda_{k+1}, \\ 0 \leqslant \lambda_{k+1} \perp w_{k+1} \geqslant 0. \end{cases} \tag{2.141}$$

Easy manipulations yield $x_{k+1} = (I_n - hA)^{-1}(x_k + hB\lambda_{k+1})$ where we assume that h is small enough to guarantee that $I_n - hA$ is an invertible matrix. Inserting this into the complementarity conditions leads to the LCP:

$$0 \leqslant \lambda_{k+1} \perp C(I_n - hA)^{-1}(x_k + hB\lambda_{k+1}) + D\lambda_{k+1} \geqslant 0, \tag{2.142}$$

with unknown λ_{k+1} and LCP matrix $hC(I_n - hA)^{-1}B + D$. As one may guess a lot depends on whether or not this LCP possesses a unique solution.

[10] See Sect. 2.4.2.

Assumption 2.84 *There exists $h^* > 0$ such that for all $h \in (0, h^*)$ the LCP(M, b_{k+1}) has a unique solution for all b_{k+1}.*

Assumption 2.85 *The system (A, B, C, D) is minimal (the pair (A, B) is controllable, the pair (C, A) is observable), and B is of full column rank.*

The approximation of the Dirac measure at $t = 0$ is given by $h\lambda_0 \approx \delta_0$. Assumption 2.84 secures that the one-step-nonsmooth-problem algorithm to solve the LCP generates a unique output at each step, for $h > 0$ small enough.

Let us now state a convergence result taken from Camlibel et al. (2002a). The interval of integration is $[0, T]$, $T > 0$. The convergence is understood as $\lim_{h\to 0}\langle x^N(t) - x(t), \varphi(t)\rangle = 0$ for all $\varphi \in \mathscr{L}^2([0, T]; \mathbb{R}^n)$ and all $t \in [0, T]$, which is the weak convergence in $\mathscr{L}^2([0, T]; \mathbb{R}^n)$.

Theorem 2.86 *Consider the LCS in (2.53) with $D \geqslant 0$ and let Assumption 2.84 hold. Let $(\lambda_k^N, x_k^N, w_k^N)$ be the output of the one-step-nonsmooth-problem solver, with the initial impulsive term being approximated by $(h\lambda_0, hx_0, hw_0)$. Assume that there exists a constant $\alpha > 0$ such that for $h > 0$ small enough,one has $\|h\lambda_0\| \leqslant \alpha$ and $\|\lambda_k^N\| \leqslant \alpha$ for all $k \geqslant 0$. Then for any sequence $\{h_k\}_{k\geqslant 0}$ that converges to zero, one has:*

(i) *There exists a subsequence $\{h_{k_l}\} \subseteq \{h_k\}_{k\geqslant 0}$ such that $(\{\lambda^N\}_{k_l}, \{w^N\}_{k_l})$ converges weakly to some (λ, w) and $\{x^N\}_{k_l}$ converges to some $x(\cdot)$.*
(ii) *The triple $(\lambda, x(\cdot), w)$ is a solution of the LCS in (2.53) on $[0, T]$ with initial data $x(0) = x_0$.*
(iii) *If the LCS has a unique solution for $x(0) = x_0$, the whole sequence $(\{\lambda^N\}_k, \{w^N\}_k)$ converges weakly to (λ, w) and the whole sequence $\{x^N\}_k$ converges to $x(\cdot)$.*

If the quadruple (A, B, C, D) is such that Assumption 2.85 holds and is passive, then (iii) *holds.*

We emphasize the notation $x(\cdot)$ since the solutions are functions of time, whereas the notation λ and w means that these have to be considered as measures. Other results of convergence for the case $D = 0$ can be found in Shen and Pang (2007, Theorem 7), under the condition that the Markov parameter CB satisfies some relaxed positivity conditions (a condition similar to the property in (2.91) which implies that $CB = B^T PB \geqslant 0$).

Remark 2.87 What happens when the system to be simulated does not enjoy the uniqueness of solutions property? Let us consider for instance the Filippov's differential inclusion in (2.74) with $g(t) \equiv 0$, which has three solutions starting from $x(0) = 0$, $x(t) \equiv 0$, $x(t) = t$ and $x(t) = -t$. Its implicit Euler discretization is:

$$x_{k+1} - x_k \in h\,\mathrm{sgn}(x_{k+1}). \tag{2.143}$$

In Fig. 2.27, we can study this generalized equation graphically as we did in Fig. 1.7.

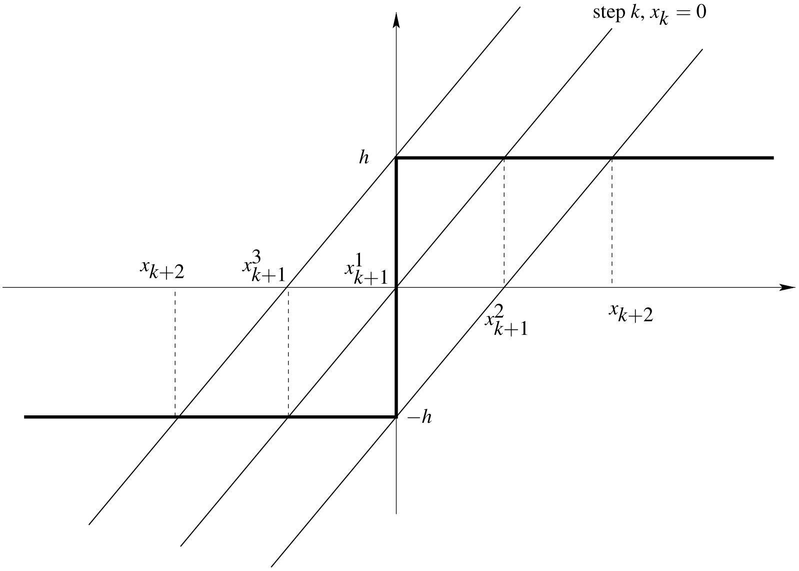


Fig. 2.27 Iterations for (2.143)

At step k there are three intersections (solutions of the generalized equation): $x^1_{k+1} = 0$, $x^2_{k+1} = h$ and $x^3_{k+1} = -h$. At step $k+1$ starting from x^2_{k+1} or x^3_{k+1} there are two solutions: $x^{2,1}_{k+2} = 0$ or $x^{2,2}_{k+2} = 2h$, and $x^{3,1}_{k+2} = 0$ or $x^{3,2}_{k+2} = -2h$. After that the solutions are unique. We conclude from this simple example that despite non-uniqueness holds, the backward Euler method still performs well in the sense that its output is made of three approximated solutions: one that stays around zero and two that diverge as t. In practice either the implemented solver chooses one of them more or less randomly, or the designer has to add some criterion that obliges the method to choose a particular solution out of the three. Similar conclusions have been obtained in an event-driven method context in Stewart (1990, 1996).

2.6.3 Moreau's Sweeping Process

We shall focus in this section on a basic result that was obtained by Moreau (1977) for sweeping processes of bounded variations with $f(t,x) = 0$ in (2.36). Generalizations for the case where the perturbation is not zero, even multivalued, exist (Edmond and Thibault 2006) which are based on the same type of approximation. Let us therefore consider the differential inclusion:

$$-dx \in N_{C(t)}(x(t)), \quad x(0) = x_0, \tag{2.144}$$

where the set-valued map $t \mapsto C(t)$ is either absolutely continuous, or Lipschitz continuous in the Hausdorff distance, or right-continuous of bounded variation, see

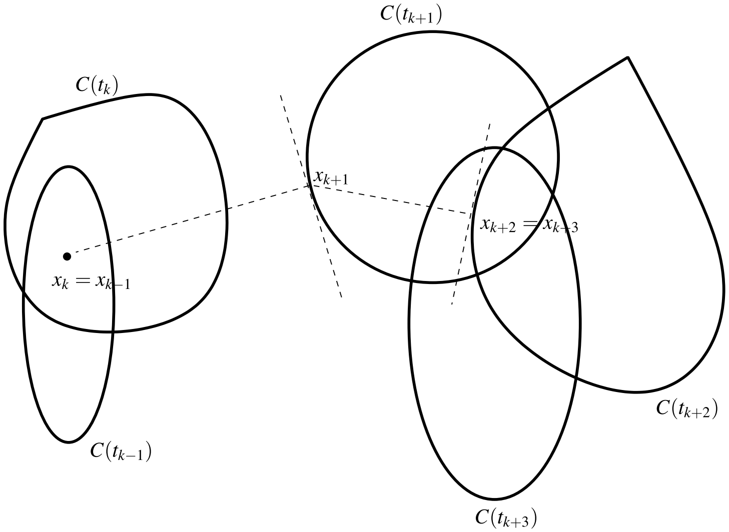


Fig. 2.28 The catching-up algorithm

Sects. A.1, A.2 and A.4 for the definitions. When the solution is absolutely continuous, then $dx = \dot{x}(t)dt$, and since the right-hand-side is a cone, the left-hand-side may be simplified to $-\dot{x}(t)$. Under suitable hypothesis on the multivalued function $t \mapsto C(t)$, numerous convergence and consistency results (Monteiro Marques 1993; Kunze and Monteiro Marquès 2000) have been given together with well-posedness results, using the so-called "Catching-up algorithm" defined in Moreau (1977):

$$-(x_{k+1} - x_k) \in \partial \psi_{C(t_{k+1})}(x_{k+1}), \tag{2.145}$$

where x_k stands for the approximation of the right limit of $x(\cdot)$. It is noteworthy that the case with a Lipschitz continuous moving set is also discretized in the same way.

By elementary convex analysis (see (2.25) or (2.14)), the inclusion (2.145) is equivalent to:

$$x_{k+1} = \text{prox}[C(t_{k+1}); x_k]. \tag{2.146}$$

Contrary to the standard backward Euler scheme with which it might be confused, the catching-up algorithm is based on the evaluation of the measure dx on the interval $(t_k, t_{k+1}]$, i.e. $dx((t_k, t_{k+1}]) = x^+(t_{k+1}) - x^+(t_k)$. Indeed, the backward Euler scheme is based on the approximation of $\dot{x}(t)$ which is not defined in a classical sense for our case. When the time step vanishes, the approximation of the measure dx tends to a finite value corresponding to the jump of $x(\cdot)$. This remark is crucial for the consistency of the scheme. Particularly, this fact ensures that we handle only finite values.

Figure 2.28 depicts the evolution of the discretized sweeping process. The name *catching-up* is clear from the figure: the algorithm makes x_k catch-up with the moving set $C(t_k)$, so that it stays inside the moving set.

We give below a brief account on the properties of the discretized sweeping process. More may be found in Monteiro Marques (1993) and Kunze and Monteiro Marquès (2000). Let us first deal with the Lipschitz continuous sweeping process.

Theorem 2.88 *Suppose that the mapping $t \mapsto C(t)$ is Lipschitz continuous in the Hausdorff distance with constant l, and $C(t)$ is non empty, closed and convex for every $t \in [0, T]$. Let $x_0 \in C(0)$. Consider the algorithm in (2.145), with a fixed time step $h = \frac{T}{N} > 0$. Let $m \in \mathbb{N}$ be such that $mT < N$. Then:*

(a) $\mathrm{var}_{[0,T]}(x^N) \leqslant \|x^N(0)\| + lT$*, for all* $t \in [t_k, t_{k+1}]$ *and all* $N \in \mathbb{N}$,
(b) $\|x^N(t) - x^N(s)\| \leqslant l(|t-s| + \frac{2}{m})$*, for all* $t, s \in [t_k, t_{k+1}]$,
(c) *from which it follows that* $\|x(t) - x(s)\| \leqslant l|t-s|$ *for all* $t, s \in [0, T]$*, where* $(x(t) - x(s))$ *is the limit in the weak sense of* $\{x^N(t) - x^N(s)\}_{N\in\mathbb{N}}$,
(d) $\|\dot{x}^N(t)\| \leqslant l$ *for all* $t \neq t_k$*, where* $\dot{x}^N(t) = \frac{1}{h}(x_{k+1} - x_k)$ *for* $t \in [t_k, t_{k+1})$,
(e) *the "velocity"* $\dot{x}^N(\cdot)$ *converges weakly to* $\dot{x}^*(\cdot)$*, i.e. for all* $\varphi(\cdot) \in \mathscr{L}^1([0, T]; \mathbb{R}^n)$ *one has*

$$\int_0^T \langle \dot{x}^N(t), \varphi(t)\rangle dt \quad \to \quad \int_0^T \langle \dot{x}^*(t), \varphi(t)\rangle dt,$$

(f) $x^N(\cdot) \to x(\cdot)$ *uniformly and* $\dot{x}(\cdot) = \dot{x}^*(\cdot)$ *almost everywhere in* $[0, T]$,
(g) *the limit satisfies* $\dot{x}(t) \in N_{C(t)}(x(t))$ *almost everywhere in* $[0, T]$.

In the absolutely continuous and the bounded variations cases, the catching-up algorithm may be used also to prove Theorems 2.55 and 2.56, with similar steps as in Theorem 2.88. In the BV case the formalism has to be that of measure differential inclusions (see Moreau 1977, §3 for a proof of existence of solutions). This book is dedicated to electrical circuits, for more details on the numerical simulation of mechanical systems please see Acary and Brogliato (2008).

2.7 Conclusions and Recapitulation

Chapters 1 and 2 introduce simple examples of circuits with nonsmooth electronic devices, and the main mathematical tools one needs to understand, analyze and simulate them. Despite they possess simple topologies, these circuits are embedded into a variety of mathematical formalisms (some of which being equivalent):

- complementarity systems,
- Filippov's differential inclusions,
- differential inclusions with a maximal monotone multivalued part,
- dynamical variational inequalities,
- Moreau's sweeping processes (perturbed, first order),
- measure differential inclusions,
- piecewise-linear systems.

The solutions (*i.e.* the trajectories) of such systems ususally are absolutely continuous, or right-continuous of local bounded variations (with possible occurrence of jumps, *i.e.* state discontinuities). In the more general situation where the dynamical equations are obtained from an automatic equations generation tool, the dynamics will not exactly fit within these classes of multivalued systems, however, but will contain them as particular cases. Mainly because the obtained dynamics will contain equalities stemming from Kirschhoff's laws in current and voltage, which make it belong to the descriptor systems family.

As we have seen many of these circuits can be written as complementarity systems as in (2.57). A crucial parameter is the relative degree of the quadruplet (A, B, C, D). Let $u(t) = 0$ and let the initial data satisfy $Cx(0) + D\lambda(0) \geqslant 0$.

- If $r = 0$ the solutions are continuously differentiable ($D \neq 0$), see (1.3), (1.38), (1.39).
- If $r = 1$ the solutions are continuous ($D = 0$ and $CB \neq 0$), see (1.16), (1.40), (1.41).
- If $r = 2$ the solutions are discontinuous ($D = CB = 0$ and $CAB \neq 0$), see (1.58).
- If $r \geqslant 3$ the solutions are Schwarz' distributions (Dirac measure and its derivatives) ($D = CB = CAB = CA^{i-1}B = 0$ and $CA^{r-1}B \neq 0$), see (2.44) and Acary et al. (2008).

The solutions regularity is therefore intimately linked to the relative degree between the two slack variables.

Why such nonsmooth models? Mainly because conventional (say SPICE-like) solvers are not adequate for the analog simulation of switched circuits (see Maffezzoni's counterexample in Chap. 7, Sect. 7.1). This is advocated in many publications (Maffezzoni et al. 2006; Wang et al. 2009; Mayaram et al. 2000; Maksimovic et al. 2001; Valsa and Vlach 1995; Biolek and Dobes 2007; Lukl et al. 2006). On the other hand working with nonsmooth models implies to take into account inconsistent initial data treatment, and thus creates new challenges. NSDS takes care of all this and is a suitable solution for the simulation of circuits with a large number of events. The price to pay is low order on smooth portions of the state trajectories.

The NSDS method (which we could also name the Moreau-Jean's method (Jean 1999; Acary et al. 2010)) is a "package" which comprises:

- modeling with nonsmooth electronic devices (multivalued and piecewise-linear current/voltage characteristics),
- Moreau's time-stepping scheme (originally called the *catching-up* algorithm in the context of contact mechanics),
- OSNSP solvers (complementarity problems, quadratic programs).

In the remaining chapters of this book the NSDS method will be presented in detail.

Part II
Dynamics Generation and Numerical Algorithms

Chapter 3
Conventional Circuit Equation Formulation and Simulation

In this chapter, some basic facts on the circuit equation formulation and simulation which are shared by most of the analog SPICE-like simulators are presented. The formulation of the circuit equations is based on two basic ingredients:

- conservative laws given by the Kirchhoff laws in currents and voltages,
- constitutive equations of the electrical components,

which lead to a Differential Algebraic Equation (DAE). We will give some details on the DAEs in Sects. 3.1 to 3.6 and the chapter will end in Sect. 3.7 on conventional techniques for the numerical analog simulation of circuits.

3.1 Circuit Topology and Kirchhoff's Laws

3.1.1 The Circuit Network as a Connected Oriented Graph

Let us consider a circuit composed of b branches denoted by the branch set

$$\mathsf{B} = \{1, \dots, b\}, \tag{3.1}$$

and n nodes denoted by the node set

$$\mathsf{N} = \{1, \dots, n\}. \tag{3.2}$$

Let us assume for the moment that the circuit is composed of simple two-terminal (one-port) elements. The circuit topology is usually represented by a connected oriented graph. This graph is built by associating a vertex to each circuit node and an edge to each branch between two nodes. The orientation is arbitrarily chosen for each branch. In the sequel, we assume that the graph contains no self-loop (or loop in the sense of the graph theory) that is, there is no edge that connects a vertex to itself. A loop (in the sense of electrical network theory) is identified to a directed cycle in graph theory. Finally, in graph theory, a cut is a partition of the vertices of a graph into two disjoint subsets. The cut-set is the set of all edges whose end points are in the different sets generated by the cut.

V. Acary et al., *Nonsmooth Modeling and Simulation for Switched Circuits,*
Lecture Notes in Electrical Engineering 69,
DOI 10.1007/978-90-481-9681-4_3, © Springer Science+Business Media B.V. 2011

In order to write Kirchhoff's laws in currents and voltages, a rooted spanning tree is chosen. The set of branches is divided into two sets: the tree branches which corresponds to the edges of the spanning tree and the links which are the remaining edges of the connected graph. The number of tree branches in the spanning tree with n node is $n-1$, therefore the number of links is $b-n+1$. By definition of a tree, adding one link to the spanning tree forms a unique cycle. A basic loop or a mesh (in the sense of electrical network theory) is identified to each cycle generated by adding a link. There are therefore $b-n+1$ basic loops in the circuit which are oriented as the associated link. A basic cut of the circuit is associated to each branch of the spanning tree and the associated basic cut-set contains its tree branch and the corresponding links. There are $n-1$ basic cuts and the associated cut-sets.

In the following sections, some fundamental matrices associated with this connected oriented graph and a chosen rooted spanning tree are presented. Kirchhoff's laws in currents and branch voltages are expressed in terms of these fundamental matrices and this fact shows that these laws depend solely on the topology of the circuit.

3.1.2 The Incidence Matrix A and Kirchhoff's Current Laws

The first fundamental matrix is called the incidence matrix and is denoted by $A \in \mathbb{R}^{(n-1)\times b}$. This matrix A is an unimodular matrix which allows to construct the graph of the corresponding circuit. Each row i of A specifies if the branch j is connected to the node i. The value $A_{ij}=1$ specifies that the current of the branch j comes into the node i and the value $A_{ij}=-1$ when it leaves the node. If the branch j is not connected to the node i the value is set to $A_{ij}=0$. A column of A represents the terminal nodes of a given branch. Note that in our convention the incidence matrix consists of $n-1$ rows. A root node is arbitrarily chosen as a reference node (ground node) where the topology is not expressed, in order to obtain full row rank matrix.

It holds that $\operatorname{rank}(A)=n-1$ (even with the reference node), therefore the incidence matrix $A \in \mathbb{R}^{(n-1)\times b}$ of the circuit has full row rank. The incidence matrix A can be split into $[A_{\mathsf{Tree}}, A_{\mathsf{Link}}]$. The matrix $A_{\mathsf{Tree}} \in R^{(n-1)\times(n-1)}$ corresponds to the tree branches, it is square and invertible, and $A_{\mathsf{Link}} \in R^{(n-1)\times(b-n+1)}$ corresponds to the links.

By construction of the incidence matrix A, Kirchhoff's Current Laws (KCL) are given by

$$AI=0, \tag{3.3}$$

where $I \in \mathbb{R}^b$ is the vector composed of the current in each branch. The KCL express that the sum of currents at a node taking into account the orientation, must be zero.

3.1.3 The Loop Matrix B and Kirchhoff's Voltage Laws

The second fundamental matrix is called the loop matrix and is denoted by $B \in \mathbb{R}^{(b-n+1)\times b}$. The loop matrix B is a totally unimodular matrix and is built as follows. For each basic loop (as it has been defined above), the entry B_{ij} specifies if the branch j is concerned by the basic loop i. The value $B_{ij} = 1$ or $B_{ij} = -1$ is set depending on the orientation of the branch with respect to the given orientation of the basic loop. The value $B_{ij} = 0$ specifies that the branch is not concerned by the basic loop.

By construction of the incidence matrix B, the KVL are given by:

$$BU = 0, \tag{3.4}$$

where $U \in \mathbb{R}^b$ is the vector composed of the branch voltages. The KVL express that the sum of branch voltages along a basic loop, taking into account the orientation, must be zero.

3.1.4 KVL in Terms of Nodes Voltages

A fundamental relation in standard graph theory asserts that

$$AB^T = 0 \quad \text{or} \quad BA^T = 0. \tag{3.5}$$

By choosing the node voltages $V \in \mathbb{R}^{n-1}$ except on the reference node (ground node), such that:

$$A^T V = U, \tag{3.6}$$

the KVL are automatically satisfied due to (3.5). Indeed, one has

$$BU = BA^T V = 0. \tag{3.7}$$

In the sequel, the equation (3.6) will be preferably used to express the KVL. This is mainly due to the fact that the incidence matrix is usually directly given by the user input and the form (3.6) does not need the computation of the loop matrix B.

Others fundamental matrices can also be introduced in the study of the network topology, such as the node-to-reference path matrix and the cut-set matrix. These matrices are interesting in more thorough studies of the circuits as in the index investigation of the generated equations (Tischendorf 1999; Günther and Feldmann 1993, Bächle and Ebert 2005a, 2005b; Günther et al. 2005). For more details on network topology, we refer to Seshu and Reed (1961) and Branin (1967).

Remark 3.1 The case of multi-terminal elements can also be included in the equations (3.3), (3.4) or (3.6) for expressing the Kirchhoff laws. Clearly, the introduction of multi-terminal elements such as transistors avoids to speak of standard connected graphs. Nevertheless, equivalent circuits with only two-port elements can be formulated. Note for a multi-terminal element that the Kirchhoff laws are valid in the following sense. For t-terminal elements, the branch currents for each terminal and the branch voltages across any pair of terminals are well-defined. The sum of all branch currents flowing into the element and the sum of branch voltages along a closed terminal loop must be zero.

3.2 The Sparse Tableau Analysis (STA)

The Sparse Tableau Analysis (STA) (Hachtel et al. 1971) is one of the most basic ways to express the equations of a circuit. As we said before, Kirchhoff's Laws depend solely on the topology of the circuits. To complete the mathematical model, the physical behavior of the electrical components in each branch has to be characterized. In transient analysis, this behavior is described by differential equations of the unknowns, I, U and V which are called the Branch Constitutive Equation (BCE). The BCE for all branches can be formally written as:

$$F(I, U, \dot{I}, \dot{U}) = 0, \tag{3.8}$$

where $\dot{I}$ and $\dot{U}$ respectively denote the time-derivatives of I and U. Thanks to (3.6), the BCE can be equivalently written as

$$F(I, A^T V, \dot{I}, A^T \dot{V}) = 0. \tag{3.9}$$

In practice, the BCE for a particular branch involves only a subset of the unknowns. For instance, for standard two-terminal elements (resistive, capacitive or inductive elements), only the currents and voltages of the single branch are used to describe the behavior of the element.

The STA considers simply that the whole circuit equations are given by:

$$\begin{cases} AI = 0, \\ A^T V = U, \\ F(I, U, \dot{I}, \dot{U}) = 0. \end{cases} \tag{3.10}$$

The terminology of this approach is justified by the fact that it keeps the intrinsic sparsity of the description of the network. Let us consider for instance that the BCE are given a linear time invariant equation such that $F(I, U, \dot{I}, \dot{U}) = Z\,I + Y\,V = 0$. If we have only two-terminal elements, each raw of Z and respectively Y, will have a single non zero entry. The matrix of the linear system to be solved for (I, V, U) of the form

$$\begin{bmatrix} A & 0 & 0 \\ 0 & A^T & -I_d \\ Z & 0 & Y \end{bmatrix}, \tag{3.11}$$

will be therefore very sparse. More generally, the number of terminals of an element is usually limited and an equivalent linear time-invariant behavior will be obtained after the time-discretization and a Newton linearization around the current point. The sparsity of the description which is mainly due to the topology of the circuits will be conserved.

3.3 The Modified Nodal Analysis

In order to reduce the number of unknowns for a given circuit, Ho et al. (1975) introduced the Modified Nodal Analysis (MNA) which favors the nodal unknowns, i.e. the node voltages. Unlike the STA which can be used for very general circuits, some underlying assumptions have to be made. The main assumption states that the BCE can be explicitly written for a part of the unknowns in each branch. This leads to the following classification of branches.

3.3.1 Classification of the Branches

In the MNA, the branches are assumed to be classified in one of the following types:

1. current-defined branches denoted by the current-defined branch set $\mathsf{I} \subset \mathsf{B}$.
2. voltage-defined branches denoted by the voltage-defined branch set $\mathsf{U} \subset \mathsf{B}$.

Let us assume that $\mathsf{B} = \mathsf{I} \cup \mathsf{U}$ and $\mathsf{I} \cap \mathsf{U} = \emptyset$. For a current-defined branch $k \in \mathsf{I}$, the current I_k is defined as an explicit relation of the form

$$I_k = i_k(U, I_{\mathsf{U}}, \dot{U}, \dot{I}_{\mathsf{U}}, t), \tag{3.12}$$

where i_k is a given function which characterizes the branch. For a voltage-defined branch $k \in \mathsf{U}$, the voltage U_k is defined as an explicit relation of the form:

$$U_k = u_k(I_{\mathsf{U}}, U_{\mathsf{B}\backslash\{k\}}, \dot{I}_{\mathsf{U}}, \dot{U}_{\mathsf{B}\backslash\{k\}}, t), \tag{3.13}$$

where u_k is a given function which characterizes the branch.

The conventional MNA assumes that all the node potentials $\{V_k\}_{k\in\mathsf{N}\backslash\{0\}}$ and the currents in the voltage-defined branches, $I_{\mathsf{U}} = \{I_k, k \in \mathsf{U}\}$ form a sufficient set of unknowns to describe the circuit. The laws (3.12) in the current-defined branches are substituted in the KCL (3.3) and the laws in the voltage-defined branches (3.13) are kept in the set of equations. By splitting the incidence matrix with respect to current-defined and voltage-defined laws such that

$$A = [A_{\mathsf{I}}\ A_{\mathsf{U}}], \tag{3.14}$$

one gets:

$$\begin{cases} A_{\mathsf{I}} i(A^T V, I_{\mathsf{U}}, A^T \dot{V}, \dot{I}_{\mathsf{U}}, t) + A_{\mathsf{U}} I_{\mathsf{U}} = 0, \\ A_{\mathsf{U}}^T V - u(I_{\mathsf{U}}, A_{\mathsf{B}\backslash\{k\}}^T V_{\mathsf{B}\backslash\{k\}}, \dot{I}_{\mathsf{U}}, A_{\mathsf{B}\backslash\{k\}}^T \dot{V}_{\mathsf{B}\backslash\{k\}}, t) = 0, \end{cases} \tag{3.15}$$

where the functions $i(\cdot) = [i_k(\cdot), k \in \mathsf{I}]^T$ and $u(\cdot) = [u_k(\cdot), k \in \mathsf{U}]^T$ collect the functions respectively defining the current-defined and the voltage-defined branches.

3.3.2 Standard Resistive, Capacitive and Inductive Branches

It can be interesting to enter into more details of the description of the branches to structure the formulation (3.15). To this end, three main types of branches are usually considered:

1. Resistive elements in the branches indexed by $\mathsf{R} \subset \mathsf{I}$ which are characterized by branch constitutive equations of the form

$$I_k = S_k(U_\mathsf{R}, t), \quad \text{for all } k \in \mathsf{R} \subset \mathsf{I}. \tag{3.16}$$

   For these branches the conductance matrix $G(U_\mathsf{R}, t)$ can be defined as the Jacobian matrix of S with respect to U_R, that is

$$G(U_\mathsf{R}, t) = \nabla_{U_\mathsf{R}} S(U_\mathsf{R}, t). \tag{3.17}$$

2. Capacitive elements in the branches indexed by $\mathsf{C} \subset \mathsf{I}$ which are characterized by branch constitutive equations of the form

$$I_k = \frac{d}{dt} q_k(U_\mathsf{C}, t), \quad \text{for all } k \in \mathsf{C} \subset \mathsf{I}, \tag{3.18}$$

   where q_k is the charge into the capacitive branch k. Let q be the vector with entries $q_k, k \in \mathsf{C}$. The capacitance matrix, $C(U_\mathsf{C}, t)$ is generally defined by the Jacobian matrix of q with respect to U_C, that is

$$C(U_\mathsf{C}, t) = \nabla_{U_\mathsf{C}} q(U_\mathsf{C}, t). \tag{3.19}$$

3. Inductive elements in the branches indexed by $\mathsf{L} \subset \mathsf{U}$ which are characterized by branch constitutive equations of the form

$$U_k = \frac{d}{dt} \phi_k(I_\mathsf{L}, t), \quad \text{for all } k \in \mathsf{L} \subset \mathsf{U}, \tag{3.20}$$

   where ϕ_k is the flux in the inductive branch k. Let ϕ be the vector with entries $\phi_k, k \in \mathsf{L}$. The inductance matrix $L(I_\mathsf{L}, t)$ can be defined as the Jacobian matrix of ϕ with respect to I_L, that is

$$L(I_\mathsf{L}, t) = \nabla_{I_\mathsf{L}} \phi(I_\mathsf{L}, t). \tag{3.21}$$

Remark 3.2 In the remaining part of this book, the matrices $G(U_\mathsf{R}, t)$, $C(U_\mathsf{C}, t)$ and $L(I_\mathsf{L}, t)$ will be assumed to be positive-definite matrices. In some pathological cases, this assumption is not satisfied, for instance with some negative resistive elements. With only two-terminal elements, the matrices are diagonal. The symmetry of these matrices cannot generally be assumed with multi-terminal elements.

Usually, the remaining branches of the circuits are considered as voltage-controlled sources, and current-controlled sources and correspond respectively to the branches indexed by

$$\mathsf{V} = \mathsf{U} \setminus \mathsf{L} \tag{3.22}$$

and

$$\mathsf{J} = \mathsf{I} \setminus \{\mathsf{R} \cup \mathsf{C}\}. \tag{3.23}$$

The incidence matrix A can be split following the branch sets R, C, J, L and V such that:

$$A = [A_\mathsf{R}\ A_\mathsf{C}\ A_\mathsf{J}\ A_\mathsf{L}\ A_\mathsf{V}]\,. \tag{3.24}$$

The conventional MNA leads therefore to

$$\begin{cases} A_{\mathsf{C}}\frac{d}{dt}q(A_{\mathsf{C}}^T V,t) + A_{\mathsf{R}}S(A_{\mathsf{R}}^T V,t) + A_{\mathsf{L}}I_{\mathsf{L}} + A_{\mathsf{V}}I_{\mathsf{V}} \\ \quad + A_{\mathsf{J}}i(A^T V, I_{\mathsf{L}}, I_{\mathsf{V}}, A^T\dot{V}, \dot{I}_{\mathsf{L}}, \dot{I}_{\mathsf{V}}, t) = 0, \\ A_{\mathsf{L}}^T V - \frac{d}{dt}\phi(I_{\mathsf{L}},t) = 0, \\ A_{\mathsf{V}}^T V - u(I_{\mathsf{L}}, I_{\mathsf{V}}, A^T V, \dot{I}_{\mathsf{L}}, \dot{I}_{\mathsf{V}}, A^T\dot{V}, t) = 0. \end{cases} \tag{3.25}$$

Using the definitions of the capacitance matrix (3.19) and the inductance matrix (3.21), the time derivative of the charge q and the flux ϕ can be expressed as

$$\begin{aligned} \frac{d}{dt}q(A_{\mathsf{C}}^T V,t) &= C(U_{\mathsf{C}},t)A_C^T\frac{d}{dt}V + q_t(A_{\mathsf{C}}^T V,t), \\ \frac{d}{dt}\phi(I_{\mathsf{L}},t) &= L(I_{\mathsf{L}},t)\frac{d}{dt}I_{\mathsf{L}} + \phi_t(A_C^T V,t), \end{aligned} \tag{3.26}$$

where q_t, respectively ϕ_t, denotes the partial derivative of q, respectively ϕ,with respect to t. The system (3.25) can be simplified as follows

$$\begin{cases} A_{\mathsf{C}}C(A^T V,t)A_{\mathsf{C}}^T\frac{dV}{dt} + A_{\mathsf{C}}q_t(A_{\mathsf{C}}^T V,t) + A_{\mathsf{R}}S(A_{\mathsf{R}}^T V,t) \\ \quad + A_{\mathsf{L}}I_{\mathsf{L}} + A_{\mathsf{V}}I_{\mathsf{V}} + A_{\mathsf{J}}i(A^T V, I_{\mathsf{L}}, I_{\mathsf{V}}, A^T\dot{V}, \dot{I}_{\mathsf{L}}, \dot{I}_{\mathsf{V}}, t) = 0, \\ -A_{\mathsf{L}}^T V + L(I_{\mathsf{L}},t)\frac{dI_{\mathsf{L}}}{dt} + \phi_t(I_{\mathsf{L}},t) = 0, \\ A_{\mathsf{V}}^T V - u(I_{\mathsf{L}}, I_{\mathsf{V}}, A^T V, \dot{I}_{\mathsf{L}}, \dot{I}_{\mathsf{V}}, A^T\dot{V}, t) = 0. \end{cases} \tag{3.27}$$

3.4 The Charge/Flux Oriented MNA

In the framework of the MNA, another choice of the unknowns can be made by adding the charge of capacitors q, and the flux of the inductors ϕ in the unknown vector (Estèvez Schwarz and Tischendorf 2000; Günther et al. 2005). This results in the so-called charge/flux oriented MNA formulation:

$$\begin{cases} A_{\mathsf{C}}\frac{d}{dt}q(A_{\mathsf{C}}^T V,t) + A_{\mathsf{R}}S(A_{\mathsf{R}}^T V,t) + A_{\mathsf{L}}I_{\mathsf{L}} + A_{\mathsf{V}}I_{\mathsf{V}} \\ \quad + A_{\mathsf{J}}i(A^T V, I_{\mathsf{L}}, I_{\mathsf{V}}, A^T\dot{V}, \dot{I}_{\mathsf{L}}, \dot{I}_{\mathsf{V}}, t) = 0, \\ A_{\mathsf{L}}^T V - \frac{d}{dt}\phi(I_{\mathsf{L}},t) = 0, \\ A_{\mathsf{V}}^T V - u(I_{\mathsf{L}}, I_{\mathsf{V}}, A^T V, \dot{I}_{\mathsf{L}}, \dot{I}_{\mathsf{V}}, A^T\dot{V}, t) = 0, \\ q - q_C(A_{\mathsf{C}}^T V,t) = 0, \\ \phi - \phi_{\mathsf{L}}(I_{\mathsf{L}},t) = 0. \end{cases} \tag{3.28}$$

3.5 Standard DAEs Stemming from the MNA

In this section, the dynamics (3.27) which stems from the MNA is analyzed in the light of the DAE theory. We will not enter into a thorough description of what is a DAE and its fundamental properties. For more details, we refer the reader to Brenan et al. (1989), Hairer and Wanner (1996), and Ascher and Petzold (1998).

3.5.1 *Various Forms of DAEs*

3.5.1.1 Linear Time Invariant (LTI) Case with Independent Sources

In the case that the voltage-controlled and the current-controlled sources are some inputs which only depend on t, the sources are called independent sources. If we consider the Linear Time Invariant (LTI) case with independent sources, (3.27) can be written as

$$\begin{cases} A_{\mathsf{C}} C A_{\mathsf{C}}^T \frac{dV}{dt} + A_{\mathsf{R}} G A_{\mathsf{R}}^T V + A_{\mathsf{L}} I_{\mathsf{L}} + A_{\mathsf{V}} I_{\mathsf{V}} + A_{\mathsf{J}} i(t) = 0, \\ -A_{\mathsf{L}}^T V + L \frac{dI_{\mathsf{L}}}{dt} = 0, \\ A_{\mathsf{V}}^T V - u(t) = 0, \end{cases} \tag{3.29}$$

with the standard abuse of notation such that $C(A^T V, t) = C$, $L(I_{\mathsf{L}}, t) = L$ and $S(A_{\mathsf{R}}^T V, t) = G A_{\mathsf{R}}^T V$ in the LTI case. The system (3.29) forms the following LTI DAE for the unknowns $X = [V, I_{\mathsf{L}}, I_{\mathsf{V}}]$:

$$M\dot{X} = JX + U(t), \tag{3.30}$$

with

$$M = \begin{bmatrix} A_{\mathsf{C}} C A_{\mathsf{C}}^T & 0 & 0 \\ 0 & L & 0 \\ 0 & 0 & 0 \end{bmatrix}, \tag{3.31}$$

$$J = \begin{bmatrix} -A_{\mathsf{R}} G A_{\mathsf{R}}^T & -A_{\mathsf{L}} & -A_{\mathsf{V}} \\ A_{\mathsf{L}}^T & 0 & 0 \\ A_{\mathsf{V}}^T & 0 & 0 \end{bmatrix}, \tag{3.32}$$

$$U(t) = \begin{bmatrix} -A_{\mathsf{J}} i(t) \\ 0 \\ -u(t) \end{bmatrix}. \tag{3.33}$$

The system is most of the time a DAE and not only an implicit ODE. Indeed, the requirements to obtain an ODE is the invertibility of the matrix M. Obviously, if there are some voltage-controlled sources, then M cannot be invertible. When there is no voltage-controlled sources, the matrix M is also not necessarily invertible. This point will be discussed in Sect. 3.6.

3.5.1.2 Nonlinear Case with Independent Sources

With independent sources, the following nonlinear DAE is obtained from (3.27):

$$M(X, t)\dot{X} = D(X, t) + U(t) \tag{3.34}$$

where $X = [V^T, I_{\mathsf{L}}^T, I_{\mathsf{V}}^T]^T$ and

$$M(X,t) = \begin{bmatrix} A_{\mathsf{C}} C(A^T V, t) A_{\mathsf{C}}^T & 0 & 0 \\ 0 & L(I_{\mathsf{L}}, t) & 0 \\ 0 & 0 & 0 \end{bmatrix}, \tag{3.35}$$

$$D(X,t) = \begin{bmatrix} -A_{\mathsf{C}} q_t(A_{\mathsf{C}}^T V, t) - A_{\mathsf{R}} S(A_{\mathsf{R}}^T V, t) - A_{\mathsf{L}} I_{\mathsf{L}} - A_{\mathsf{V}} I_{\mathsf{V}} \\ A_{\mathsf{L}}^T V - \phi_t(I_{\mathsf{L}}, t) \\ A_{\mathsf{V}}^T V \end{bmatrix}, \tag{3.36}$$

$$U(t) = \begin{bmatrix} -A_{\mathsf{J}} i(t) \\ 0 \\ -u(t) \end{bmatrix}. \tag{3.37}$$

3.5.1.3 Nonlinear Case with General Controlled Sources

Up to this point, we have assumed that the current and voltage sources are independent sources, which are only given by some functions of time t. Under this assumption, the obtained DAE formulation (3.34) is linear in $\dot{X}$. This is one of the main interests of the MNA besides the classification of branches which allows one to drastically reduce the number of unknowns.

If the controlled current and voltage sources are general function of unknowns as they have been defined in (3.12) and (3.13), the structure of the DAE (3.34) is lost. In practice, the controlled current and voltage sources are assumed to be sufficiently smooth allowing us to use the linearized behavior around the current point $\bar{X} = [\bar{U}, \bar{I}_{\mathsf{L}}, \bar{I}_{\mathsf{V}}]^T$, $\bar{\dot{X}} = [\bar{\dot{U}}, \bar{\dot{I}}_{\mathsf{L}}, \bar{\dot{I}}_{\mathsf{V}}]^T$. Thanks to this linearization, the sources are inserted in the same formulation.

In the numerical practice, the linearization is used in the Newton-Raphson loop and leads to the evaluation of the Jacobian of the function $i(\cdot)$ in (3.12) (and respectively $u(\cdot)$ in (3.13)) with respect to their arguments:

$$\begin{aligned} A_{\mathsf{J}} i(U, I_{\mathsf{L}}, I_{\mathsf{V}}, \dot{U}, \dot{I}_{\mathsf{L}}, \dot{I}_{\mathsf{V}}, t) = {} & A_{\mathsf{J}} i(\bar{U}, \bar{I}_{\mathsf{L}}, \bar{I}_{\mathsf{V}}, \bar{\dot{U}}, \bar{\dot{I}}_{\mathsf{L}}, \bar{\dot{I}}_{\mathsf{V}}, \bar{t}) \\ & + A_{\mathsf{J}} \nabla_U i(\bar{U}, \bar{I}_{\mathsf{L}}, \bar{I}_{\mathsf{V}}, \bar{\dot{U}}, \bar{\dot{I}}_{\mathsf{L}}, \bar{\dot{I}}_{\mathsf{V}}, \bar{t})(A^T(V - \bar{V})) \\ & + A_{\mathsf{J}} \nabla_{I_{\mathsf{L}}} i(\bar{U}, \bar{I}_{\mathsf{L}}, \bar{I}_{\mathsf{V}}, \bar{\dot{U}}, \bar{\dot{I}}_{\mathsf{L}}, \bar{\dot{I}}_{\mathsf{V}}, \bar{t})(I_{\mathsf{L}} - \bar{I}_{\mathsf{L}}) \\ & + A_{\mathsf{J}} \nabla_{I_{\mathsf{V}}} i(\bar{U}, \bar{I}_{\mathsf{L}}, \bar{I}_{\mathsf{V}}, \bar{\dot{U}}, \bar{\dot{I}}_{\mathsf{L}}, \bar{\dot{I}}_{\mathsf{V}}, \bar{t})(I_{\mathsf{V}} - \bar{I}_{\mathsf{V}}) \\ & + A_{\mathsf{J}} \nabla_{\dot{U}} i(\bar{U}, \bar{I}_{\mathsf{L}}, \bar{I}_{\mathsf{V}}, \bar{\dot{U}}, \bar{\dot{I}}_{\mathsf{L}}, \bar{\dot{I}}_{\mathsf{V}}, \bar{t})(A^T(\dot{V} - \bar{\dot{V}})) \\ & + A_{\mathsf{J}} \nabla_{\dot{I}_{\mathsf{L}}} i(\bar{U}, \bar{I}_{\mathsf{L}}, \bar{I}_{\mathsf{V}}, \bar{\dot{U}}, \bar{\dot{I}}_{\mathsf{L}}, \bar{\dot{I}}_{\mathsf{V}}, \bar{t})(\dot{I}_{\mathsf{L}} - \bar{\dot{I}}_{\mathsf{L}}) \\ & + A_{\mathsf{J}} \nabla_{\dot{I}_{\mathsf{V}}} i(\bar{U}, \bar{I}_{\mathsf{L}}, \bar{I}_{\mathsf{V}}, \bar{\dot{U}}, \bar{\dot{I}}_{\mathsf{L}}, \bar{\dot{I}}_{\mathsf{V}}, \bar{t})(\dot{I}_{\mathsf{V}} - \bar{\dot{I}}_{\mathsf{V}}). \end{aligned} \tag{3.38}$$

Inserting the linearized behavior of the controlled current sources in (3.27) yields new terms into the formulation (3.34). For instance, the matrix $M(X,t)$ is modified such that

$$M(X,t) = \begin{bmatrix} A_{\mathsf{C}} C(A^T V, t) A_{\mathsf{C}}^T + A_{\mathsf{J}} \nabla_{\dot{U}} i(\bar{X}, \bar{\dot{X}}) A^T & A_{\mathsf{J}} \nabla_{I_{\mathsf{L}}} i(\bar{X}, \bar{\dot{X}}) & A_{\mathsf{J}} \nabla_{\dot{I}_{\mathsf{V}}} i(\bar{X}, \bar{\dot{X}}) \\ 0 & L(I_{\mathsf{L}}, t) & 0 \\ \nabla_U u(\bar{X}, \bar{\dot{X}}) A^T & \nabla_{I_{\mathsf{L}}} u(\bar{X}, \bar{\dot{X}}) & \nabla_{I_{\mathsf{V}}} u(\bar{X}, \bar{\dot{X}}) \end{bmatrix}. \tag{3.39}$$

These new terms in $M(X,t)$ can be interpreted as capacitive-like or inductive-like elements. The same augmentation has also to be done for $D(X,t)$ and $U(t)$ to obtain a similar expression to (3.34) where the sources have been linearized.

We will not enter into deeper details of this technical aspect of the current and voltage sources. In the sequel we will only keep the expression (3.34) for the sake of simplicity. Another numerical way to take into account this source is to explicitly evaluate the nonlinear terms with the previous values of the state. The controlled current and voltage sources can also be used to model the non-linear constitutive behavior of some particular components. The linearization of their behavior allows us to integrate their formulation in the MNA and to interpret their contribution to the circuit.

3.5.2 Index and Solvability

For a discussion of the solvability and the index of the obtained DAE, we refer to Günther and Feldmann (1993), Reissig and Feldmann (1996), März and Tischendorf (1997), Tischendorf (1999), Estèvez Schwarz and Tischendorf (2000), and Günther et al. (2005). In the sequel, we summarize the major results of the latter works.

3.5.2.1 Steady State Solutions

If the circuit contains neither capacitive branches nor inductive branches, or if we consider steady-state analysis of the circuits, the DAE reduces to a purely algebraic systems of nonlinear equations as

$$0 = D(X,t) + U(t), \tag{3.40}$$

where $X = [V, I_{\mathsf{L}}, I_{\mathsf{V}}]^T$ and

$$D(X,t) = \begin{bmatrix} -A_{\mathsf{R}} S(A_{\mathsf{R}}^T V, t) - A_{\mathsf{L}} I_{\mathsf{L}} - A_{\mathsf{V}} I_{\mathsf{V}} \\ A_{\mathsf{L}}^T V \\ A_{\mathsf{V}}^T V \end{bmatrix}, \tag{3.41}$$

$$U(t) = \begin{bmatrix} -A_{\mathsf{J}} i(t) \\ 0 \\ -u(t) \end{bmatrix}. \tag{3.42}$$

According to (3.17), the Jacobian matrix of the system is then given by

$$J(X,t) = \nabla_X D(X,t) = \begin{bmatrix} -A_{\mathsf{R}} G(A_{\mathsf{R}}^T V, t) A_{\mathsf{R}}^T & -A_{\mathsf{L}} & -A_{\mathsf{V}} \\ A_{\mathsf{L}}^T & 0 & 0 \\ A_{\mathsf{V}}^T & 0 & 0 \end{bmatrix}. \tag{3.43}$$

For the solvability of the steady-state systems (3.40), the following theorem on the incidence matrix is recalled.

Theorem 3.3 (Theorem 1.1 in Estèvez Schwarz and Tischendorf 2000) *The following relations are satisfied for the incidence matrix A split into $[A_{\mathsf{C}}\ A_{\mathsf{L}}\ A_{\mathsf{R}}\ A_{\mathsf{V}}\ A_{\mathsf{J}}]$:*

1. *the cut-sets of current sources are forbidden by the KCL; this implies that the matrix $[A_{\mathsf{C}}\ A_{\mathsf{L}}\ A_{\mathsf{R}}\ A_{\mathsf{V}}]$ has full row rank,*
2. *the loops of voltage sources are forbidden by the KVL; this implies that the matrix $[A_{\mathsf{V}}]$ has full column rank,*
3. *The matrix $[A_{\mathsf{C}}\ A_{\mathsf{R}}\ A_{\mathsf{V}}]$ has full row rank if and only if the circuit does not contain a cut-set consisting only of inductors and/or currents sources.*
4. *Let Q_{C} be any projector onto $\ker A_{\mathsf{C}}^T$. Then, the matrix $Q_{\mathsf{C}}^T A_{\mathsf{V}}$ has full column rank if and only if the circuit does not contain a loop consisting only of capacitors and voltage sources.*

We recall the definition of a projector. For $\mathbb{R}^n = R_1 \oplus R_2$, the linear application Q is a projector onto R_1 along R_2 if and only if $Q^2 = Q$, $\operatorname{im} Q = R_1$, and $\ker Q = R_2$. For a linear application A, $\ker A$ denotes the Kernel (or null) space of A *i.e.*

$$\ker A = \{x \mid Ax = 0\}, \tag{3.44}$$

and $\operatorname{im} A$ denotes the range space of A *i.e.*

$$\operatorname{im} A = \{y \mid \exists x,\ y = Ax\}. \tag{3.45}$$

Proof Let us recall that A has full row rank. If there is no cut-set of current sources, we can choose a spanning tree of the circuit with only branches that belong to $\mathsf{C} \cup \mathsf{L} \cup \mathsf{R} \cup \mathsf{V}$. Then the matrix $[A_{\mathsf{C}}\ A_{\mathsf{L}}\ A_{\mathsf{R}}\ A_{\mathsf{V}}]$ has full row rank. The same argument is valid if the circuit does contain a cut-set consisting only of inductors and/or currents sources. In this case, $[A_{\mathsf{C}}\ A_{\mathsf{R}}\ A_{\mathsf{V}}]$ has full row rank.

If the circuit does not contain any loop of voltage sources, the incidence matrix A_{V} is also an incidence matrix of the corresponding forest (a set of trees) and then is full column rank. If the circuit does not contain any loop of voltage sources and capacitors, the incidence matrix $[A_{\mathsf{V}}\ A_{\mathsf{C}}]$ is also a full column rank matrix. For the last point in Theorem 3.3, let us assume there is a loop consisting only of capacitors and voltage sources. Then the matrix $[A_{\mathsf{C}}\ A_{\mathsf{V}}]$ has dependent columns. It exists a nontrivial vector x, y such that

$$A_{\mathsf{C}} x + A_{\mathsf{V}} y = 0, \tag{3.46}$$

and since A_{V} has full column rank, one necessarily has $x \neq 0$. If $y = 0$, we have $A_{\mathsf{C}} x = 0$ and the circuit contains at least a loop with capacitors only. This is excluded by assumption. Multiplying (3.46) by Q_{C}^T, we get

$$Q_{\mathsf{C}}^T A_{\mathsf{V}} y = 0. \tag{3.47}$$

Since $y \neq 0$, the matrix $Q_{\mathsf{C}}^T A_{\mathsf{V}}$ does not have full column rank. Conversely, let us assume that there exists $y \neq 0$ such that (3.47) is satisfied. Then $A_{\mathsf{V}} y \in \ker Q_{\mathsf{C}}^T = \operatorname{im} A_{\mathsf{C}}$. There exists x such that (3.46) holds, and then there is a loop consisting only of capacitors and/or voltage sources. Since $A_{\mathsf{V}} y \neq 0$, the loop contains at least one voltage source. □

We can add the following result on the solvability of (3.40) based on some topological considerations.

Theorem 3.4 *Let us assume that G defined in* (3.17) *is positive definite. The matrix* (3.43) *is invertible if and only if the circuit does not contain neither a loop of independent voltage sources and/or inductors, nor cut-set of independent current sources and/or capacitors.*

Proof The following inclusion trivially holds for $\ker J(X,t)$,

$$\ker \begin{bmatrix} -A_{\mathsf{R}} G(A_{\mathsf{R}}^T V, t) A_{\mathsf{R}}^T \\ A_{\mathsf{L}}^T \\ A_{\mathsf{V}}^T \end{bmatrix} \times \ker [\, -A_{\mathsf{L}} \quad -A_{\mathsf{V}} \,] \subset \ker J(X,t). \tag{3.48}$$

Remarking that $\ker A_{\mathsf{R}}^T \subset \ker A_{\mathsf{R}} G(A_{\mathsf{R}}^T V, t) A_{\mathsf{R}}^T$, we obtain the following inclusion

$$\ker \begin{bmatrix} A_{\mathsf{R}}^T \\ A_{\mathsf{L}}^T \\ A_{\mathsf{V}}^T \end{bmatrix} \times \ker [\, -A_{\mathsf{L}} \quad -A_{\mathsf{V}} \,] \subset \ker J(X,t). \tag{3.49}$$

If the circuit contains a loop of independent voltage sources and/or inductors, the incidence matrix $[A_{\mathsf{L}}\ A_{\mathsf{V}}]$ has dependent columns, that is $\ker[A_{\mathsf{L}}\ A_{\mathsf{V}}] \neq \{0\}$. If the circuit contains a cut-set of independent current sources and/or capacitors, the graph made of the branches in $\mathsf{R} \cup \mathsf{L} \cup \mathsf{V}$ has c components with $c \geqslant 2$. The rank of the (reduced) incidence matrix $[A_{\mathsf{L}}\ A_{\mathsf{V}}]$ is the rank of the graph, that is $n_G - c$ where n_G is the number of vertices of the graph. The number of rows in $[A_{\mathsf{L}}\ A_{\mathsf{V}}]$ is $n_G - 1$, therefore $[A_{\mathsf{L}}\ A_{\mathsf{V}}]$ does not have full row rank since $n_G - 1 > n_G - c$. Due to the inclusion (3.49), we can conclude if the Jacobian matrix $J(X,t)$ is invertible, $\ker[A_{\mathsf{L}}\ A_{\mathsf{V}}] = \{0\}$ and $\ker[\, -A_{\mathsf{L}}\ -A_{\mathsf{V}} \,] = \{0\}$ and the circuit does not contain neither a loop of independent voltage sources and/or inductors nor a cut-set of independent current sources and/or capacitors.

Conversely, let $[x, y]^T$ be a vector such that $J(X,t)\left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right] = 0$:

$$A_{\mathsf{R}} G(A_{\mathsf{R}}^T V, t) A_{\mathsf{R}}^T x + [A_{\mathsf{L}}\ A_{\mathsf{V}}] y = 0 \quad \text{and} \quad x \in \ker[A_{\mathsf{L}} A_{\mathsf{V}}]^T. \tag{3.50}$$

Let us denote by Q_{LV} a projector on $\ker[A_{\mathsf{L}}\ A_{\mathsf{V}}]^T$. By definition of the projector onto a null space, we have

$$[A_{\mathsf{L}}\ A_{\mathsf{V}}]^T Q_{\mathsf{LV}} = 0, \tag{3.51}$$

and then

$$Q_{\mathsf{LV}}^T [A_{\mathsf{L}}\ A_{\mathsf{V}}] = 0. \tag{3.52}$$

Multiplying (3.50) by Q_{LV}^T, we get

$$Q_{\mathsf{LV}}^T A_{\mathsf{R}} G(A_{\mathsf{R}}^T V, t) A_{\mathsf{R}}^T x = 0. \tag{3.53}$$

Since $x \in \ker[A_{\mathsf{L}}\ A_{\mathsf{V}}]^T$, $Q_{\mathsf{LV}} x = x$ and we can write

$$x^T Q_{\mathsf{LV}}^T A_{\mathsf{R}} G(A_{\mathsf{R}}^T V, t) A_{\mathsf{R}}^T Q_{\mathsf{LV}} x = 0. \tag{3.54}$$

Since G is positive definite, we obtain $A_{\mathsf{R}}^T Q_{\mathsf{LV}} x = A_{\mathsf{R}}^T x = 0$ and

$$x \in \ker[A_{\mathsf{R}}\ A_{\mathsf{L}}\ A_{\mathsf{V}}]^T. \tag{3.55}$$

If the circuit does not contain cut-set of independent current sources and/or capacitors, the matrix $[A_{\mathsf{R}}\ A_{\mathsf{L}}\ A_{\mathsf{V}}]$ has full row rank, then $x = 0$. The condition (3.50) reduces to $y \in \ker[A_{\mathsf{L}}\ A_{\mathsf{V}}]$. If the circuit does not contain a loop of independent voltage sources and/or inductors, the matrix $[A_{\mathsf{L}}\ A_{\mathsf{V}}]$ has full column rank and $y = 0$. □

Standard analog simulators refuse to simulate in a steady-state analysis systems that do not fulfill the assumptions of Theorem 3.4.

3.5.2.2 Notion of Differential Index

Let us now consider the DAE (3.34). One of the main notions concerning the solvability of a DAE is the index. There are many notions of index for DAEs which are not necessarily equivalent. For a detailed presentation, we refer to the following standard books (Brenan et al. 1989; Hairer and Wanner 1996; Griepentrog and März 1986). To fix the ideas, we give the definition of the differential index.

Definition 3.5 (Differential index, see Brenan et al. 1989) The differential index ν of the general nonlinear and sufficiently smooth DAE

$$F(\dot{y}, y, t) = 0, \tag{3.56}$$

is the smallest integer such that

$$\begin{cases} F(\dot{y}, y, t) = 0, \\ \frac{d}{dt} F(\dot{y}, y, t) = 0, \\ \vdots \\ \frac{d^\nu}{dt^\nu} F(\dot{y}, y, t) = 0, \end{cases} \tag{3.57}$$

uniquely determines the variable $\dot{y}$ as a function of (y, t).

Briefly speaking, the differential index corresponds to the number of differentiations with respect to time to obtain a differential system which is similar to an ODE. In the case of a semi-explicit LTI DAE

$$\begin{cases} \dot{x} = Ax + Bz, \\ 0 = Cx + Dz, \end{cases} \tag{3.58}$$

the evaluation of the differential index amounts to checking the regularity of the system's Markov parameters: $D, CB, CAB, CA^2B, \ldots$

1. if D is a regular matrix, a single differentiation of the algebraic equation in (3.58) determines uniquely $\dot{x}$ and $\dot{z}$ as

$$\begin{cases} \dot{x} = Ax + Bz \\ \dot{z} = -D^{-1}(CAx + CBz). \end{cases} \tag{3.59}$$

Then the differential index is one;

2. if D is a singular matrix, a complete orthogonal decomposition (Golub and Van Loan 1996) can be performed

$$QDZ = \bar{D} = \begin{bmatrix} \bar{D}_{11} & 0 \\ 0 & 0 \end{bmatrix}, \tag{3.60}$$

such that Q and Z are orthogonal matrices and $\bar{D}_{11}$ a regular matrix of the same rank as D. Applying the change of variable

$$\begin{bmatrix} \bar{x} \\ \bar{z} \end{bmatrix} = Z^T \begin{bmatrix} x \\ z \end{bmatrix}, \tag{3.61}$$

we obtain

$$\dot{\bar{x}} = \bar{A}\bar{x} + \bar{B}\bar{z}, \tag{3.62a}$$

$$0 = \bar{C}_1\bar{x} + \bar{D}_{11}\bar{z}_1, \tag{3.62b}$$

$$0 = \bar{C}_2\bar{x}. \tag{3.62c}$$

Since D_{11} is regular, $\bar{z}_1$ appears to be an index-1 variable because a single derivation of (3.62b) yields

$$\dot{\bar{z}}_1 = -\bar{D}_{11}^{-1}\bar{C}_1[\bar{A}, \bar{B}]Z^T \begin{bmatrix} x \\ z \end{bmatrix}. \tag{3.63}$$

For $\bar{z}_2$, we need to derive (3.62c) twice and we get

$$\dot{\bar{z}}_2 = -(\bar{C}_2\bar{B}_2)^{-1}\bar{C}_2\bar{A}[\bar{A}, \bar{B}]Z^T \begin{bmatrix} x \\ z \end{bmatrix}, \tag{3.64}$$

since $\bar{C}_2\bar{B}_2$ is also a regular matrix. After two differentiations with respect to time, we obtain an explicit evaluation $\dot{x}$ and $\dot{z}$ as

$$\begin{cases} \dot{x} = Ax + Bz, \\ \dot{z} = Z_2 \begin{bmatrix} I_d \\ -\bar{D}_{11}^{-1}\bar{C}_1 \\ -(\bar{C}_2\bar{B}_2)^{-1}\bar{C}_2\bar{A} \end{bmatrix} [\bar{A}, \bar{B}]Z^T \left[\begin{smallmatrix} x \\ z \end{smallmatrix}\right]. \end{cases} \tag{3.65}$$

We conclude that the differential index is 2.

3. Similar transformations can be performed for higher index systems by stating additional assumptions on the regularity of $CA^{\nu}B$ for a DAE of index $\nu + 2$, $\nu \geqslant 1$.

In the case of an LTI DAE

$$M\dot{y} = Jy + f, \tag{3.66}$$

another transformation is often used. If the matrix pencil $\lambda M + J$ is regular, there exist non singular matrices P and Q such that

$$PMQ = \begin{bmatrix} I_d & 0 \\ 0 & N \end{bmatrix} \quad \text{and} \quad PJQ = \begin{bmatrix} C & 0 \\ 0 & I_d \end{bmatrix}, \tag{3.67}$$

where N is a matrix of nilpotency k. This transformation leads to a DAE of the form

$$\begin{cases} \dot{y}_1 = Cy_1 + f_1, \\ N\dot{y}_2 = Cy_1 + f_2. \end{cases} \tag{3.68}$$

A simple calculation shows that y_2 may be expressed as

$$y_2 = \sum_{i=0}^{k-1} (-1)^i N^i \frac{d^i f_2}{dt^i}. \tag{3.69}$$

We can therefore conclude that the degree of nilpotency k of N is the differential index of the DAE.

3.5.2.3 Topological Index Results for the MNA

To summarize the results on the index based on topological considerations, we cite two main theorems of Tischendorf (1999) and Estèvez Schwarz and Tischendorf (2000).

Theorem 3.6 (Part of Theorem 3.2 in Estèvez Schwarz and Tischendorf (2000)) *Let us consider the DAE* (3.34) *given by the MNA formulation of the circuit with only independent sources. Let us assume that the Jacobian matrices $G(U_\mathsf{R}, t)$, $C(U_\mathsf{C}, t)$ and $L(I_\mathsf{L}, t)$ are positive definite.*

1. *If the circuit contains neither cut-sets with only inductive and/or independent current sources, nor loops with capacitive elements and independent voltage sources, then the DAE* (3.34) *is of differential index one.*
2. *If the circuit contains cut-sets with only inductive and/or independent current sources, or loops with capacitive elements and independent voltage sources, then the DAE* (3.34) *is of differential index two.*

Theorem 3.6 is completed by the statements of the explicit constraints in the index-1 case and the hidden constraints in the index-2 case. Moreover, the result is given in a more general framework with controlled current and voltage sources. The following theorem shows that the conditions of Theorem 3.6 are necessary and sufficient conditions.

Theorem 3.7 (Theorem 3.3 in Estèvez Schwarz and Tischendorf 2000) *If the differential index of the DAE* (3.34) *is one, then the network contains neither cut-sets with only inductive and/or independent current sources, nor loops with capacitive elements and independent voltage sources. If the differential index of the DAE* (3.34) *is two, then the network contains at least a cut-set with only inductive and/or independent current sources, or a loop with capacitive elements and independent voltage sources.*

Similar results have also been obtained for the charge-oriented MNA (3.28) and using the tractability index introduced in Griepentrog and März (1986) and März (1992).

3.6 Semi-Explicit DAE Forms

When we examine the structure of the DAE (3.34), a question is raised about the possibility to obtain a semi-explicit DAE of the form

$$\begin{cases} N(x,t)\dot{x} = f(x,z,t), \\ 0 = g(x,z,t), \end{cases} \tag{3.70}$$

where $N(x,t)$ is a regular matrix.

When the DAE (3.34) is of index one, as in most of the cases, the semi-explicit form of the DAE (3.34) is of low interest because standard numerical time integration schemes correctly work on index one DAEs. Nevertheless, we are interested in this question for our further developments with nonsmooth elements.

The question is also closely related to the notion of index. Indeed, if we consider a DAE in the implicit form as (3.56), a semi-explicit form can be straightforwardly obtained by setting

$$\begin{cases} \dot{y} = z, \\ 0 = F(z,y,t). \end{cases} \tag{3.71}$$

By doing such a change of variable, the resulting differential index of (3.71) is equal to the index of the original DAE (3.56) plus one. The quest for a semi-explicit form (3.70) is therefore constrained by the increase of the index and in some sense the question of the redundancy of unknowns.

3.6.1 *A First Naive Attempt*

To outline a part of the algebraic equations in (3.34), we may split the vector of unknowns X as follows: $X = [x^T, z^T]^T$ with $x = [V^T, I_{\mathsf{L}}^T]^T$ and $z = [I_{\mathsf{V}}]^T$. One gets the following equivalent system

$$\begin{cases} N(x,t)\dot{x} = f(x,z,t), \\ 0 = g(x,z,t) \end{cases} \tag{3.72}$$

with

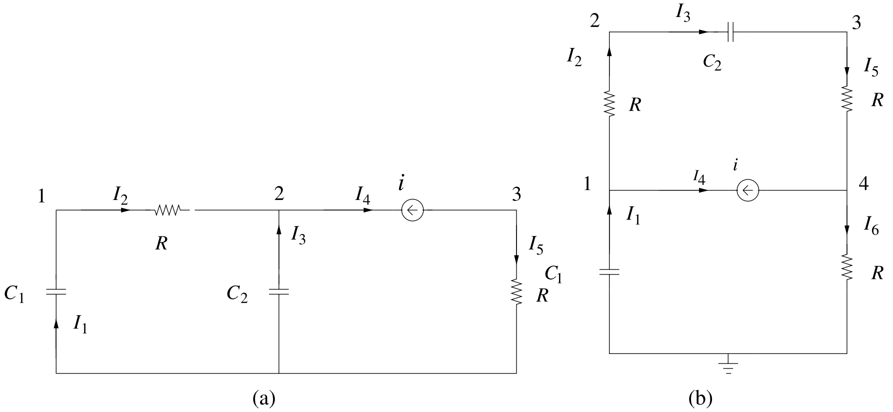


Fig. 3.1 Simple circuits illustrating DAE formulations

$$N(x,t) = \begin{bmatrix} A_{\mathsf{C}} C(A^T V, t) A_C^T & 0 \\ 0 & L(I_{\mathsf{L}}, t) \end{bmatrix}, \tag{3.73}$$

$$f(x,z,t) = \begin{bmatrix} -A_{\mathsf{C}} q_t(A_{\mathsf{C}}^T V, t) - A_{\mathsf{R}} S(A_{\mathsf{R}}^T V, t) - A_{\mathsf{L}} I_{\mathsf{L}} - A_{\mathsf{V}} I_{\mathsf{V}} - A_{\mathsf{J}} i(t) \\ A_{\mathsf{L}}^T V - \phi_t(I_{\mathsf{L}}, t) \end{bmatrix}, \tag{3.74}$$

$$g(x,z,t) = [\, A_{\mathsf{V}}^T V - u(t) \,]. \tag{3.75}$$

Unfortunately, this formulation does not lead to a semi-explicit DAE, because the matrix $N(x,t)$ is not necessarily regular due to the matrix $A_{\mathsf{C}} C(A^T V, t) A_C^T$ which is almost never invertible. Assuming that the matrices C and L are positive-definite, the matrix $A_{\mathsf{C}} C A_{\mathsf{C}}^T$ is regular if a spanning tree of the circuit with only capacitive elements can be found. In others terms, for each node, there is path to the reference node (ground) with only capacitive elements.

Example 3.8 To illustrate the various formulations, let us consider the simple circuit depicted in Fig. 3.1(a). The reduced incidence matrix is given by

$$A = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix}, \tag{3.76}$$

which can be split into

$$A_{\mathsf{C}} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, \qquad A_{\mathsf{R}} = \begin{bmatrix} -1 & 0 & 0 \\ 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}, \quad \text{and} \quad A_{\mathsf{J}} = \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix}. \tag{3.77}$$

The matrix

$$A_{\mathsf{C}} C A_{\mathsf{C}}^T = \begin{bmatrix} C_1 & 0 & 0 \\ 0 & C_2 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \tag{3.78}$$

is obviously singular since there is no tree of capacitive branches that spans the whole circuit. The equation of the circuit in the standard MNA form is given by

$$\begin{bmatrix} C_1 & 0 & 0 \\ 0 & C_2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{V}_1 \\ \dot{V}_2 \\ \dot{V}_3 \end{bmatrix} = \begin{bmatrix} -1/R & 1/R & 0 \\ 1/R & -1/R & 0 \\ 0 & 0 & -1/R \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \\ V_3 \end{bmatrix} + \begin{bmatrix} 0 \\ -i \\ i \end{bmatrix}. \tag{3.79}$$

The DAE (3.79) is of differential index 1 since a single differentiation of the third equation allows us to uniquely define $\dot{V}$ in terms of V.

3.6.2 A Second Attempt

Assuming that $C(A^T V, t)$ is positive definite and $L(I_\mathsf{L})$ is positive definite, another method that may be used to obtain a regular matrix is to put all the capacitor branch voltages with the inductive currents in the unknown vector, that is $x = [U_\mathsf{C}^T, I_\mathsf{L}^T]^T$ and to write the law $C\frac{dU_\mathsf{C}}{dt} = I_\mathsf{C}$ to fill the matrix N such that

$$N(x,t) = \begin{bmatrix} C(x,t) & 0 \\ 0 & L(x,t) \end{bmatrix}, \tag{3.80}$$

such that $N(x,t)$ is regular. In order to state the remaining equations, we have to choose $z = [V^T, I_\mathsf{V}^T, I_\mathsf{C}^T]^T$ and one gets

$$f(x,z,t) = \begin{bmatrix} I_\mathsf{C} - q_t(U_\mathsf{C}, t) \\ A_\mathsf{L}^T V - \phi_t(I_\mathsf{L}, t) \end{bmatrix}, \tag{3.81}$$

and

$$\begin{aligned} g(x,z,t) = {} & \begin{bmatrix} -I_d & 0 \\ 0 & A_\mathsf{L} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} U_\mathsf{C} \\ I_\mathsf{L} \end{bmatrix} \\ & + \begin{bmatrix} A_\mathsf{C}^T V \\ A_\mathsf{C} I_\mathsf{C} + A_\mathsf{R} S(A_\mathsf{R}^T V, t) + A_\mathsf{V} I_\mathsf{V} + A_\mathsf{J} i(t) \\ A_\mathsf{V}^T V - u(t) \end{bmatrix}. \end{aligned} \tag{3.82}$$

In the LTI case, one obtains

$$\begin{cases} \begin{bmatrix} C & 0 \\ 0 & L \end{bmatrix} \begin{bmatrix} \dot{U}_\mathsf{C} \\ \dot{I}_\mathsf{L} \end{bmatrix} = \begin{bmatrix} I_\mathsf{C} \\ A_\mathsf{V}^T V \end{bmatrix}, \\ 0 = \begin{bmatrix} -I_d & 0 \\ 0 & A_\mathsf{L} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} U_\mathsf{C} \\ I_\mathsf{L} \end{bmatrix} + \begin{bmatrix} 0 & A_\mathsf{C}^T & 0 \\ A_\mathsf{C} & A_\mathsf{R} G A_\mathsf{R}^T & A_\mathsf{V} \\ 0 & A_\mathsf{V}^T & 0 \end{bmatrix} \begin{bmatrix} I_\mathsf{C} \\ V \\ I_\mathsf{V} \end{bmatrix} + \begin{bmatrix} 0 \\ A_\mathsf{J} i(t) \\ -u(t) \end{bmatrix}. \end{cases} \tag{3.83}$$

This method is not appropriate because it increases the size of the vector of unknowns. Although this change of unknowns is very similar to the method evoked in the beginning of this section to pass from (3.56) to (3.71), it does not lead to an increase of the index of the problem.

Theorem 3.9 *If the circuit does not contain neither cut-sets with only inductive and/or independent current sources, nor loops with capacitive elements and independent voltage sources, then:*

1. *the DAE (3.83) is of differential index one,*
2. *the matrix*
$$\begin{bmatrix} 0 & A_{\mathsf{C}}^T & 0 \\ A_{\mathsf{C}} & A_{\mathsf{R}} G A_{\mathsf{R}}^T & A_{\mathsf{V}} \\ 0 & A_{\mathsf{V}}^T & 0 \end{bmatrix} \tag{3.84}$$
is regular.

Proof Let
$$\begin{bmatrix} x \\ y \\ z \end{bmatrix}$$
be a vector such that
$$A_{\mathsf{C}}^T y = 0, \tag{3.85}$$
$$A_{\mathsf{V}}^T y = 0, \tag{3.86}$$
$$A_{\mathsf{C}} x + A_{\mathsf{R}} G A_{\mathsf{R}}^T y + A_{\mathsf{V}} z = 0, \tag{3.87}$$
i.e.
$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} \in \ker \tilde{D}.$$

Let us denote Q_{CV} a projector on $\ker[A_{\mathsf{C}}\ A_{\mathsf{V}}]^T$. By definition, we recall that
$$[A_{\mathsf{C}}\ A_{\mathsf{V}}]^T Q_{\mathsf{CV}} = 0 \quad \text{and} \quad Q_{\mathsf{CV}}^T [A_{\mathsf{C}}\ A_{\mathsf{V}}] = 0. \tag{3.88}$$
Multiplying (3.87) by Q_{CV}^T, we obtain
$$Q_{\mathsf{CV}}^T [A_{\mathsf{C}}\ A_{\mathsf{V}}] \begin{bmatrix} x \\ z \end{bmatrix} + Q_{\mathsf{CV}}^T A_{\mathsf{R}} G A_{\mathsf{R}}^T y = 0. \tag{3.89}$$

By definition, $Q_{\mathsf{CV}}^T [A_{\mathsf{C}}\ A_{\mathsf{V}}] = 0$, and using (3.85) and (3.86), we have that $Q_{\mathsf{CV}} y = y$. Then (3.89) can be written as
$$Q_{\mathsf{CV}}^T A_{\mathsf{R}} G A_{\mathsf{R}}^T Q_{\mathsf{CV}} y = 0. \tag{3.90}$$
Multiplying by y^T and recalling that G is positive definite, we have
$$A_{\mathsf{R}}^T Q_{\mathsf{CV}} y = 0, \quad \text{or equivalently} \quad A_{\mathsf{R}}^T y = 0. \tag{3.91}$$
The relations (3.91), (3.85) and (3.86) result in $y \in \ker[A_{\mathsf{C}}\ A_{\mathsf{R}}\ A_{\mathsf{V}}]^T$. According to Theorem 3.3, if the circuit does not contain cut-set consisting of inductors and/or current sources only, the matrix $[A_{\mathsf{C}}\ A_{\mathsf{R}}\ A_{\mathsf{V}}]$ has full row rank. Then, we can conclude that $y = 0$ and
$$[A_{\mathsf{C}}\ A_{\mathsf{V}}] \begin{bmatrix} x \\ z \end{bmatrix} = 0. \tag{3.92}$$

Let us multiply (3.92) by Q_{C}^T, we get that

$$Q_{\mathsf{C}}^T A_{\mathsf{V}} z = 0. \tag{3.93}$$

According to Theorem 3.3 again, if the circuit does not contain any loop consisting of capacitors and voltage sources only, the matrix $Q_{\mathsf{C}}^T A_{\mathsf{V}}$ has full column rank, then $z = 0$. Due the fact that A_{V} has full column rank, we conclude that $x = 0$. The matrix in (3.84) is therefore regular and the DAE (3.83) is of index one. □

Example 3.10 To illustrate the second approach, let us consider the circuit depicted in Fig. 3.1(b). The reduced incidence matrix is split into

$$A_{\mathsf{C}} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \qquad A_{\mathsf{R}} = \begin{bmatrix} -1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \quad \text{and} \quad A_{\mathsf{J}} = \begin{bmatrix} -1 \\ 0 \\ 0 \\ 1 \end{bmatrix}. \tag{3.94}$$

The matrix

$$A_{\mathsf{C}} C A_{\mathsf{C}}^T = \begin{bmatrix} C_1 & 0 & 0 & 0 \\ 0 & C_2 & -C_2 & 0 \\ 0 & -C_2 & C_2 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \tag{3.95}$$

is obviously singular since there is no tree of capacitive branches that spans the whole circuit. The equations of the circuit in the standard MNA form are given by

$$\begin{bmatrix} C_1 & 0 & 0 & 0 \\ 0 & C_2 & -C_2 & 0 \\ 0 & -C_2 & C_2 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{V}_1 \\ \dot{V}_2 \\ \dot{V}_3 \\ \dot{V}_4 \end{bmatrix} = \begin{bmatrix} -1/R & 1/R & 0 & 0 \\ 1/R & -1/R & 0 & 0 \\ 0 & 0 & 1/R & -1/R \\ 0 & 0 & -1/R & 2/R \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \\ V_3 \\ V_4 \end{bmatrix} + \begin{bmatrix} -i \\ 0 \\ 0 \\ i \end{bmatrix}. \tag{3.96}$$

The DAE (3.96) is of differential index 1 since a single differentiation yields

$$\begin{bmatrix} C_1 & 0 & 0 & 0 \\ 0 & C_2 & -C_2 & 0 \\ 0 & -C_2 & C_2 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \ddot{V}_1 \\ \ddot{V}_2 \\ \ddot{V}_3 \\ \ddot{V}_4 \end{bmatrix} = \begin{bmatrix} -1/R & 1/R & 0 & 0 \\ 1/R & -1/R & 0 & 0 \\ 0 & 0 & 1/R & -1/R \\ 0 & 0 & -1/R & 2/R \end{bmatrix} \begin{bmatrix} \dot{V}_1 \\ \dot{V}_2 \\ \dot{V}_3 \\ \dot{V}_4 \end{bmatrix}. \tag{3.97}$$

Combining (3.96) and (3.97), we determine the derivatives of the unknowns as

$$\begin{bmatrix} \dot{V}_1 \\ \dot{V}_2 \\ \dot{V}_3 \\ \dot{V}_4 \end{bmatrix} = 1/R \begin{bmatrix} -1/C_1 & 1/C_1 & 0 & 0 \\ -1/C_1 & 1/C_1 & -1/(2C_2) & -1/(2C_2) \\ -1/C_1 & 1/C_1 & 1/(2C_2) & -1/(2C_2) \\ -1/(2C_1) & 1/(2C_1) & 1/(4C_2) & -1/(4C_2) \end{bmatrix}$$
$$+ \begin{bmatrix} V_1 \\ V_2 \\ V_3 \\ V_4 \end{bmatrix} \begin{bmatrix} -i/C_1 \\ -i/C_1 \\ -i/C_1 \\ -i/2C_1 \end{bmatrix}. \tag{3.98}$$

The index can be also be exhibited by performing a row compression of the matrix $\tilde{C} = A_{\mathsf{C}} C A_{\mathsf{C}}^T$ using a transformation P based on the range of $\tilde{C}$ and $\tilde{C}^T$ as

$$P = \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & -1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}. \tag{3.99}$$

The following change of unknown

$$X = P^T \begin{bmatrix} x \\ z \end{bmatrix}, \tag{3.100}$$

yields

$$\begin{cases} \begin{bmatrix} C_1 & 0 \\ 0 & C_2 \end{bmatrix} \dot{x} = \frac{1}{R} \begin{bmatrix} -1 & -1/2 \\ -1/2 & 0 \end{bmatrix} x + \frac{1}{R} \begin{bmatrix} 1/2 & 0 \\ 1/2 & -1/2 \end{bmatrix} z, \\ 0 = \frac{1}{R} \begin{bmatrix} 1/2 & 1/2 \\ 0 & -1/2 \end{bmatrix} x + \frac{1}{R} \begin{bmatrix} 0 & -1/2 \\ -1/2 & 1 \end{bmatrix} z. \end{cases} \tag{3.101}$$

The fact that the matrix $\frac{1}{R}\begin{bmatrix} 0 & -1/2 \\ -1/2 & 1 \end{bmatrix}$ is invertible shows that the DAE (3.101) is of index 1. In the semi-explicit form (3.101), the variable $x = [V_1, V_3 - V_2]^T$ appears to be an "index-0 variable", due to the fact that $\dot{x}$ is explicitly given as a function of the unknown variables x and z. On the contrary, the variable $z = [V_3 + V_2, V_4]^T$ appears as an "index-1 variable", because we have to derive with respect to time the algebraic equations to obtain $\dot{z}$ in terms of x and z.

Let us now introduce the branch voltages U_{C_1} and U_{C_2}. The formulation given by (3.80)– (3.82) yields

$$\begin{cases} \begin{bmatrix} C_1 & 0 \\ 0 & C_2 \end{bmatrix} \begin{bmatrix} \dot{U}_{C_1} \\ \dot{U}_{C_2} \end{bmatrix} = \begin{bmatrix} I_{C_1} \\ I_{C_2} \end{bmatrix}, \\ \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} U_{C_1} \\ U_{C_2} \end{bmatrix} + \begin{bmatrix} 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 \\ 1 & 0 & -1/R & 1/R & 0 & 0 \\ 0 & 1 & 1/R & -1/R & 0 & 0 \\ 0 & -1 & 0 & 0 & 1/R & -1/R \\ 0 & 0 & 0 & 0 & -1/R & 2/R \end{bmatrix} \begin{bmatrix} I_{C_1} \\ I_{C_2} \\ V_1 \\ V_2 \\ V_3 \\ V_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ i \\ 0 \\ 0 \\ -i \end{bmatrix}, \end{cases} \tag{3.102}$$

and yields also an index-1 DAE.

3.6.3 The Proposed Solution

In the previous section, a semi-explicit formulation which does not increase the index (at least in the index-1 case) is proposed. The main drawback of the formulation (3.80) is the increase of the number of unknowns. For each capacitive branch of the circuit two variables are added, *i.e.* U_{C} and I_{C}. We propose in this section to reduce this number of unknowns by substitution when it is directly possible. Let us consider a splitting of the incidence matrix of the capacitive branches as:

$$A_{\mathsf{C}} = \begin{bmatrix} \tilde{A}_{\mathsf{C_F}} & \tilde{A}_{\mathsf{C_L}} \\ \hat{A}_{\mathsf{C_F}} & \hat{A}_{\mathsf{C_L}} \end{bmatrix}, \tag{3.103}$$

such that $\tilde{A}_{\mathsf{C_F}}$ is invertible. This splitting is built by firstly splitting the set of capacitive branches in the set $\mathsf{C_F}$ and the set $\mathsf{C_L}$ by choosing a spanning forest of the capacitive branches. The subscript F stands for "forest" and the subscript L stands for "links". If the circuit does not contain a loop of capacitors, then $\mathsf{C_L} = \emptyset$. Secondly, the set of nodes N is split into two subsets $\tilde{\mathsf{N}}$ and $\hat{\mathsf{N}}$. The set $\tilde{\mathsf{N}}$ is built by choosing the nodes contained in the spanning forest and removing for each component (connected graph) a reference node. If the ground node is already in the forest, it is not removed again. According to this construction, $\tilde{A}_{\mathsf{C_F}}$ is invertible. Let us now apply this splitting to the semi-explicit DAE (3.80) with (3.81) and (3.82). The split KCL yields:

$$\begin{aligned} I_{\mathsf{C_F}} &= -\tilde{A}_{\mathsf{C_F}}^{-1}\left[\tilde{A}_{\mathsf{C_L}} I_{\mathsf{C_L}} + \tilde{A}_{\mathsf{R}} S(A_{\mathsf{R}}^T V, t) + \tilde{A}_{\mathsf{V}} I_{\mathsf{V}} + \tilde{A}_{\mathsf{L}} I_{\mathsf{L}} + \tilde{A}_{\mathsf{J}} i(t)\right], \\ 0 &= (\hat{A}_{\mathsf{C_L}} - \hat{A}_{\mathsf{C_F}} \tilde{A}_{\mathsf{C_F}}^{-1} \tilde{A}_{\mathsf{C_L}}) I_{\mathsf{C_L}} + (\hat{A}_{\mathsf{R}} - \hat{A}_{\mathsf{C_F}} \tilde{A}_{\mathsf{C_F}}^{-1} \tilde{A}_{\mathsf{R}}) S(A_{\mathsf{R}}^T V, t) \\ &\quad + (\hat{A}_{\mathsf{L}} - \hat{A}_{\mathsf{C_F}} \tilde{A}_{\mathsf{C_F}}^{-1} \tilde{A}_{\mathsf{L}}) I_{\mathsf{L}} + (\hat{A}_{\mathsf{V}} - \hat{A}_{\mathsf{C_F}} \tilde{A}_{\mathsf{C_F}}^{-1} \tilde{A}_{\mathsf{V}}) I_{\mathsf{V}} + (\hat{A}_{\mathsf{J}} - \hat{A}_{\mathsf{C_F}} \tilde{A}_{\mathsf{C_F}}^{-1} \tilde{A}_{\mathsf{J}}) i(t). \end{aligned} \tag{3.104}$$

By identification with (3.70), we get

$$N(x,t) = \begin{bmatrix} C(x,t) & 0 \\ 0 & L(x,t) \end{bmatrix}, \tag{3.105}$$

$$f(x,z,t) = \begin{bmatrix} -\tilde{A}_{\mathsf{C_F}}^{-1}[\tilde{A}_{\mathsf{C_L}} I_{\mathsf{C_L}} + \tilde{A}_{\mathsf{R}} S(A_{\mathsf{R}}^T V, t) + \tilde{A}_{\mathsf{V}} I_{\mathsf{V}} + \tilde{A}_{\mathsf{L}} I_{\mathsf{L}}] - q_t(U_{\mathsf{C_F}}, t) \\ I_{\mathsf{C_L}} - q_t(U_{\mathsf{C_L}}, t) \\ A_{\mathsf{L}}^T V - \phi_t(I_{\mathsf{L}}, t) \end{bmatrix}, \tag{3.106}$$

and

$$\begin{aligned} g(x,z,t) = &\begin{bmatrix} -I_d & 0 \\ 0 & \hat{A}_{\mathsf{L}} - \hat{A}_{\mathsf{C_F}} \tilde{A}_{\mathsf{C_F}}^{-1} \tilde{A}_{\mathsf{L}} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} U_{\mathsf{C}} \\ I_{\mathsf{L}} \end{bmatrix} \\ &+ \begin{bmatrix} A_{\mathsf{C}}^T V \\ (\hat{A}_{\mathsf{C_L}} - \hat{A}_{\mathsf{C_F}} \tilde{A}_{\mathsf{C_F}}^{-1} \tilde{A}_{\mathsf{C_L}}) I_{\mathsf{C_L}} + (\hat{A}_{\mathsf{R}} - \hat{A}_{\mathsf{C_F}} \tilde{A}_{\mathsf{C_F}}^{-1} \tilde{A}_{\mathsf{R}}) S(A_{\mathsf{R}}^T V, t) \\ + (\hat{A}_{\mathsf{V}} - \hat{A}_{\mathsf{C_F}} \tilde{A}_{\mathsf{C_F}}^{-1} \tilde{A}_{\mathsf{L}}) I_{\mathsf{V}} + (\hat{A}_{\mathsf{J}} - \hat{A}_{\mathsf{C_F}} \tilde{A}_{\mathsf{C_F}}^{-1} \tilde{A}_{\mathsf{J}}) i(t) \\ A_{\mathsf{V}}^T V - u(t) \end{bmatrix}. \end{aligned} \tag{3.107}$$

With this substitution, we have eliminated the currents $I_{\mathsf{C_F}}$. It is worth noting that this substitution in only based on topological considerations. It can be done only one time in the initialization process. In the same vein, a substantial part of the node potential V can be eliminated by writing

$$\tilde{V} = \tilde{A}_{\mathsf{CF}}^{-T}(U_{\mathsf{CF}} + \hat{A}_{\mathsf{CF}}^{T}\hat{V}). \tag{3.108}$$

This latter substitution is done in practice but for the sake of readability it is not reported here. We end with the following variable definition:

$$x = \begin{bmatrix} U_{\mathsf{C}} \\ I_{\mathsf{L}} \end{bmatrix}, \qquad z = \begin{bmatrix} \hat{V} \\ I_{\mathsf{C_L}} \\ I_{\mathsf{V}} \end{bmatrix}. \tag{3.109}$$

3.7 Basics on Standard Circuit Simulation

In this section, we recall the basic ingredients of the conventional approach for simulating a circuit in the standard SPICE-like analog approach. Further details can be found in any standard textbooks on simulation of electrical circuits (Chua et al. 1991) or in the SPICE reference manual. In the comprehensive review of Günther et al. (2005), the reader will find more informations on less conventional approaches for the simulations of circuits.

Three main ingredients are at the heart of the approach:

- Computation of the initial conditions,
- Time-discretization of the DAE resulting from the MNA,
- Solving a nonlinear systems by a Newton-like method.

The following sections will focus on each of these points, starting from the MNA formulation with independent sources ((3.34) to (3.37)).

3.7.1 Computation of the Initial Conditions

The computation of consistent initial conditions amounts to performing a steady state analysis (DC operating point). The problem that we have to solve is the system of nonlinear equations (3.40) for the initial condition X_0 at the given initial time t_0, *i.e.*:

$$0 = D(X_0, t_0) + U(t_0). \tag{3.110}$$

As we said in Sect. 3.5.2.1, the solvability of this problem depends on the topology of the circuit if the branch constitutive equations are well defined. Theorem 3.4 gives the condition on the circuit topology for the solvability of (3.110). In practice, the user can prescribe a part of the initial conditions. In this case, the prescribed known values are then dropped from the unknown vector.

3.7.2 Time-Discretization of the MNA

The approximate solution of the DAE (3.34) is computed at discrete time points, t_k by numerical integration usually using linear implicit multi-step methods (Hairer and Wanner 1996). To simplify the presentation, let us rewrite (3.34) in a more compact form as:

$$\begin{cases} 0 = E\dot{y}(t) + f(x(t), t), \\ 0 = y(t) - g(x(t)), \end{cases} \tag{3.111}$$

with $y = [q^T, \phi^T]^T$, $x = [V^T, I_\mathsf{V}^T, I_\mathsf{L}^T]^T$,

$$E = \begin{bmatrix} A_\mathsf{C} & 0 \\ 0 & I_d \\ 0 & 0 \end{bmatrix}, \qquad g(x,t) = \begin{bmatrix} q_\mathsf{C}(A_\mathsf{C}^T V) \\ \phi_\mathsf{L}(I_\mathsf{L}) \end{bmatrix}, \tag{3.112}$$

and

$$f(x,t) = \begin{bmatrix} A_\mathsf{C} q_t(A_\mathsf{C}^T V, t) + A_\mathsf{R} S(A_\mathsf{R}^T V, t) + A_\mathsf{L} I_\mathsf{L} + A_\mathsf{V} I_\mathsf{V} + A_\mathsf{J} i(t) \\ A_\mathsf{L}^T V - \phi_t(I_\mathsf{L}, t) \\ A_\mathsf{V}^T V - u(t) \end{bmatrix}. \tag{3.113}$$

Briefly speaking, the approach consists in an approximation of $\dot{y}$ by a Backward Differentiation Formulas (BDF) pioneered by Gear (1971) as:

$$\dot{y}_k = \frac{1}{h_k} \sum_{i=0}^{\rho} \gamma_{k,i} y_{k-i} - \sum_{i=0}^{\rho} \beta_{k,i} \dot{y}_{k-i}, \tag{3.114}$$

where ρ is the number of steps required by the integrator, h_k is the length of the time-step k, $\gamma_{k,i}, \beta_{k,i}$ are the coefficients of the method to ensure a certain order of consistency, $y_{k-i}, \dot{y}_{k-i}$ are the already computed values for $i = 1, \ldots, \rho$. The formula (3.114) is expressed in terms of the unknowns at step k by

$$\dot{y}_k = \alpha_k y_k + r_k, \tag{3.115}$$

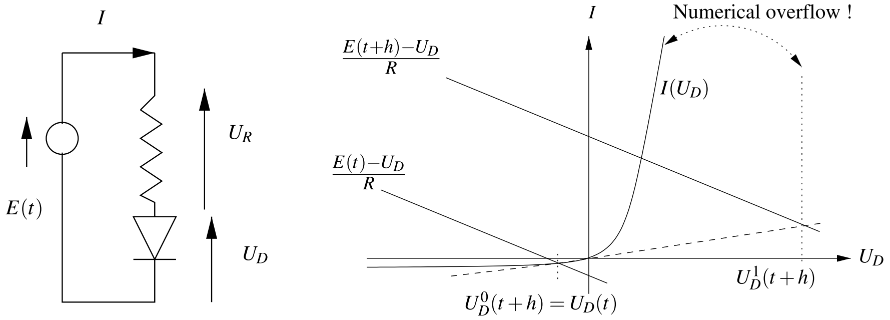
where α_k and r_k can be easily identified from (3.114). Without going into deeper details, the substitution of $\dot{y}$ in (3.111) by (3.115) yields a system of nonlinear equations:

$$0 = E\dot{y}_k + f(x_k, t) = E(\alpha_k g(x_k) + r_k) + f(x_k, t_k). \tag{3.116}$$

3.7.3 Solving Nonlinear Systems

The system of nonlinear equations (3.116) (as well for (3.110)) is usually solved with the Newton's method. The solution is sought as the limit of the sequence x_k^α for α defined by

$$\nabla_x^T \mathscr{R}(x_k^\alpha)(x_k^{\alpha+1} - x_k^\alpha) = -\mathscr{R}(x_k^\alpha), \tag{3.117}$$



(a) Polarization of a diode

(b) Newton's method iterates

Fig. 3.2 Newton's method failure

where the nonlinear residue is defined as

$$\mathscr{R}(x) = E(\alpha_k g(x) + r_k) + f(x, t_k). \tag{3.118}$$

The nonlinear system (3.117) yields

$$(\alpha_k E \nabla_x^T g(x_k^\alpha) + \nabla_x^T f(x_k^\alpha, t_k))(x_k^{\alpha+1} - x_k^\alpha) = -\mathscr{R}(x_k^\alpha). \tag{3.119}$$

The solvability of (3.117) relies on the regularity of the matrix pencil $\{\nabla_x^T g(x_k^\alpha), \nabla_x^T f(x_k^\alpha, t_k)\}$. This regularity is satisfied if the DAE (3.111) is of index one. In Example 3.11, we illustrate one of the motivations to deal with nonsmooth models rather than nonlinear stiff models.

Example 3.11 (Numerical overflow with stiff nonlinear model) The exponential characteristic of the diode (or the quadratic response of a MOS transistor to V_{GS} when it goes from the weak inversion region to the strong inversion region) may cause convergence problems when a DC analysis tries to find an equilibrium point belonging to a region different from the initial guess, or when a transient analysis tries to compute the evolution of the device's current across two regions.

This will be illustrated in the case of the diode when the Newton method tries to find two successive polarization points of a circuit with a voltage source $E(t)$, a resistor R and a diode D (see Fig. 3.2(a) for the circuit's description and Fig. 3.2(b) for the algorithm).

The process for finding the new polarization point at time $t+h$ starts with the initial guess $U_D^0(t+h)$, *i.e.* the polarization point obtained for time $t : U_D(t)$. The first iterate $U_D^1(t+h)$ is given by the intersection between the tangent to the diode's characteristic and the source-resistor characteristic. Due to the stiffness of the diode's characteristic, this value $U_D^1(t+h)$ will result in a numerical overflow during the computation of the next iterate $U_D^2(t+h)$.

Several solutions try to overcome these problems, yielding an increase of the number of iterations. Note that the secant method replacing Newton-Raphson algorithm is also prone to such kind of drawback. The huge number of iterations required to compute the solution of a set of nonlinear equations is a key problem when one wants to simulate a circuit involving a large number of components with stiff characteristics for a large number of cycles. This is the case for instance in power electronics.

3.7.4 Implementation Details and the Stamping Method

In the standard implementation of the MNA as in the SPICE family simulators, the incidence matrices A_{C}, A_{L}, A_{R} are never formed, but the product with the capacitance matrix, the inductance matrix or the conductance matrix is directly inserted into $M(X,t)$, $D(X,t)$ and $U(t)$. This algorithm is usually called the "stamp method", which is an algorithmic method used to fill the table equation from the components. It consists in writing a sub-table for each type of component, this sub-table being the contribution of the component in the tableau equation.

For the sake of simplicity, let us consider the LTI case of (3.27)

$$\begin{cases} A_{\mathsf{C}} C A_C^T \frac{dV}{dt} + A_{\mathsf{R}} S A_{\mathsf{R}}^T V + A_{\mathsf{L}} I_{\mathsf{L}} + A_{\mathsf{V}} I_{\mathsf{V}} + A_{\mathsf{J}} i(t) = 0 & (KCL), \\ -A_{\mathsf{L}}^T V + L \frac{dI_{\mathsf{L}}}{dt} = 0 & (BCE_{\mathsf{L}}), \\ A_{\mathsf{V}}^T V - u(t) = 0 & (BCE_{\mathsf{V}}), \end{cases} \tag{3.120}$$

which forms the following linear time invariant DAE equivalent to (3.34):

$$M\dot{X} = JX + U(t), \tag{3.121}$$

with

$$M = \begin{bmatrix} A_{\mathsf{C}} C A_C^T & 0 & 0 \\ 0 & L & 0 \\ 0 & 0 & 0 \end{bmatrix}, \tag{3.122}$$

$$J = \begin{bmatrix} -A_{\mathsf{R}} S A_{\mathsf{R}}^T & -A_{\mathsf{L}} & -A_{\mathsf{V}} \\ A_{\mathsf{L}}^T & 0 & 0 \\ A_{\mathsf{V}}^T & 0 & 0 \end{bmatrix}, \tag{3.123}$$

$$U(t) = \begin{bmatrix} -A_{\mathsf{J}} i(t) \\ 0 \\ -u(t) \end{bmatrix}. \tag{3.124}$$

Let us show several standard examples of stamp in the MNA.

1. Resistive element stamp which corresponds to an element of $A_{\mathsf{R}} S A_{\mathsf{R}}^T$ in J:

$$\left(\begin{array}{lcc} & V_i \text{ (column in } J) & V_j \text{ (column in } J) \\ \hline KCL(i) \text{ (line in } J) & -\frac{1}{R_k} & \frac{1}{R_k} \\ KCL(j) \text{ (line in } J) & \frac{1}{R_k} & -\frac{1}{R_k} \end{array} \right) \tag{3.125}$$

where $R_k = \frac{1}{S_k}$ is the resistance of the branch k.

2. Capacitive element stamp which corresponds to an element of $A_{\mathsf{C}} C A_{\mathsf{C}}^T$ in M:

$$\left(\begin{array}{lcc} & \dot{V}_i \text{ (column in } M\text{)} & \dot{V}_j \text{ (column in } M\text{)} \\ \hline KCL(i) \text{ (line in } M\text{)} & C_k & -C_k \\ KCL(j) \text{ (line in } M\text{)} & -C_k & C_k \end{array}\right) \tag{3.126}$$

where $C_k = \frac{1}{S_k}$ is the capacitance of the branch k.

3. Inductive element stamp which corresponds to an element of L in M and of A_{L}^T and A_{L} in J:

$$\left(\begin{array}{lcccc} & V_i \text{ (column in } J\text{)} & V_j \text{ (column in } J\text{)} & \dot{I}_{\mathsf{L},k} \text{ (column in } M\text{)} & I_{\mathsf{L},k} \text{ (column in J)} \\ \hline KCL(i) \text{ (line in } M \text{ or } J\text{)} & & & & -1 \\ KCL(j) \text{ (line in } M \text{ or } J\text{)} & & & & 1 \\ BCE_{\mathsf{L}}(k) \text{ (line in } M\text{)} & -1 & 1 & L_k & \end{array}\right) \tag{3.127}$$

where L_k is the inductance of the branch k.

4. Voltage independent sources ($u = f(t)$) which correspond to elements of A_{V}^T and A_{V} in J and an element in $U(t)$:

$$\left(\begin{array}{lcccc} & V_i & V_j & I_k & RSH \\ \hline KCL(i_c) & & & -1 & \\ KCL(j_c) & & & 1 & \\ BCE_{\mathsf{L}}(k) & -1 & 1 & & f(t) \end{array}\right). \tag{3.128}$$

5. Current independent sources ($i = f(t)$) which correspond to elements in $U(t)$:

$$\left(\begin{array}{lc} & RSH \\ \hline KCL(i_c) & f(t) \\ KCL(j_c) & -f(t) \end{array}\right). \tag{3.129}$$

The fully nonlinear case is treated in the same way by directly evaluating the components of F element by element. The matrices $M(X,t)$, $D(X,t)$ and $U(t)$ and the Jacobian matrices of F and U are filled following the same linearization procedure as in Sect. 3.5.1.3. The main consequence is the cheap evaluation of the Jacobians for a circuit which is a very sparse matrix. The evaluation is only slightly more expensive than evaluating the right-hand-side in (3.117). This is one of the reasons why the modified Newton method is almost never used in practice.

Chapter 4
Nonsmooth Modeling of Electrical Components

In the NonSmooth Dynamical Systems (NSDS) approach, the standard description of elements by means of explicit and smooth functions is enriched by new elements described by generalized equations. The characteristics of the electronic devices can be then nonsmooth and even multivalued. These new elements are called the electrical "nonsmooth elements". Some examples have already been studied in Chaps. 1 and 2. The description of nonsmooth components relies a lot on mathematical notions from Convex Analysis and the Mathematical Programming theory. A significant amount of informations on these aspects has already been provided in Chap. 2, starting from the very simple academic circuits examples presented in Chap. 1. In this chapter we take advantage of the material of the foregoing chapters to arrive at the general mathematical formalisms which are used in the NSDS approach to simulate the electrical circuits of Chaps. 7 and 8. The first subsections briefly recall some basic facts which are exposed in more details in Chap. 2, in particular Sects. 2.1 and 2.3. In a way similar to the foregoing chapter, the time argument is dropped from the state variables, in order to lighten the presentation of the dynamics.

4.1 General Nonsmooth Electrical Element

In order to precise what can be the constitutive laws of nonsmooth electronic devices, let us start with a very general definition of a generalized equation.

Definition 4.1 Let $y \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}^n$ be two vectors. A generalized equation (Robinson 1979) between y and λ is given by the following inclusion:

$$0 \in F(y, \lambda) + T(y, \lambda), \tag{4.1}$$

where $F : \mathbb{R}^{n \times n} \to \mathbb{R}^n$ is assumed to be a continuously differentiable mapping and $T : \mathbb{R}^{n \times n} \rightsquigarrow \mathbb{R}^n$ a multivalued mapping with a closed graph.

V. Acary et al., *Nonsmooth Modeling and Simulation for Switched Circuits*,
Lecture Notes in Electrical Engineering 69,
DOI 10.1007/978-90-481-9681-4_4, © Springer Science+Business Media B.V. 2011

This definition clearly extends (2.13) since a supplementary variable λ is added, and the multivalued mapping $T(\cdot,\cdot)$ may not be a normal cone. Using this definition of a generalized equation, a nonsmooth element will be defined in its full generality by the following constitutive equations:

$$\left.\begin{array}{l} y = g_{\mathsf{NS}}(I_{\mathsf{NS}}, U_{\mathsf{NS}}, \lambda, t) \\ 0 = h_{\mathsf{NS}}(I_{\mathsf{NS}}, U_{\mathsf{NS}}, \lambda, t) \end{array}\right] \text{Input/Output Relations,} \\ 0 \in F(y,\lambda,t) + T(y,\lambda,t) \quad]\,\text{Inclusion rule} \tag{4.2}$$

where $I_{\mathsf{NS}} \in \mathbb{R}^m$ and $U_{\mathsf{NS}} \in \mathbb{R}^{m-1}$ are two vectors which collect the currents and voltages at the m port of the nonsmooth elements. These currents and voltages are the controlling variables of the nonsmooth elements and are added to the set of unknowns of the circuit model. The variables y and λ are not necessarily physical values but are used to describe the internal behavior of the components.

The first two equations in (4.2) defined by the functions $g_{\mathsf{NS}}(\cdot)$ and $h_{\mathsf{NS}}(\cdot)$ will be called the input/output relations of the electrical nonsmooth components as they define the relations between the external variable $I_{\mathsf{NS}}, U_{\mathsf{NS}}$ with respect to the internal component variables y and λ. Inserting these new nonsmooth elements in the standard MNA (3.27), we obtain the following system:

$$\begin{cases} A_{\mathsf{C}} C(A^T V, t) A_C^T \frac{dV}{dt} + A_{\mathsf{C}} q_t(A_C^T V, t) + A_{\mathsf{R}} S(A_{\mathsf{R}}^T V, t) \\ \quad + A_{\mathsf{L}} I_{\mathsf{L}} + A_{\mathsf{V}} I_{\mathsf{V}} + A_{\mathsf{J}} i(A^T V, I_{\mathsf{L}}, I_{\mathsf{V}}, A^T \dot V, \dot I_{\mathsf{L}}, \dot I_{\mathsf{V}}, t) + A_{\mathsf{NS}} I_{\mathsf{NS}} = 0, \\ -A_{\mathsf{L}}^T V + L(I_{\mathsf{L}}, t) \frac{dI_{\mathsf{L}}}{dt} + \phi_t(I_{\mathsf{L}}, t) = 0, \\ A_{\mathsf{V}}^T V - u(I_{\mathsf{L}}, I_{\mathsf{V}}, A^T V, \dot I_{\mathsf{L}}, \dot I_{\mathsf{V}}, A^T \dot V, t) = 0, \\ y = g_{\mathsf{NS}}(I_{\mathsf{NS}}, A_{\mathsf{NS}}^T V, \lambda, t), \\ 0 = h_{\mathsf{NS}}(I_{\mathsf{NS}}, A_{\mathsf{NS}}^T V, \lambda, t), \\ 0 \in F(y,\lambda,t) + T(y,\lambda,t), \end{cases} \tag{4.3}$$

where A_{NS} is the incidence matrix of the branches concerned by nonsmooth element. In the sequel, we will specialize this general nonsmooth element and its associated generalized equation in order to precise what type of behavior may be modeled.

4.2 Nonsmooth Elements as Inclusions into the Subdifferential of Convex Functions and Variational Inequality (VI)

The set of relations in (4.2) is too generic to be useful in applications. It has to be refined in order to yield tractable mathematical formalisms. A standard example of a multivalued function with a closed graph is the subdifferential of a proper convex and lower semi-continuous function $\varphi : \mathbb{R}^n \to \mathbb{R}$, which is denoted by

$$\partial\varphi(x) = \{\gamma \in \mathbb{R}^n \mid \varphi(s) - \varphi(x) \geqslant \gamma^T(s-x) \text{ for all } s\}. \tag{4.4}$$

The subdifferential is the set of all subgradients, see Definition 2.22. There are other definitions of generalized gradients of nonsmooth functions, see for instance Clarke

(1975) and Mordukhovich (1994). For the sake of simplicity, we will consider only the subgradient in the sense of Convex Analysis, which has been presented in Chap. 2.

A first class of nonsmooth electrical components is given by the inclusion into the subdifferential of convex functions such as

$$-F(y) \in \partial\varphi(\lambda), \tag{4.5}$$

which we have named an electrical superpotential in Sect. 2.5.4. For the numerical tractability, this inclusion (4.5) is often transformed into a variational inequality of the form:

$$F^T(y)(x-\lambda) + \varphi(x) - \varphi(\lambda) \geqslant 0 \quad \text{for all } x, \tag{4.6}$$

which is also equivalent to

$$\lambda = P_\varphi(\lambda - F(y)), \tag{4.7}$$

where P_φ is the proximation operator associated to $\varphi(\cdot)$. The proximation operator is defined by the following equivalence

$$x = P_\varphi(z) \quad \Leftrightarrow \quad x = \underset{u}{\operatorname{argmin}} \frac{1}{2}\|z-u\|^2 + \varphi(u). \tag{4.8}$$

For the inclusion (4.7) we obtain a characterization of the nonsmooth element by means of an optimization problem given by

$$\lambda = \underset{u}{\operatorname{argmin}} \frac{1}{2}\|\lambda - F(y) - u\|^2 + \varphi(u). \tag{4.9}$$

All these equivalences are an extension of the results presented in Sects. 2.3.3, 2.3.4 and in Proposition 2.37. They allow one to work with different formulations of the same objects, which proves to be convenient for numerical simulation.

Example 4.2 (Relay with dead-zone component) Let us consider the following convex proper continuous function depicted in Fig. 4.1(a):

$$\varphi(x) = \begin{cases} -x & \text{for } x \leqslant -1, \\ 1 & \text{for } -1 \leqslant x \leqslant 1, \\ x & \text{for } x \geqslant 1. \end{cases} \tag{4.10}$$

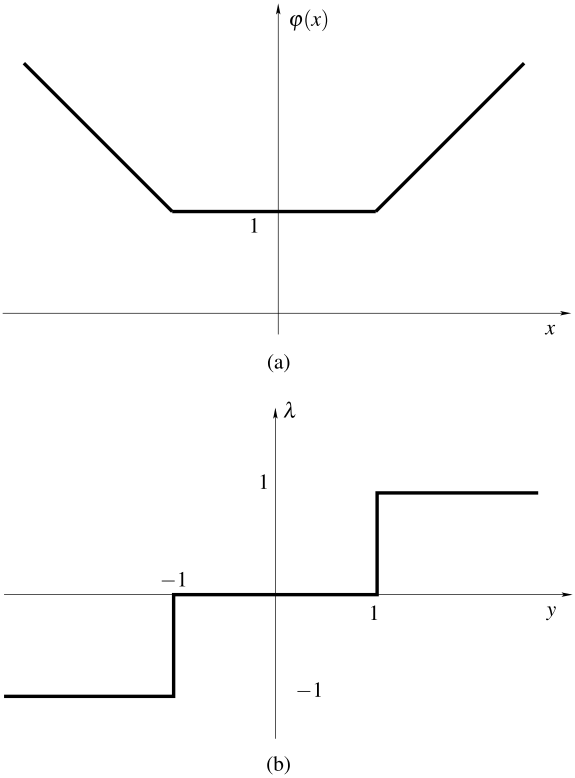
The sub-differential of $\varphi(\cdot)$ is depicted in Fig. 4.1(b), and it is given by

$$\partial\varphi(x) = \begin{cases} -1 & \text{for } x < -1, \\ [-1, 0] & \text{for } x = -1, \\ 0 & \text{for } -1 < x < 1, \\ [0, 1] & \text{for } x = 1, \\ 1 & \text{for } x > 1. \end{cases} \tag{4.11}$$

The model which is obtained in this example corresponds to a relay component with a dead-zone given by the following inclusion:

$$y \in \partial\varphi(\lambda). \tag{4.12}$$

This is a simple extension of the "basic" relay multifunction of Sect. 2.4.6. Notice that the multivalued mapping $y \mapsto \lambda$ of Fig. 4.1(b) is maximal monotone (see Sect. 2.1.2.2, in particular Theorem 2.34).



(a)

(b)

Fig. 4.1 The relay with dead-zone multivalued mapping

4.3 Nonsmooth Elements as Inclusions into Normal Cones and Variational Inequalities

As introduced in Sect. 2.1.2.3, another type of generalized equations is given under the form of an inclusion into a normal cone to a set C:

$$-F(y) \in N_{C(t,y)}(\lambda). \tag{4.13}$$

As for subgradients, there are also multiple definitions of the normal cone to a set, however we restrict ourselves in this book to normal cone to convex sets and we will assume that $C(t, y) \subset \mathbb{R}^n$ is a closed non empty convex set. The normal cone is given in this case by

$$N_C(\lambda) = \{s \in \mathbb{R}^n, s^T(x - \lambda) \leqslant 0 \text{ for all } x \in C(t, y)\}. \tag{4.14}$$

Due to the definition of the normal cone, the inclusion can be equivalently stated in terms of the following variational inequality:

$$F(y)^T(x - \lambda) \geqslant 0 \quad \text{for all } x \in C(t, y), \tag{4.15}$$

which is also equivalent to the projection form

$$\lambda = P_{C(t,y)}(\lambda - F(y)). \tag{4.16}$$

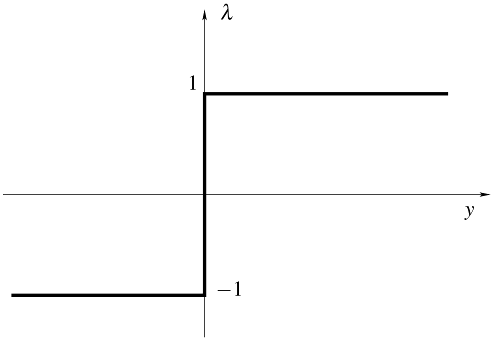


Fig. 4.2 The relay multivalued mapping

The projection form is particularly interesting for numerical applications in the design of numerical solvers.

Example 4.3 (Relay component) Let us consider the normal cone to the interval $[-1, 1]$ which is explicitly given by

$$N_{[-1,1]}(\lambda) = \begin{cases} \mathbb{R}_+ & \text{for } \lambda = 1, \\ 0 & \text{for } -1 < \lambda < 1, \\ \mathbb{R}_- & \text{for } \lambda = -1. \end{cases} \tag{4.17}$$

The inclusion $-F(y) \in N_{[-1,1]}(\lambda)$ yields therefore

$$\lambda = \begin{cases} 1 & \text{for } F(y) < 0, \\ [-1, 1] & \text{for } F(y) = 0, \\ -1 & \text{for } F(y) > 0. \end{cases} \tag{4.18}$$

When $F(\cdot)$ is the identity, the inclusion $y \in N_{[-1,1]}(\lambda)$ which is a multivalued mapping $\lambda \mapsto y$, models the relay multivalued mapping depicted in Fig. 4.2. Note that the equivalence in (2.90) holds and may be used to derive these results.

4.4 Complementarity Problems

Let us specialize a little bit more the formulation (4.13). If C is supposed to be a cone, the inclusion $-F(y) \in N_C(\lambda)$ is equivalent to a complementarity problem of the form

$$C^* \ni F(y) \perp \lambda \in C, \tag{4.19}$$

where C^* is the dual cone of C (see Remark 2.4). A particularly interesting cone is the non negative orthant, $\mathbb{R}^n_+$. In this case, we obtain standard complementarity problem of the form

$$0 \leqslant F(y) \perp \lambda \geqslant 0, \tag{4.20}$$

where the non-negativity inequality on vectors has to be understood componentwise. See Sect. 2.3 for more details on complementarity theory, and the relationships between complementarity problems and other formalisms.

4.5 The Linear Input/Output Relation Case

In the LTI case, the input/output relations in the constitutive equations in (4.2) of the nonsmooth elements reduce to

$$\begin{aligned} [y] &= K \begin{bmatrix} I_{\mathsf{NS}} \\ U_{\mathsf{NS}} \end{bmatrix} + L\lambda + a(t), \\ [0] &= E \begin{bmatrix} I_{\mathsf{NS}} \\ U_{\mathsf{NS}} \end{bmatrix} + F\lambda + b(t). \end{aligned} \tag{4.21}$$

In the numerical practice, most of the nonsmooth elements define a part or all of their currents and voltages I_{NS} and U_{NS} as an explicit function of λ. More precisely, we can perform a splitting of the vector I_{NS} and U_{NS} as:

$$I_{\mathsf{NS}} = \begin{bmatrix} \tilde{I}_{\mathsf{NS}} \\ \hat{I}_{\mathsf{NS}} \end{bmatrix}, \qquad U_{\mathsf{NS}} = \begin{bmatrix} \tilde{U}_{\mathsf{NS}} \\ \hat{U}_{\mathsf{NS}} \end{bmatrix}, \tag{4.22}$$

yielding an explicit rewriting of the linear input/output relations as:

$$\begin{aligned} [y] &= K \begin{bmatrix} I_{\mathsf{NS}} \\ U_{\mathsf{NS}} \end{bmatrix} + L\lambda + a(t), \\ \begin{bmatrix} \tilde{I}_{\mathsf{NS}} \\ \tilde{U}_{\mathsf{NS}} \end{bmatrix} &= \hat{E} \begin{bmatrix} \hat{I}_{\mathsf{NS}} \\ \hat{U}_{\mathsf{NS}} \end{bmatrix} + F\lambda + b(t). \end{aligned} \tag{4.23}$$

We will see that when this reformulation is possible the number of equations in the MNA can be reduced by substituting $\tilde{I}_{\mathsf{NS}}$ into the KCL.

4.5.1 *Some Instances of Linear Nonsmooth Components*

Together with the various formulations of the generalized equations, we obtain various types of linear components with complementarity and/or inclusion into sets. For instance

- Mixed linear complementarity component:

$$\begin{cases} [y] = K\left[\begin{smallmatrix} I_{\mathsf{NS}} \\ U_{\mathsf{NS}} \end{smallmatrix}\right] + L\lambda + a(t), \\ [0] = E\left[\begin{smallmatrix} I_{\mathsf{NS}} \\ U_{\mathsf{NS}} \end{smallmatrix}\right] + F\lambda + b(t), \\ 0 \leqslant y \perp \lambda \geqslant 0. \end{cases} \tag{4.24}$$

- Mixed linear relay component:

$$\begin{cases} [y] = K\left[\begin{smallmatrix} I_{\mathsf{NS}} \\ U_{\mathsf{NS}} \end{smallmatrix}\right] + L\lambda + a(t), \\ [0] = E\left[\begin{smallmatrix} I_{\mathsf{NS}} \\ U_{\mathsf{NS}} \end{smallmatrix}\right] + F\lambda + b(t), \\ -y \in N_{[-1,1]^p}(\lambda), \end{cases} \tag{4.25}$$

  where $[-1,1]^p = [-1,1] \times \cdots \times [-1,1]$, p times (here p generically denotes the dimension of the vectors y and λ).

Notice that (4.24) is a particular case of the MLCP in (2.19). In (4.25) one finds a generalized relay mapping, where the variables y and λ do not only satisfy the last inclusion, but also a set of linear constraints. One may use (2.25) or (2.14) in order to study further (4.25). The reason why we have written $-y$ and not y then clearly appears from a simple analogy. If the matrix $L = L^T > 0$ then one deduces from (4.25) that:

$$\lambda = \mathrm{proj}_L \left([-1, 1]^p; -L^{-1} K \begin{bmatrix} I_{\mathsf{NS}} \\ U_{\mathsf{NS}} \end{bmatrix} - L^{-1} a(t) \right) \tag{4.26}$$

with $[-1, 1]^p = C$ in (2.14). Inserting (4.26) into the second line of (4.25) yields a nonlinear equation for the unknowns I_{NS} and U_{NS}.

4.6 Generic Piecewise-Linear Components

As we said in Sect. 1.5, the literature on piecewise-linear modeling of electrical components is vast. A lot of expressions based mainly on absolute value function (Kang and Chua 1978; Chua and Ying 1983; Chua and Dang 1985) have been developed. In this section, we are interested in implicit expressions of piecewise-linear models in the complementarity framework. To this end, we report some of the main models introduced in the pioneering works of Leenaerts and Van Bokhoven (1998).

4.6.1 The First Model Description of van Bokhoven

In van Bokhoven (1981), a first model of piecewise-linear function $z = f_{\mathsf{pwl}}(x)$ is presented as

$$\begin{cases} z = Ax + B\lambda + f, \\ y = Cx + D\lambda + g, \\ 0 \leqslant y \perp \lambda \geqslant 0, \end{cases} \tag{4.27}$$

with $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{m\times k}$, $C \in \mathbb{R}^{k\times n}$ and $D \in \mathbb{R}^{k\times k}$. The second equation in (4.27) defines k hyperplanes in $\mathbb{R}^n$ parametrized by x and then in which state the model is. It may define 2^k polytopes of $\mathbb{R}^n$. In each polytope, a linear mapping defined by the first line of (4.27) is defined. The model of piecewise-linear component is then equivalent to (4.24).

Example 4.4 Let us consider a piecewise-linear continuous function $f_{\mathsf{pwl}} : \mathbb{R} \to \mathbb{R}$ defined by

$$f_{\mathsf{pwl}}(x) = \begin{cases} a_1 x + f_1 & \text{if } cx + g \leqslant 0, \\ a_2 x + f_2 & \text{if } cx + g \geqslant 0. \end{cases} \tag{4.28}$$

The continuity of $f_{\mathsf{pwl}}(\cdot)$ on the surface $cx + g = 0$ implies

$$a_1 x + f_1 = a_2 x + f_2 \quad \text{for } x = -\frac{g}{c}. \tag{4.29}$$

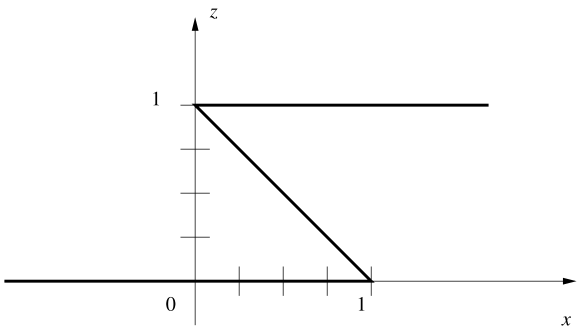


Fig. 4.3 Multivalued mapping. Example of Leenaerts and Van Bokhoven (1998)

A simple evaluation of the model (4.27) yields

$$A = [a_1], \qquad B = \left[-\frac{a_2 - a_1}{c} \right], \qquad C = [c], \qquad D = [1]. \tag{4.30}$$

Example 4.5 The model can also handle mappings that are not one-to-one as it has been shown in Leenaerts and Van Bokhoven (1998) with this example:

$$\begin{cases} z = [-1]x + \begin{bmatrix} -1 & 1 \end{bmatrix}\lambda + [1], \\ y = \begin{bmatrix} -1 \\ 1 \end{bmatrix}x + \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}\lambda + \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \\ 0 \leqslant y \perp \lambda \geqslant 0. \end{cases} \tag{4.31}$$

The corresponding function is depicted in Fig. 4.3.

4.6.2 The Second Model Description of van Bokhoven

In order to have more insight on the model, another model in presented in Leenaerts and Van Bokhoven (1998) with only definition of hyperplanes and a symmetrization of the unknowns x and z. The model is written as

$$\begin{cases} 0 = I_d z + Ax + B\lambda + f, \\ y = Dz + Cx + I_d\lambda + g, \\ 0 \leqslant y \perp \lambda \geqslant 0, \end{cases} \tag{4.32}$$

where I_d is the identity matrix.

Example 4.6 The function depicted in Fig. 4.3 can be written in the model (4.32) as:

$$\begin{cases} 0 = I_d z + [0]x + [-2\ 2]\lambda + [0], \\ y = \begin{bmatrix} -1 \\ -1 \end{bmatrix}z + \begin{bmatrix} -1/2 \\ -1/2 \end{bmatrix}x + I_d\lambda + \begin{bmatrix} -1/2 \\ 1 \end{bmatrix}, \\ 0 \leqslant y \perp \lambda \geqslant 0. \end{cases} \tag{4.33}$$

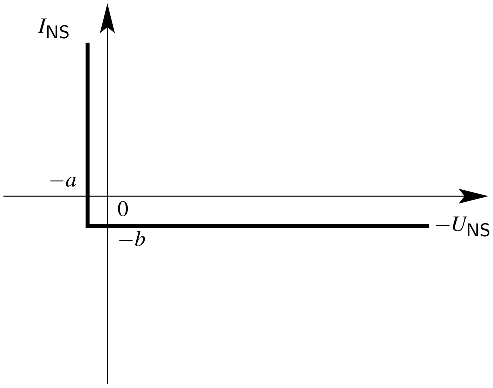


Fig. 4.4 Ideal diode with residual current and voltage

4.7 Special Instances of Nonsmooth Components

We give in this section some examples of ideal components or idealized components that can be put under the form described in the previous sections. Similarly as above, some of them (the simplest ones) have been introduced in Chap. 1 (the ideal diode, the Zener diode, the ideal switch).

4.7.1 Ideal Diode

Let us start with the most simple nonsmooth component which is the ideal diode. The ideal diode with residual current and voltage can be defined by

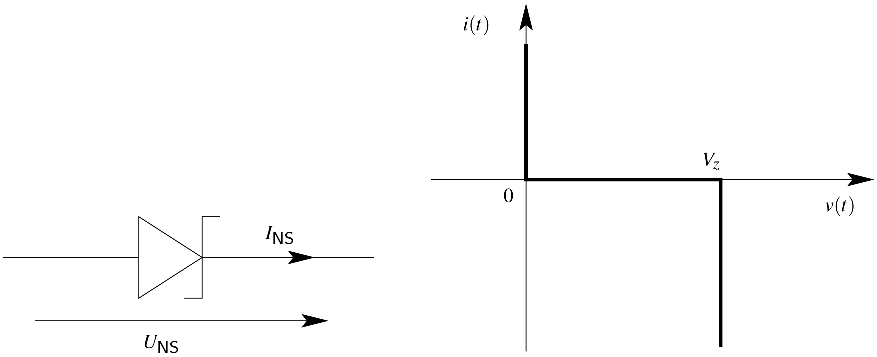
$$\begin{cases} y = -U_{\mathsf{NS}} + a, \\ 0 = -I_{\mathsf{NS}} + \lambda - b, \\ 0 \leqslant y \perp \lambda \geqslant 0. \end{cases} \tag{4.34}$$

The characteristic between I_{NS} and $-U_{\mathsf{NS}}$ is depicted in Fig. 4.4. The set of relations in (4.34) is a simple instance of (4.24). It can easily be rewritten equivalently as an inclusion into a normal cone, using (2.23).

4.7.2 Zener Diode

Let us consider, now, another electrical device: the ideal Zener diode whose schematic symbol is depicted in Fig. 4.5(a). A Zener diode is a type of diode that permits the 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_{NS} and the voltage U_{NS} can be seen in Fig. 4.5(b). The Zener diode can be put into the form of an inclusion into a sub-differential with

$$\varphi(x) = \begin{cases} V_z x & \text{if } x \geqslant 0, \\ 0 & \text{if } x < 0. \end{cases} \tag{4.35}$$



(a) The Zener diode schematic symbol.

(b) The ideal characteristic of a Zener diode.

Fig. 4.5 The Zener diode

One gets

$$U_{\mathsf{NS}} \in \partial\varphi(-I_{\mathsf{NS}}) \tag{4.36}$$

for some convex lower semi-continuous function $\varphi(\cdot)$. Equivalently, the Zener diode model can be written as an inclusion into the normal cone of the interval $[0, V_z]$ by

$$-I_{\mathsf{NS}} \in N_{[0,V_z]}(U_{\mathsf{NS}}), \tag{4.37}$$

so that $\varphi(\cdot)$ and $\psi_{[0,V_z]}(\cdot)$ are conjugate functions. The nonsmooth component can be therefore written as

$$\begin{cases} y = I_{\mathsf{NS}}, \\ 0 = -U_{\mathsf{NS}} + \lambda, \\ -y \in N_{[0,V_z]}(\lambda), \end{cases} \tag{4.38}$$

or

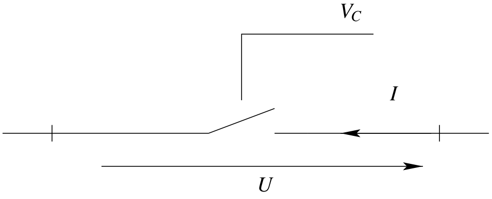
$$\begin{cases} y = U_{\mathsf{NS}}, \\ 0 = I_{\mathsf{NS}} + \lambda, \\ y \in \partial\varphi(\lambda). \end{cases} \tag{4.39}$$

One sees that (4.38) is a special instance of (4.25).

4.7.3 Ideal Switch

A switch is modeled by a piecewise-linear two-port element having a zero threshold V_c. When V_c is positive, the switch is ON, and is equivalent to a small resistance R_{on}. When V_c is negative, the switch is OFF, and is equivalent to a large resistance R_{off}. The notation for the current and the potentials at the ports of the

Fig. 4.6 Ideal switch



switch is depicted in Fig. 4.6 (compared with the convention in (1.46) here we have $U(t) = -u(t)$):

$$U = \begin{cases} R_{\text{on}} I & \text{if } V_c < 0, \\ R_{\text{off}} I & \text{if } V_c \geqslant 0. \end{cases}$$

The ideal switch is modeled with nonlinear complementarity relations:

$$\begin{cases} U = (\lambda_2 + R_{\text{on}})I, \\ y_1 = R_{\text{off}} - \lambda_2 - R_{\text{on}}, \\ y_2 = V_c + \lambda_1, \\ 0 \leqslant \binom{y_1}{y_2} \perp \binom{\lambda_1}{\lambda_2} \geqslant 0. \end{cases} \tag{4.40}$$

The set of relations in (4.40) is a particular instance of (4.24) with $I_{\mathsf{NS}} = [I]$, $U_{\mathsf{NS}} = [U]$ and $a(t)$ is the controlling tention $V_c(t)$ if these tension is an external source.

4.7.4 Explicit Ideal Switch. Glocker's Model

The model proposed by Glocker (2005) is based on the sign multifunction as follows

$$U \in V_T \operatorname{sgn}(-I), \tag{4.41}$$

where the parameter $V_T \geqslant 0$ is *a priori* chosen by the user depending on the status of the switch. For $V_T = 0$, the switch is a perfect conductor; we obtain Ifree, $U = 0$. For $V_T \to +\infty$, the switch is a perfect isolator with a saturation for $|V| = V_T$; we obtain for $|U| \geqslant V_T$ $I = 0$, U free.

In practice, the value of V_T is set *a priori* by the user and this necessarily yields an explicit evaluation in the numerical practice. In order to bind the value with a control voltage V_c as in Fig. 4.6, the following complementarity problem may be added

$$\begin{cases} y = \frac{1}{\alpha} V_T + V_c, \\ 0 \leqslant y \perp V_T \geqslant 0, \end{cases} \tag{4.42}$$

where $\alpha > 0$ is a user defined parameter. With this added complementarity, we obtain that if $V_c < 0$, $V_T = -\alpha V_c > 0$ and if $V_c \geqslant 0$, $V_T = 0$. The saturation voltage V_T is a function of V_c through the coefficient α that can be set sufficiently large to avoid saturation effect if needed.

Table 4.1 Parameters for Sah's model of nMOS transistor

Symbol	Definition	Typical values
μ	Mobility of majority carriers	$750\ \mathrm{cm^2\,V^{-1}\,s^{-1}}$ for a NMOS $250\ \mathrm{cm^2\,V^{-1}\,s^{-1}}$ for a PMOS
ϵ_{OX}	Permittivity of the silicon oxide	$\epsilon_{r\ \mathrm{SiO_2}} \cdot \epsilon_0$
$\epsilon_{r\ \mathrm{SiO_2}}$	Relative permittivity of the silicon oxide	$\epsilon_{r\ \mathrm{SiO_2}} \approx 3.9\ \mathrm{F\,m^{-1}}$
ϵ_0	Vacuum permittivity	$8.8542.10^{-12}\ \mathrm{F\,m^{-1}}$
t_{OX}	Oxide thickness	≈4 nm in a 180 nm technology
W	Channel width	≈130 nm in a 180 nm technology
L	Channel length	≈180 nm in a 180 nm technology
V_T	Threshold voltage	Depending on technology

The complete model of the ideal switch can be written as

$$\begin{cases} y = \frac{1}{\alpha}\lambda + V_c, \\ 0 \leqslant y \perp \lambda \geqslant 0, \\ U \in \lambda\,\mathrm{sgn}(-I), \end{cases} \tag{4.43}$$

with $\alpha > 0$ and

$$U_{\mathsf{NS}} = \begin{bmatrix} U \\ V_c \end{bmatrix}, \quad \text{and} \quad I_{\mathsf{NS}} = I. \tag{4.44}$$

4.7.5 MOSFET Transistor

One could benefit from a simplification of devices models (*e.g.* MOS models) in the form of a piecewise-linear representation instead of the complex model implemented in SPICE-like simulators. For instance, in Leenaerts and Van Bokhoven (1998), the authors considered the Sah model of the nMOS static characteristic:

$$I_{DS} = \frac{K}{2} \cdot \left(f(V_G - V_S - V_T) - f(V_G - V_D - V_T)\right), \tag{4.45}$$

where the function $f : \mathbb{R} \longrightarrow \mathbb{R}$ is defined as:

$$f(x) = \begin{cases} 0 & \text{if } x < 0, \\ x^2 & \text{if } x \geqslant 0, \end{cases} \tag{4.46}$$

and

$$K = \mu \frac{\epsilon_{OX}}{t_{OX}} \frac{W}{L}. \tag{4.47}$$

The parameters are defined in Table 4.1. The notation for the currents and the potentials at the ports of the nMOS is depicted in Fig. 4.7.

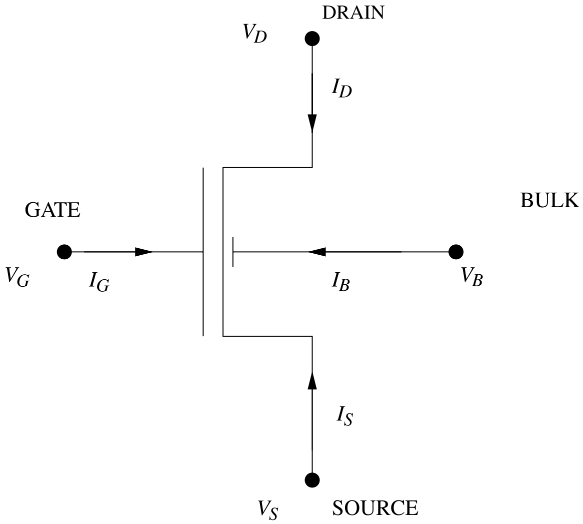


Fig. 4.7 nMOS transistor symbol

The piecewise and quadratic nature of this function is approximated by the following $s+2$ segments piecewise-linear function (Leenaerts and Van Bokhoven 1998):

$$f_{\mathsf{pwl}}(x) = \alpha_i x + \beta_i, \quad \text{for } a_i \leqslant x \leqslant a_{i+1},\ i = -1, \dots, s+1, \tag{4.48}$$

with $a_{-1} = -\infty$ and $a_{s+1} = +\infty$. The complete model of the piecewise-linear nMOS transistor with $s+2$ segments in (4.48) can be recast under the following mixed linear complementarity form:

$$\begin{aligned}
y(t) &= \begin{bmatrix} \underbrace{\begin{matrix} 0 & \dots & 0 \\ -b & \dots & -b \end{matrix}}_{\times s+1} & \underbrace{\begin{matrix} -b & \dots & -b \\ 0 & \dots & 0 \end{matrix}}_{\times s+1} \end{bmatrix}^T U_{\mathsf{NS}}(t) + \lambda(t) \\
&\quad + [h_1 \ \dots \ h_{s-1} \ h_1 \ \dots \ h_{s-1}]^T, \\
0 &= \mathsf{I}_3\, I_{\mathsf{NS}}(t) + \begin{bmatrix} -c_1 & \dots & -c_{s-1} & c_1 & \dots & c_{s-1} \\ 0 & 0 & 0 & 0 & & 0 \\ c_1 & \dots & c_{s-1} & -c_1 & \dots & -c_{s-1} \end{bmatrix} \lambda(t), \\
0 &\leqslant y(t) \perp \lambda(t) \geqslant 0, \\
U_{\mathsf{NS}} &= \begin{bmatrix} V_{GD}(t) = V_G(t) - V_D(t) \\ V_{GS}(t) = V_G(t) - V_S(t) \end{bmatrix}, \qquad I_{\mathsf{NS}} = \begin{bmatrix} I_D(t) \\ I_G(t) \\ I_S(t) \end{bmatrix}.
\end{aligned} \tag{4.49}$$

The parameters are given as follows: $b = \frac{\mathsf{K}}{2}$, $h_i = b(V_T + a_i)$, $i = 1 \dots s$. The values c_i are computed from the linear approximation in (4.48). Using some basic convex analysis, one obtains the compact formulation of (4.49):

$$\begin{cases} -y(t) \in N_K(\lambda(t)), \\ y(t) = BU_{\mathsf{NS}}(t) + \lambda(t) + h(t), \\ 0 = I_{\mathsf{NS}}(t) + C\lambda(t) \end{cases} \tag{4.50}$$

with $K = (\mathbb{R}_+)^{2(s+1)}$. In the case of the MOSFET transistor, the inclusion is an equality as expected since its piecewise-linear characteristic is single valued. The pMOS transistor is represented in the same way, changing the values of h_i, $i(t)$ to $-i(t)$ and b to $-b$.

Contrarily to the other models of components, the complementarity variables y and λ in (4.50) have no direct physical meaning. They are just slackness variables which permit us to express the presence of the operating point in the different segments of the model. For more details on the construction and the calibration of such a model, we refer to Leenaerts and Van Bokhoven (1998).

Remark 4.7 The piecewise-linear model in (4.48) has $s + 2$ segments. Multiple choices are possible in order to adjust the number of slack variables and consequently the size of the OSNSP-MLCP to be solved at each step with respect to the accuracy. In practice one should therefore be very careful about choosing a reasonable piecewise-linear approximation of the devices so that the MLCP size does not increase too much.

For instance, the function $f(\cdot)$ may be approximated by the following 6-segment piecewise-linear function in Leenaerts and Van Bokhoven (1998) (see Fig. 4.8):

$$f_{\mathsf{pwl}}(x) = \begin{cases} 0 & \text{if } x < 0, \\ 0.09 \cdot x & \text{if } 0 \leqslant x < 0.1, \\ 0.314055 \cdot x - 0.0224055 & \text{if } 0.1 \leqslant x < 0.2487, \\ 0.780422 \cdot x - 0.138391 & \text{if } 0.2487 \leqslant x < 0.6185, \\ 1.94107 \cdot x - 0.856254 & \text{if } 0.6185 \leqslant x < 1.5383, \\ 4.82766 \cdot x - 5.29668 & \text{if } 1.5383 \leqslant x \end{cases}$$

and the coefficients are computed as

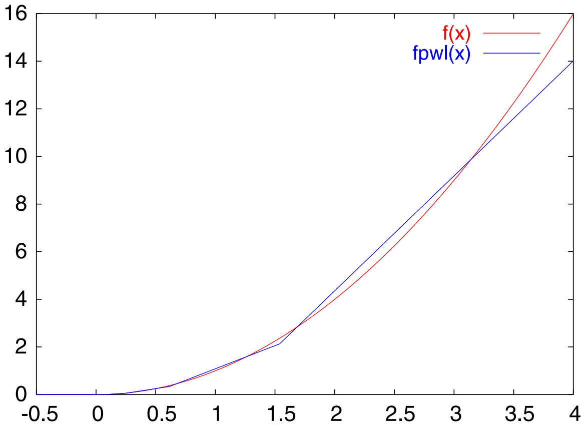
$$\begin{cases} c_1 = 0.09, \qquad c_2 = 0.2238, \qquad c_3 = 0.4666, \\ c_4 = 1.1605, \qquad c_5 = 2.8863, \\ a_1 = 0, \qquad a_2 = 0.1, \qquad a_3 = 0.2487, \\ a_4 = 0.6182, \qquad a_5 = 1.5383 \end{cases} \tag{4.51}$$

with the following parameter values

$$\begin{aligned} &\epsilon_{r\ \mathrm{SiO_2}} = 3.9, \\ &t_{OX} = 20\ \mathrm{nm}, \\ &\mu = 750\ \mathrm{cm}^2\,\mathrm{V}^{-1}\,\mathrm{s}^{-1}, \\ &W = 1\ \mu\mathrm{m}, \\ &L = 1\ \mu\mathrm{m}, \\ &V_T = 1\ \mathrm{V}. \end{aligned}$$

The relative error between $f(\cdot)$ and $f_{\mathsf{pwl}}(\cdot)$ is kept below 0.1 for $0.1 \leqslant x < 3.82$. The absolute error is less than $2 \cdot 10^{-3}$ for $0 \leqslant x < 0.1$ and 0 for negative x. In practice, the values of V_G, V_S, V_D, V_T in logic integrated circuits allow a good approximation of $f(\cdot)$ by $f_{\mathsf{pwl}}(\cdot)$. Figure 4.9 displays the static characteristic $I_{DS}(V_{GS}, V_{DS})$ of an nMOS obtained with the SPICE level 1 model and the

Fig. 4.8 Piecewise-linear approximation of $f(\cdot)$



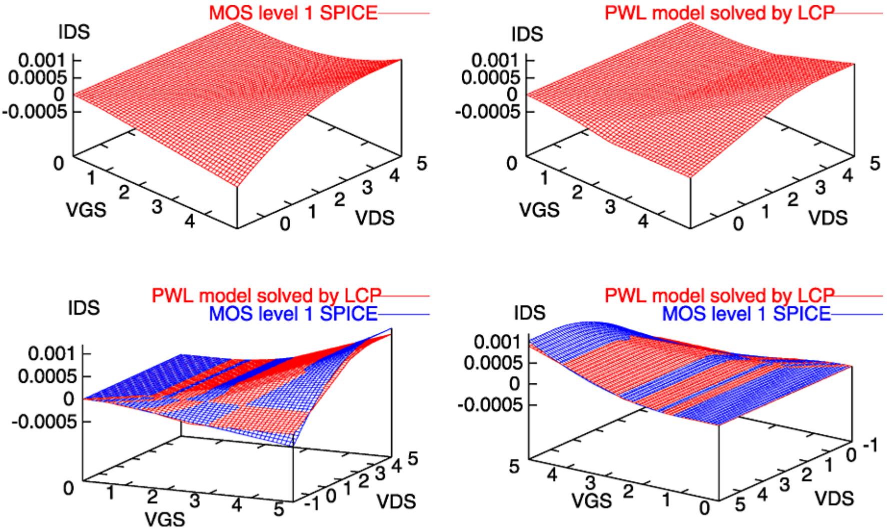


Fig. 4.9 Static characteristic of an nMOS transistor with a simple piecewise-linear model and SPICE level 1 model

piecewise-linear approximation of the Sah model. Bottom figures include both models results with two different viewpoints to display the regions where differences appear.

The largest differences occur for large V_{DS} and either a large positive V_{GS} or a large positive V_{GD}. Indeed this yields a high value of x in one of the $f(x)$ in $f(V_{GS} - V_T) - f(V_{GD} - V_T)$ and the linear approximation $f_{\mathrm{pwl}}(\cdot)$ differs from $f(\cdot)$. For small values of V_{DS}, errors compensate due to the difference $f(V_{GS} - V_T) - f(V_{GD} - V_T)$. For simulating cMOS logic circuits, the useful operating region is the square $(V_{GS}, V_{DS}) \in [0,5]^2$ and the error is moderate. The piecewise-linear model results in Fig. 4.9 were reached by an LCP algorithm.

This means that the definition of $f_{\mathsf{pwl}}(\cdot)$ is turned into an LCP formulation. For a given value of (V_D, V_G, V_S), the λ values corresponding to the intervals in which $V_{GS} - V_T$ and $V_{GD} - V_T$ fall are computed, allowing then to compute I_{DS}.

Once again one recognizes that (4.49) is a particular instance of (4.24).

4.7.6 Nonlinear and Nonsmooth MOS Transistor

Like it was described in (4.45), it consists in modeling the MOS considering two domains, $V_{GS}(t) > V_T$ and $V_{GD}(t) > V_T$. In this case, the MOS design can be described with the equations (4.52) below. Like in the previous ideal model, $I_G(t)$ is supposed equal to zero. It leads to define $I_{DS}(t) = I_D(t) = -I_S(t)$, the current through the MOS transistor:

$$\begin{cases} I_{DS}(t) = I_1(t) + I_2(t), \\ I_1(t) = \begin{cases} \frac{K}{2}(V_{GS}(t) - V_T)^2 & \text{if } V_{GS}(t) > V_T, \\ 0 & \text{if } V_{GS}(t) \leqslant V_T, \end{cases} \\ I_2(t) = \begin{cases} \frac{-K}{2}(V_{GD}(t) - V_T)^2 & \text{if } V_{GS}(t) > V_T, \\ 0 & \text{if } V_{GD}(t) \leqslant V_T. \end{cases} \end{cases} \tag{4.52}$$

It is equivalent to the developed system:

$$I_{DS}(t) = \begin{cases} 0 & \text{if } V_{GS}(t) < V_T \wedge V_{GD}(t) < V_T, \\ \frac{K}{2}(V_{GS}(t) - V_T)^2 & \text{if } V_{GS}(t) > V_T \wedge V_{GD}(t) < V_T, \\ K((V_{GS}(t) - V_T)V_{DS}(t) \\ \quad -\frac{1}{2}V_{DS}(t)^2) & \text{if } V_{GS}(t) > V_T \wedge V_{GD}(t) > V_T, \\ \frac{-K}{2}(V_{GD}(t) - V_T)^2 & \text{if } V_{GS}(t) < V_T \wedge V_{GD}(t) > V_T. \end{cases}$$

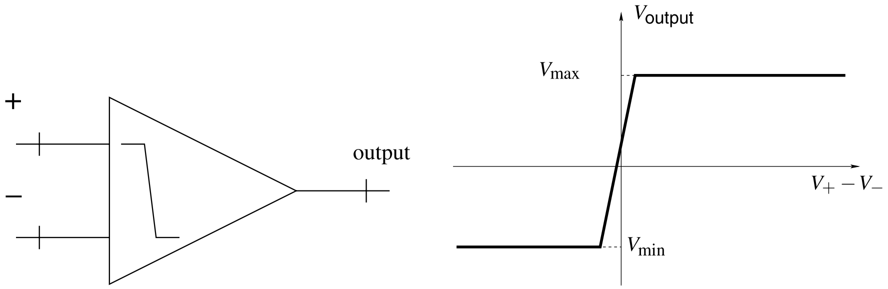
The system (4.52) can be reformulated as the complementarity system:

$$\begin{cases} I_{DS}(t) = \frac{K}{2}(\lambda_4(V_{GS}(t) - V_T)^2 - \lambda_2(V_{GD}(t) - V_T)^2), \\ \begin{pmatrix} y_1(t) \\ y_2(t) \\ y_3(t) \\ y_4(t) \end{pmatrix} = \begin{pmatrix} 1-\lambda_2 \\ V_T - V_{GD}(t) + \lambda_1 \\ 1-\lambda_4 \\ V_T - V_{GS}(t) + \lambda_3(t) \end{pmatrix}, \\ 0 \leqslant \begin{pmatrix} \lambda_1(t) \\ \lambda_2(t) \\ \lambda_3(t) \\ \lambda_4(t) \end{pmatrix} \perp \begin{pmatrix} y_1(t) \\ y_2(t) \\ y_3(t) \\ y_4(t) \end{pmatrix} \geqslant 0. \end{cases} \tag{4.53}$$

The set of relations in (4.53) can be recast into the general form in (4.2) with

$$U_{\mathsf{NS}}(t) = \begin{bmatrix} V_{GD}(t) \\ V_{GS}(t) \end{bmatrix}, \qquad I_{\mathsf{NS}}(t) = [I_{DS}(t)].$$

Comparison between the piecewise-linear model and the piecewise-nonlinear model is given in Sect. 8.1.3.6.



(a) Comparator symbol

(b) Comparator behavior

Fig. 4.10 The comparator component

4.7.7 Comparator Component

This device is usually modeled with a differentiable single-valued function. For example, the arctangent function can be used. In this section we propose to model a comparator with a piecewise-linear function:

$$V_{\mathrm{output}} = \begin{cases} V_{\mathrm{max}} & \text{if } V_+ - V_- < \epsilon, \\ V_{\mathrm{min}} & \text{if } V_+ - V_- > -\epsilon. \end{cases}$$

The component symbol and behavior are described in Fig. 4.10. Note that it differs from the relay described in Fig. 4.2. In this model, there is a ramp between the two constant parts. Such a ramp is sometimes called a regularization of the relay nonsmooth multifunction. The piecewise-linear model of this design is described with the following complementarity system:

$$\begin{cases} y_1 = V_+ - V_- + \lambda_1 + \epsilon, \\ y_2 = V_+ - V_- + \lambda_2 - \epsilon, \\ V_{\mathrm{output}} = V_{\mathrm{max}} + \frac{V_{\mathrm{max}} - V_{\mathrm{min}}}{2\epsilon}(\lambda_1 - \lambda_2), \\ 0 \leqslant y \perp \lambda \geqslant 0. \end{cases} \tag{4.54}$$

The set of relations in (4.54) can be recast into the general form in (4.2) with I_{ns} empty and

$$U_{\mathrm{NS}} = \begin{bmatrix} V_{\mathrm{output}} \\ V_+ - V_- \end{bmatrix}.$$

Chapter 5
Time-Stepping Schemes and One Step Solvers

5.1 Summary of the Mathematical Formalisms

It has been seen in Chaps. 1, 2 and 3 that electrical circuits with nonsmooth multivalued electronic devices, can be recast under various mathematical formalisms (see Sect. 2.7 for a summary). For the sake of the numerical integration of those circuits, one needs a small set of general formulations which are suitable for a subsequent time-discretization. In other words, the simple examples that are analyzed in details in Chaps. 1 and 2 possess a too simple dynamics to be characteristic representatives of the general issue of nonsmooth circuits. Especially, the material of Chap. 3 teaches us that DAEs are ubiquitous in circuits (a well-known fact, indeed). It is therefore necessary to obtain mathematical formalisms that incorporate not only the nonsmooth and multivalued models of the electronic devices (ideal diodes, Zener diodes, etc.), but also the equations obtained from the MNA (see Sects. 3.5 and 3.6).

5.1.1 Nonsmooth DAE Formulation. Differential Generalized Equation (DGE)

Starting from (3.34), the extended MNA formulation with nonsmooth components and nonlinear behavior can be written as:

$$\boxed{\begin{array}{ll}
\multicolumn{2}{l}{\textsf{Problem} \quad \textsf{(DGE)}} \\
M(X,t)\dot{X} = D(X,t) + U(t) + R & \text{] Differential Algebraic Equations} \\
\left.\begin{array}{l} y = G(X,\lambda,t) \\ R = H(X,\lambda,t) \end{array}\right] & \begin{array}{l}\text{Input/output relations} \\ \text{on nonsmooth components}\end{array} \\
0 \in F(y,\lambda,t) + T(y,\lambda,t) & \text{] Generalized equation} \\
X = [V^T, I_{\mathsf{L}}^T, I_{\mathsf{V}}^T, I_{\mathsf{NS}}^T]^T & \text{] Variable definition}
\end{array}} \tag{5.1}$$

V. Acary et al., *Nonsmooth Modeling and Simulation for Switched Circuits,*
Lecture Notes in Electrical Engineering 69,
DOI 10.1007/978-90-481-9681-4_5, © Springer Science+Business Media B.V. 2011

The nonlinear feature refers here to the differential-algebraic part of the dynamics, *i.e.* the first line in (5.1). Compared to (3.35), (3.36), (3.37) and (4.2), the definitions of M, F, U, G and H are extended as follows:

$$M(X,t)=\begin{bmatrix} A_{\mathsf{C}}C(A^TV,t)A_C^T & 0 & 0 & 0\\ 0 & L(I_{\mathsf{L}},t) & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}, \tag{5.2}$$

$$D(X,t)=\begin{bmatrix} -A_{\mathsf{C}}q_t(A_C^TV,t)-A_{\mathsf{R}}S(A_{\mathsf{R}}^TV,t)-A_{\mathsf{L}}I_{\mathsf{L}}-A_{\mathsf{V}}I_{\mathsf{V}}\\ A_{\mathsf{L}}^TV-\phi_t(I_{\mathsf{L}},t)\\ A_{\mathsf{V}}^TV\\ 0\end{bmatrix}, \tag{5.3}$$

$$U(t)=\begin{bmatrix} -A_{\mathsf{J}}i(t)\\ 0\\ -u(t)\\ 0\end{bmatrix}, \tag{5.4}$$

$$\begin{aligned} R=H(X,\lambda,t)&=\begin{bmatrix} -A_{\mathsf{NS}}I_{\mathsf{NS}}\\ 0\\ 0\\ h_{\mathsf{NS}}(I_{\mathsf{NS}},A_{\mathsf{NS}}^TV,\lambda,t)\end{bmatrix},\\ G(X,\lambda,t)&=[g_{\mathsf{NS}}(I_{\mathsf{NS}},A_{\mathsf{NS}}^TV,\lambda,t)]. \end{aligned} \tag{5.5}$$

The functions $h(\cdot)$, $g(\cdot)$ and $F(\cdot)$ and the multivalued mapping $T(\cdot)$ are built by concatenation of the corresponding functions of the nonsmooth components, *i.e.*:

$$\begin{cases} h_{\mathsf{NS}}(I_{\mathsf{NS}},A_{\mathsf{NS}}^TV,\lambda,t)=\left[h_k(I_{\mathsf{NS}},A_{\mathsf{NS}}^TV,\lambda,t),\ k\in\mathsf{NS}\right]^T,\\ g_{\mathsf{NS}}(I_{\mathsf{NS}},A_{\mathsf{NS}}^TV,\lambda,t)=\left[g_k(I_{\mathsf{NS}},A_{\mathsf{N}}^TV,\lambda,t),\ k\in\mathsf{NS}\right]^T,\\ F(I_{\mathsf{NS}},A_{\mathsf{NS}}^TV,\lambda,t)=\left[F_k(I_{\mathsf{NS}},A_{\mathsf{NS}}^TV,\lambda,t),\ k\in\mathsf{NS}\right]^T,\\ T(I_{\mathsf{NS}},A_{\mathsf{NS}}^TV,\lambda,t)=\left[T_k(I_{\mathsf{NS}},A_{\mathsf{NS}}^TV,\lambda,t),\ k\in\mathsf{NS}\right]^T. \end{cases} \tag{5.6}$$

In the linear-time-invariant case, the extended MNA formulation with nonsmooth components can be written as follows:

Problem (DGE)$_{\mathsf{LTI}}$

$$\begin{array}{ll} M\dot{X}=JX+U(t)+R &]\ \text{Differential Algebraic Equations}\\ \begin{array}{l} y=CX+D\lambda+a(t)\\ R=B\lambda \end{array} & \left.\right]\begin{array}{l}\text{Input/output relations}\\ \text{on nonsmooth components}\end{array}\\ 0\in F(y,\lambda,t)+T(y,\lambda,t) &]\ \text{Generalized equation}\\ X=[V^T,I_{\mathsf{L}}^T,I_{\mathsf{V}}^T,I_{\mathsf{NS}}^T]^T &]\ \text{Variable definition} \end{array} \tag{5.7}$$

Similarly to the nonlinear case the definitions of M, J and U are extended as follows:

$$\begin{aligned}
M &= \begin{bmatrix} A_{\mathrm{C}} C A_C^T & 0 & 0 & 0 \\ 0 & L & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \\
J &= \begin{bmatrix} -A_{\mathrm{R}} S A_{\mathrm{R}}^T & -A_{\mathrm{L}} & -A_{\mathrm{V}} & -A_{\mathrm{NS}} \\ A_{\mathrm{L}}^T & 0 & 0 & 0 \\ A_{\mathrm{V}}^T & 0 & 0 & 0 \\ E_U A_{\mathrm{NS}}^T & 0 & 0 & E_I \end{bmatrix}, \\
U(t) &= \begin{bmatrix} -A_{\mathrm{J}} i(t) \\ 0 \\ -u(t) \\ b \end{bmatrix}.
\end{aligned} \tag{5.8}$$

The matrix E_U and E_I are introduced as sub-matrices of E such that

$$E \begin{bmatrix} I_{\mathrm{NS}} \\ U_{\mathrm{NS}} \end{bmatrix} = E_U U_{\mathrm{NS}} + E_I I_{\mathrm{NS}}, \tag{5.9}$$

and are built from the concatenation of each nonsmooth element. The matrices B, C and D are defined by:

$$B = \begin{bmatrix} 0 \\ 0 \\ 0 \\ F \end{bmatrix}, \qquad C = [\, K_U A^T \quad 0 \quad 0 \quad K_I \,], \quad \text{and} \quad D = [L]. \tag{5.10}$$

The matrices K_U and K_I are introduced as sub-matrices of K such that:

$$K \begin{bmatrix} I_{\mathrm{NS}} \\ U_{\mathrm{NS}} \end{bmatrix} = K_U U_{\mathrm{NS}} + K_I I_{\mathrm{NS}}. \tag{5.11}$$

5.1.2 The Semi-Explicit Nonsmooth DAE: Semi-Explicit DGE

Starting from (3.70), the extended MNA formulation with nonsmooth component and nonlinear behavior can be written as follows:

Problem (SEDGE)

$$\begin{array}{l}
\left.\begin{array}{l} \dot{x} = N^{-1}(x,t)[f(x,z,t)] + r_1 \\ 0 = g(x,z,t) + r_2 \end{array}\right] \begin{array}{l}\text{Differential Algebraic}\\ \text{Equations}\end{array} \\[1ex]
\left.\begin{array}{l} 0 = h_{\mathrm{NS}}(x,z,z_{\mathrm{NS}},\lambda,t) \\ y = g_{\mathrm{NS}}(x,z,z_{\mathrm{NS}},\lambda,t) \\ r = [r_1, r_2]^T = H z_{\mathrm{NS}} \end{array}\right] \begin{array}{l}\text{Input/output relations}\\ \text{on nonsmooth components}\end{array} \\[1ex]
0 \in F(y,\lambda,t) + T(y,\lambda,t) \quad]\ \text{Generalized equation} \\[1ex]
\left.\begin{array}{l} x = [I_{\mathrm{L}}^T, U_{\mathrm{C}}^T]^T \\ \text{and } z = [\hat{V}^T, I_{\mathrm{V}}^T, I_{\mathrm{C_L}}^T]^T, z_{\mathrm{NS}} = [I_{\mathrm{NS}}] \end{array}\right] \text{Variable definition}
\end{array} \tag{5.12}$$

The definitions of the mappings N, f, g are the same as in (3.105), (3.106) and (3.107) with the substitution (3.108). Only the matrix H remains to be defined by:

$$H \triangleq \begin{bmatrix} -\tilde{A}_{\mathsf{CF}}^{-1}\tilde{A}_{\mathsf{NS}} \\ 0 \\ 0 \\ 0 \\ (\hat{A}_{\mathsf{NS}} - \hat{A}_{\mathsf{CF}}\tilde{A}_{\mathsf{CF}}^{-1}\tilde{A}_{\mathsf{NS}}) \\ 0 \end{bmatrix}. \tag{5.13}$$

With the LTI variant of nonsmooth elements, the semi-explicit formulation described in (5.12) leads to the system:

Problem (SEDGE)$_{\mathsf{LTI}}$

$$\begin{array}{ll}
\left.\begin{array}{l} \dot{x} = J_{1x}x + J_{1z}z + U(t) + r_1 \\ 0 = J_{2x}x + J_{2z}z + U_2(t) + r_2 \end{array}\right] & \text{Differential Algebraic Equations} \\
\left.\begin{array}{l} y = K_x x + K_z z + K_{z_{\mathsf{NS}}} z_{\mathsf{NS}} + L_\lambda \lambda + a(t) \\ 0 = E_x x + E_z z + E_{z_{\mathsf{NS}}} z_{\mathsf{NS}} + F_\lambda \lambda + b(t) \\ r = [r_1, r_2]^T = H z_{\mathsf{NS}} \end{array}\right] & \text{Input/output relations on nonsmooth components} \\
0 \in F(y,\lambda,t) + T(y,\lambda,t) \quad] & \text{Generalized equation} \\
x = [I_{\mathsf{L}}^T, U_{\mathsf{C}}^T]^T \text{ and } z = [V^T, I_{\mathsf{V}}^T, I_{\mathsf{CL}}^T, I_{\mathsf{NS}}^T]^T \ ] & \text{Variable definition}
\end{array} \tag{5.14}$$

For most of the nonsmooth electrical components, the good choice of y and λ permits to write the variable $z_{\mathsf{NS}} = I_{\mathsf{NS}}$ as an explicit function of x, z and λ. Its substitution in (5.14) yields the following simplified system

$$\begin{cases} \dot{x} = J_{1x}x + J_{1z}z + u(t) + r_1, \\ 0 = J_{2x}x + J_{2z}z + D_1\lambda + c(t), \\ y = C_x x + C_z z + D_2\lambda + a(t), \\ r_1 = B\lambda, \\ 0 \in F(y,\lambda,t) + T(y,\lambda,t). \end{cases} \tag{5.15}$$

With the new definition of a new variable $\hat{y}$ and $\hat{\lambda}$ as

$$\hat{y} = \begin{bmatrix} \hat{y}_e \\ \hat{y}_i \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix}, \qquad \hat{\lambda} = \begin{bmatrix} z \\ \lambda \end{bmatrix}, \tag{5.16}$$

the system (5.15) can be compacted as

$$\begin{cases} \dot{x} = Ax + u(t) + \hat{r}, \\ \hat{y} = Cx + D\lambda + \hat{a}(t), \\ \hat{r} = \hat{B}\hat{\lambda}, \\ 0 \in \hat{F}(\hat{y},\hat{\lambda},t) + \hat{T}(\hat{y},\hat{\lambda},t). \end{cases} \tag{5.17}$$

5.2 Principles of the Numerical Time-Integration Scheme

For the sake of readability, the main principles of the numerical time integration scheme are exposed on the following LTI system:

Problem (P)$_{\mathrm{LTI}}$

$$\begin{aligned}
&\dot{x} = Ax + u(t) + r && \text{] Differential Equations}\\
&y = Cx + D\lambda + a(t) && \text{] Input/output relations}\\
&R = B\lambda && \ \text{on nonsmooth components}\\
&0 \in y + N_K(\lambda) && \text{] Generalized equation}\\
&x(t_0) = x_0 && \text{] Initial conditions}
\end{aligned} \tag{5.18}$$

and its associated nonlinear version:

Problem (P)

$$\begin{aligned}
&\dot{x} = f(x,t) + u(t) + r && \text{] Differential Equations}\\
&y = g(x,\lambda,t) && \text{] Input/output relations}\\
&R = h(x,\lambda,t) && \ \text{on nonsmooth components}\\
&0 \in y + N_K(\lambda) && \text{] Generalized equation}\\
&x(t_0) = x_0 && \text{] Initial conditions}
\end{aligned} \tag{5.19}$$

The set K is assumed to be a non empty polyhedral convex set. The problems (5.18) and (5.19) are special instances of the problems presented in Sect. 5.1 and especially (5.17). Besides their simplicity, they are interesting because we are able to state some of their mathematical properties like their relative degree and then the expected nonsmoothness of their solution. The more general cases will be developed in Sect. 5.3.

The time integration methods that will be presented in the sequel are only *event-capturing time-stepping schemes*, or shortly called, time-stepping schemes. These systems can also be integrated with event-tracking schemes, also called event-driven schemes. The advantages and the drawbacks of these two classes of methods are pointed out in Acary and Brogliato (2008). Briefly speaking, event-capturing schemes are well suited when the system size is quite large with a lot of possible modes and the number of events is also large. For the simulation of switched electrical circuits, our choice is clearly to promote event-capturing schemes.

The two main principles for the design of the event-capturing schemes are:

1. The fully implicit evaluation of the generalized equation, also named the inclusion rule in (5.18) and (5.19).
2. A consistent evaluation of the unknown variables and their derivatives according to their smoothness. For instance, time-stepping schemes must not approximate high order time-derivatives of functions which are not sufficiently smooth or must not try to point-wisely evaluate distributions.

These two main principles will be illustrated and implemented in the next sections on the systems (5.18) and (5.19). The question of consistency for a certain level of smoothness of the solution is crucial in the design of a time-stepping scheme. This is the reason why the time-stepping schemes are presented according to the expected smoothness of the solution. In the meantime, the smoothness of the solution will be related to the relative degree of the systems and the consistency of the initial conditions.

The following notation is used throughout this part. We denote by $0 = t_0 < t_1 < \cdots < t_k < \cdots < t_N = T$ a finite partition (or a subdivision) of the time interval $[0, T]$ $(T > 0)$. The integer N stands for the number of time intervals in the subdivision. The length of a time step is denoted by $h_k = t_{k+1} - t_k$. For simplicity sake, we consider only in the sequel a constant time length $h = h_k$ $(0 \leqslant k \leqslant N - 1)$. Then $N = \frac{T}{h}$. The approximation of $f(t_k)$, the value of a real function $f(\cdot)$ at the time t_k, is denoted by f_k. For $\theta \in [0, 1]$, the notation $f_{k+\theta}$ stands for $\theta f_{k+1} + (1 - \theta) f_k$.

5.2.1 Time-Stepping Solutions for a Solution of Class C^1

The problem $(\mathsf{P})_{\mathsf{LTI}}$ has a unique C^1 trajectory $x(t)$ if the variable $\lambda(t)$ can be estimated as a Lipschitz continuous function of x. Then $\dot{x}$ is also Lipschitz continuous. For $K = \mathbb{R}^m_+$, we obtain a C^1 trajectory when the relative degree is equal to 0, *i.e.* the matrix D is regular and the following inclusion:

$$0 \in Cx + D\lambda + a + N_K(\lambda) \tag{5.20}$$

possesses a unique solution for all x. There is neither notion of consistent initial conditions in this case, since we assume that

$$0 \in Cx_0 + D\lambda(t_0) + a(t_0) + N_K(\lambda(t_0)) \tag{5.21}$$

has a solution, nor hard constraints on the state vector x.

The following time-stepping scheme is used for $(\mathsf{P})_{\mathsf{LTI}}$ when a solution of class C^1 is expected

$$\begin{cases} x_{k+1} - x_k = h(Ax_{k+\theta} + u_{k+\theta} + r_{k+\gamma}), \\ y_{k+1} = Cx_{k+1} + D\lambda_{k+1} + a_{k+1}, \\ r_{k+1} = B\lambda_{k+1}, \\ 0 \in y_{k+1} + N_K(\lambda_{k+1}), \end{cases} \tag{5.22}$$

with $\theta \in [0, 1]$ and $\gamma \in [0, 1]$. The initial value of $\lambda_0 = \lambda(t_0)$ is given by the solution of (5.21).

The discretized system (5.22) amounts to solving at each time-step the following OSNSP:

$$\begin{cases} y_{k+1} = M\lambda_{k+1} + q, \\ 0 \in y_{k+1} + N_K(\lambda_{k+1}), \end{cases} \tag{5.23}$$

with

$$M = D + h\gamma C(I - h\theta A)^{-1} B, \tag{5.24}$$

and

$$q = a_{k+1} + C(I - h\theta A)^{-1} \big[(I + h(1-\theta)A)x_k + hu_{k+\theta} + h(1-\gamma)B\lambda_k\big]. \tag{5.25}$$

For the nonlinear dynamics of the problem (P) in (5.19), the following scheme is implemented:

$$\begin{cases} x_{k+1} - x_k = h(f(x_{k+\theta}, t_{k+\theta}) + u_{k+\theta} + r_{k+\gamma}), \\ y_{k+1} = g(x_{k+1}, \lambda_{k+1}, t_{k+1}), \\ r_{k+1} = h(x_{k+1}, \lambda_{k+1}, \lambda_{k+1}), \\ 0 \in y_{k+1} + N_K(\lambda_{k+1}). \end{cases} \tag{5.26}$$

A OSNSP equivalent to (5.23) cannot be explicitly written but the problem still appears as a MCP. We will see how one can perform a Newton linearization of (5.26) to retrieve (5.23), in Sect. 5.2.6.

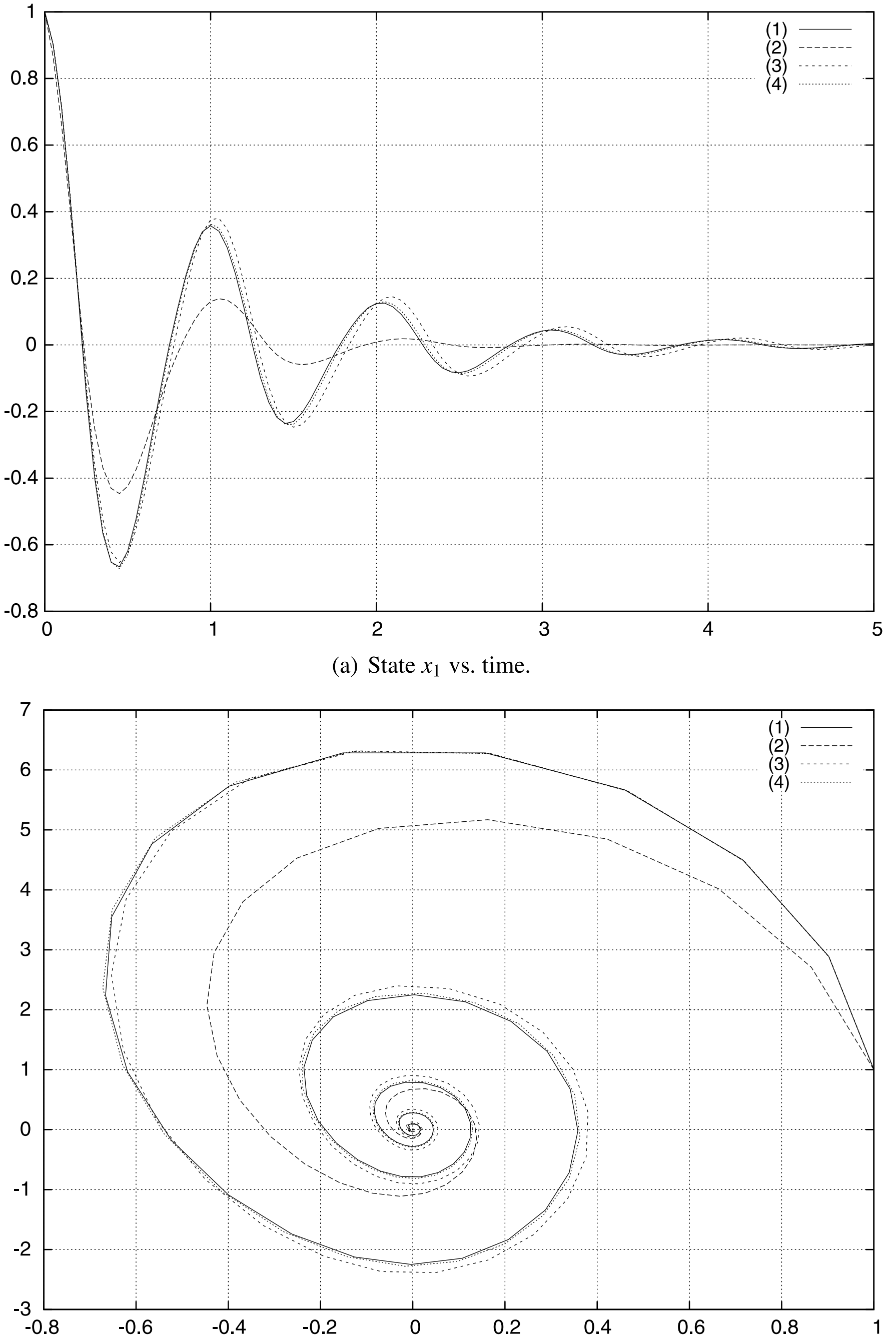
Example 5.1 Let us illustrate the influence of the parameters θ and γ on the behavior of the numerical scheme (5.22). Let us consider the circuit (b) described in Fig. 1.14. The dynamical system which models the circuit is given by (1.39) and we identify $D = [R]$. The solution $x(t)$ of the problem is of class C^1.

In Fig. 5.1, simulation results are given with the following data: $h = 5 \times 10^{-2}$, $x_0 = [1\,1]^T$, $R = 10$, $L = 1$ and $C = 1/(2\pi)^2$. Three numerical simulations are compared to the exact solution. With $\theta = 1$ and $\gamma = 1$, we retrieve a fully implicit Euler scheme. The main discrepancy when $\theta = 1/2$ and $\gamma = 1/2$ is the numerical damping, which is drastically attenuated. If the numerical damping does not hamper the convergence of the scheme, the quality of the solution for a finite time-step is modified.

Remark 5.2 Note that the order of consistency of the scheme is not necessarily improved when $\theta = 1/2$ and $\gamma = 1/2$ as we can expect for a smooth solution. Indeed, the mid-point rule is known to be of order 2 for solutions of class C^2. In the case of a solution of class C^1, the order is not necessarily achieved. Higher order time-stepping schemes, for instance Runge–Kutta schemes or BDF methods cannot achieve a higher order of consistency as it has been shown in Acary and Brogliato (2008, Chap. 9). Their usefulness for such a nonsmooth modeling is therefore doubtful. Their interest can lie in improving the efficiency over smooth phases of evolution.

5.2.2 Time-Stepping Schemes for an Absolutely Continuous Solution

In this case, the problem $(\mathsf{P})_{\mathsf{LTI}}$ is expected to have a unique trajectory $x(t)$ which is only absolutely continuous. Its time-derivative $\dot{x}(t)$ and the variable $\lambda(t)$ are as-



(a) State x_1 vs. time.

(b) Phase portrait. x_1 vs. x_2.

Fig. 5.1 Solution of (1.39) with the time-stepping scheme (5.22). (*1*) Exact solution $x(t_k)$. (*2*) x_k with $\theta = 1$, $\gamma = 1$. (*3*) x_k with $\theta = 1/2$, $\gamma = 1$. (*4*) x_k with $\theta = 1/2$, $\gamma = 1/2$

sumed to be right-continuous functions of bounded variations. Since λ may have some discontinuities, some care has to be taken for its discretization and the θ–method as in (5.22) can no longer be used. For $K = \mathbb{R}^m_+$, this situation is encountered when the relative degree of the system is equal to 1, *i.e.*, when the matrix D is rank deficient or only positive semi-definite.

The question of consistent initial conditions is also raised since the inclusion (5.20) with a rank deficient matrix D imposes some constraints on the state vector x. For instance, with $D = 0$, the initial conditions must satisfy

$$Cx_0 + a(t_0) \in K^*. \tag{5.27}$$

If the condition (5.27) does not hold, the trajectory $x(t)$ has to jump at the initial time, and the solution is no longer continuous. This case will therefore be treated in Sect. 5.2.3 when a solution of bounded variation is expected.

The following time-stepping scheme is used for $(\mathsf{P})_{\mathsf{LTI}}$ when a absolutely continuous solution is expected:

$$\begin{cases} x_{k+1} - x_k = h(Ax_{k+\theta} + u_{k+\theta} + r_{k+1}), \\ y_{k+1} = Cx_{k+1} + D\lambda_{k+1} + a_{k+1}, \\ r_{k+1} = B\lambda_{k+1}, \\ 0 \in y_{k+1} + N_K(\lambda_{k+1}), \end{cases} \tag{5.28}$$

with $\theta \in [0, 1]$.

In the time-discretization (5.28), the θ-method is not applied to variables like r and λ which are supposed to be of bounded variations. The discretized system (5.28) amounts to solving at each time-step the following OSNSP:

$$\begin{cases} y_{k+1} = M\lambda_{k+1} + q, \\ 0 \in y_{k+1} + N_K(\lambda_{k+1}), \end{cases} \tag{5.29}$$

with

$$M = D + hC(I - h\theta A)^{-1}B, \tag{5.30}$$

and

$$q = a_{k+1} + C(I - h\theta A)^{-1}\left[(I + h(1-\theta)A)x_k + hu_{k+\theta}\right]. \tag{5.31}$$

For the nonlinear dynamics of the problem (P) in (5.19), the following scheme is implemented:

$$\begin{cases} x_{k+1} - x_k = h\left(f(x_{k+\theta}, t_{k+\theta}) + u_{k+\theta} + r_{k+1}\right), \\ y_{k+1} = g(x_{k+1}, \lambda_{k+1}, t_{k+1}), \\ r_{k+1} = h(x_{k+1}, \lambda_{k+1}, \lambda_{k+1}), \\ 0 \in y_{k+1} + N_K(\lambda_{k+1}). \end{cases} \tag{5.32}$$

Example 5.3 Let us illustrate the behavior of the numerical scheme (5.22) on the simple example given by a RLCZD circuit. Let us consider the dynamics of the circuit in Fig. 1.15(c), 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:

$$\begin{cases} \dot{x}_1(t) = x_2(t), \\ \dot{x}_2(t) + \frac{R}{L}x_2(t) + \frac{1}{LC}x_1(t) = \frac{1}{L}v(t), \end{cases} \tag{5.33}$$

where $v(\cdot)$ is the voltage of the Zener diode. We saw that $v(t) \in \partial\mathscr{F}_z(-i(t))$ in (1.11), thus we get

$$\begin{cases} \dot{x}_1(t) - x_2(t) = 0, \\ \dot{x}_2(t) + \frac{R}{L}x_2(t) + \frac{1}{LC}x_1(t) \in \frac{1}{L}\partial\mathscr{F}_z(-x_2(t)), \end{cases} \tag{5.34}$$

which is a differential inclusion. The solution is sought as an absolutely continuous solutions. The variable $\lambda(t)$ is a function of bounded variations that can encounter jumps. The system can be written in the form (5.18) as

$$\begin{cases} \dot{x}(t) = Ax(t) + B\lambda(t), \\ y(t) = Cx(t) + D\lambda(t) + a, \\ 0 \in y(t) + N_{\mathbb{R}^2_+}(\lambda(t)), \end{cases} \tag{5.35}$$

with

$$\begin{gathered} A = \begin{bmatrix} 0 & 1 \\ -\frac{1}{LC} & -\frac{R}{L} \end{bmatrix}, \qquad B = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \qquad C = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \\ D = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \qquad a = \begin{bmatrix} 0 \\ V_z \end{bmatrix}. \end{gathered} \tag{5.36}$$

Note that the matrix D is a skew–symmetric matrix, which has full rank and is positive semi-definite.

In Fig. 5.2, a numerical simulation is reported with the initial conditions $x_1(0) = 1$, $x_2(0) = 1$ and $R = 0.1, L = 1, C = \frac{1}{(2\pi)^2}$, $V_z = 5$. The time step is $h = 5 \times 10^{-3}$. The effect of the choice of θ is mainly the decrease of the numerical damping of the scheme when $\theta = 1/2$.

One of the questions that can be raised is the use of the scheme (5.22) on the system (5.35) whose solutions are not of class C^1. In Fig. 5.3, we compare the solution obtained with the scheme (5.22) for two values of γ. For $\gamma = 1$, the scheme (5.22) is equivalent to (5.28). For $\gamma = 1/2$, we note that the state is almost approximated in the same way, but the variable $v(t)$ is subjected to instabilities when it reaches the multivalued part of the characteristics. This is typical of the behavior of higher order estimations of functions of bounded variations. Note that nothing is said on the convergence of the scheme (5.22) in the state variable. Only the qualitative behavior of the scheme is commented.

Other instabilities will be shown in Sect. 5.2.4.

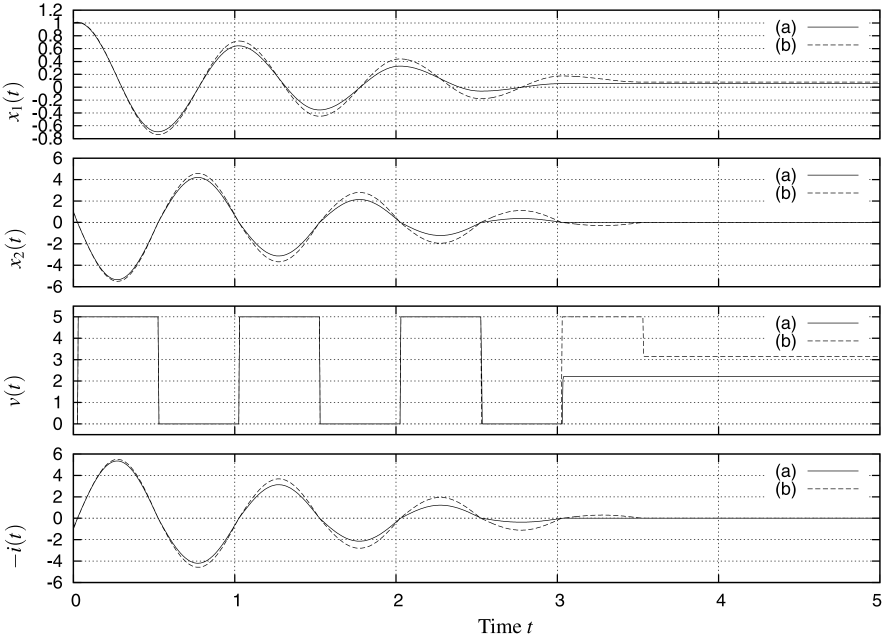


Fig. 5.2 Simulation of (5.35) with the time-stepping scheme (5.28). (**a**) $\theta = 1$ (**b**) $\theta = 1/2$

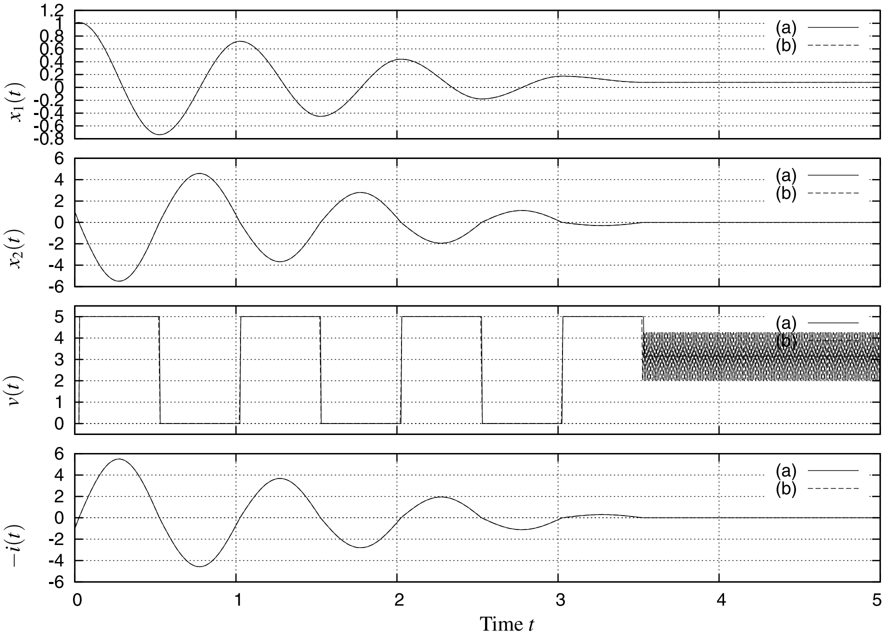


Fig. 5.3 Simulation of (5.35) with the time-stepping scheme (5.22). (**a**) $\theta = 1/2, \gamma = 1$ (**b**) $\theta = 1/2, \gamma = 1/2$

5.2.3 *Time-Stepping Solutions for a Solution of Bounded Variations*

If some jumps in the state are expected, the state $x(t)$ is usually assumed to be a right continuous function of bounded variations. This is particularly the case when the initial conditions are inconsistent or the external excitations $a(t)$ force the state to jump in order to satisfy the conditions (5.27).

As it has been illustrated in Sect. 1.1.5, the variable λ has to be replaced by a measure that can contain Dirac distributions. In the same vein, the time-derivative of the state $x(t)$ cannot be considered in the usual sense, but as a differential measure dx associated with a RCBV function. As in (2.126), the dynamics in the problem (5.18) is written in terms of a measure differential equation as

$$dx = Ax(t)dt + u(t)dt + Bdi, \tag{5.37}$$

where dx is the differential measure associated with the RCBV function $\dot{x}(t)$ and di is also a measure. The absolutely continuous function $\lambda(t)$ is the Radon-Nikodym derivative of di with respect to the Lebesgue measure, *i.e.*:

$$\frac{di}{dt} = \lambda(t). \tag{5.38}$$

If the singular part of the differential measure is neglected, a decomposition of the measure can be written as:

$$di = \lambda(t)dt + \sum_i \sigma_i \delta_{t_i} \tag{5.39}$$

where δ_{t_i} is the Dirac measure at time of discontinuities t_i and σ_i the amplitude. Thanks to this decomposition, the differential measure equation (5.37) can be written as a smooth dynamics:

$$\dot{x}(t) = Ax(t) + u(t) + B\lambda(t), \quad dt\text{—almost everywhere,} \tag{5.40}$$

and a jump dynamics at t_i:

$$x(t_i^+) - x(t_i^-) = B\sigma_i. \tag{5.41}$$

The time discretization of (5.37) has to take into account the nature of the solution to avoid point-wise evaluation of measures which is a nonsense as previously pointed out in Sect. 1.1.5. Only the measure of the time-intervals $(t_k, t_{k+1}]$ are considered such that:

$$dx((t_k, t_{k+1}]) = \int_{t_k}^{t_{k+1}} Ax(t) + u(t)\,dt + Bdi((t_k, t_{k+1}]). \tag{5.42}$$

By definition of the differential measure, we get

$$dx((t_k, t_{k+1}]) = x(t_{k+1}^+) - x(t_k^+). \tag{5.43}$$

The measure of the time-interval by di is kept as an unknown variable denoted by

$$\sigma_{k+1} = di((t_k, t_{k+1}]). \tag{5.44}$$

Finally, the remaining Lebesgue integral in (5.42) is approximated by an implicit Euler scheme

$$\int_{t_k}^{t_{k+1}} Ax(t) + u(t)\,dt \approx hAx_{k+1} + u_{k+1}. \tag{5.45}$$

As we said earlier in Remark 2.66, the matrix D in $(\mathsf{P})_{\mathsf{LTI}}$ needs to be at least rank-deficient to expect some jumps in the state. Let us start with the simplest case of $D = 0$.

The following time-stepping scheme is used for $(\mathsf{P})_{\mathsf{LTI}}$ when a solution of bounded variations is expected and $D = 0$

$$\begin{cases} x_{k+1} - x_k = h(Ax_{k+1} + u_{k+1}) + \sigma_{k+1}, \\ y_{k+1} = Cx_{k+1} + a_{k+1}, \\ r_{k+1} = B\sigma_{k+1}, \\ 0 \in y_{k+1} + N_K(\sigma_{k+1}). \end{cases} \tag{5.46}$$

The discretized system (5.46) amounts to solving at each time-step the following OSNSP

$$\begin{cases} y_{k+1} = M\sigma_{k+1} + q, \\ 0 \in y_{k+1} + N_K(\sigma_{k+1}), \end{cases} \tag{5.47}$$

with

$$M = C(I - hA)^{-1}B, \tag{5.48}$$

and

$$q = a_{k+1} + C(I - hA)^{-1}[x_k + hu_{k+1}]. \tag{5.49}$$

It is worth noting that the matrix M remains consistent when the time-step h vanishes if CB is assumed to be regular. This was not necessarily the case in (5.24).

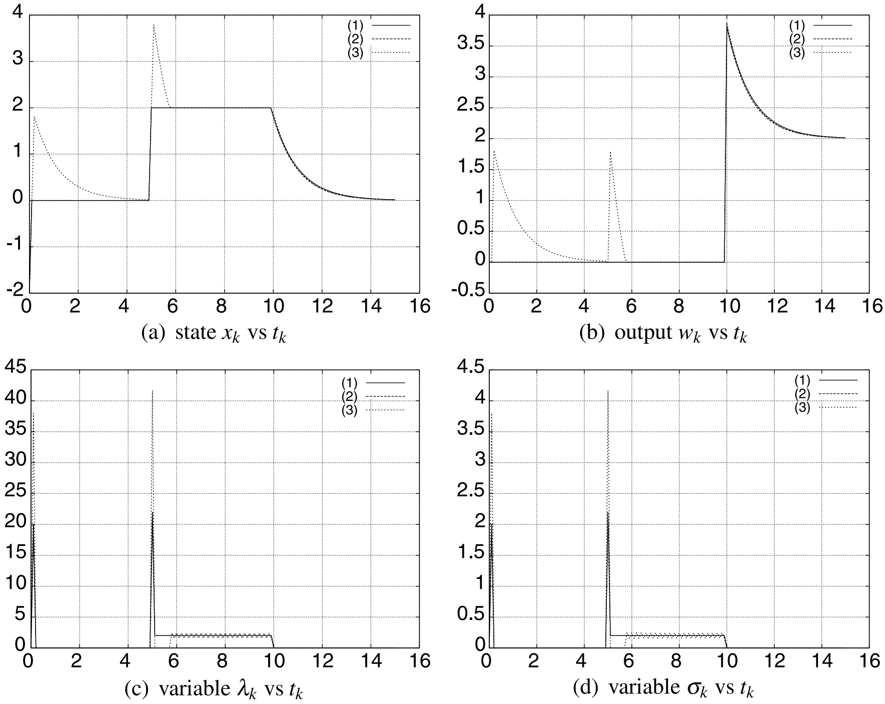
Remark 5.4 The following time-stepping scheme can also be used with an approximation of the integral term in (5.45) based on a θ-method:

$$\begin{cases} x_{k+1} - x_k = h(Ax_{k+\theta} + u_{k+\theta}) + r_{k+1}, \\ y_{k+1} = Cx_{k+1} + a_{k+1}, \\ r_{k+1} = B\sigma_{k+1}, \\ 0 \in y_{k+1} + N_K(\sigma_{k+1}). \end{cases} \tag{5.50}$$

The scheme remains consistent even if $\theta = 1/2$, but the order 2 is not retrieved if some jumps are encountered. The principal benefit of the latter scheme (5.50) is to reduce the numerical damping inherent to the backward Euler approximation.

Example 5.5 Let us consider the example in Sect. 1.1.5 of a simple RLD circuit whose dynamical equations are given by:

$$\begin{cases} L\dot{x}(t) + Rx(t) = \lambda(t), \\ 0 \leqslant y(t) = x(t) - i(t) \perp \lambda(t) \geqslant 0. \end{cases} \tag{5.51}$$



(a) state x_k vs t_k

(b) output w_k vs t_k

(c) variable λ_k vs t_k

(d) variable σ_k vs t_k

Fig. 5.4 Simulation of system (1.5). (*1*) scheme (5.46). (*2*) scheme (5.50) with $\theta = 1/2$. (*3*) scheme (5.22) with $\theta = 1/2$, $\gamma = 1/2$

As we have seen in Sect. 1.1.5, the solution jumps due to inconsistent initial conditions and due to the jump in the input $i(t)$. In Fig. 5.4, numerical simulations are performed with the same data as in Example 1.3 and by means of the schemes (5.46), (5.50) and (5.22). The values of $x_k, y_k, \lambda_k, \sigma_k$ are plotted with respect to time. When the value of λ_k (respectively σ_k) is not defined by the scheme, it is computed with σ_k/h (respectively $h\lambda_k$). The scheme (5.46) reproduces exactly the same results as the scheme in Sect. 1.1.5 and we note that the θ-scheme (5.50) can be used with $\theta = 1/2$ yielding to slightly less numerical damping. The scheme (5.22) with $\theta = 1/2$ and $\gamma = 1/2$ implies some instabilities on the values of λ_k and then σ_k. Furthermore, the trajectory is not correctly approximated and it does not to converge toward to the expected solution. This behavior forbids the use of such a scheme when a solution of bounded variations is expected.

For a general rank-deficient matrix D, the situation is more difficult. The relative degree r is non uniform and then some components in y and λ can be viewed as relative degree "0" variables and other as relative degree "1" variables. Without entering into deepest details, we will assume that the matrix D has the form

$$D = \begin{bmatrix} \tilde{D} & 0 \\ 0 & 0 \end{bmatrix}, \tag{5.52}$$

where $\tilde{D} \in \mathbb{R}^{d\times d}$ is regular, and that the set K can be written as

$$K = \tilde{K} \times \hat{K} \quad \text{with } \tilde{K} \subset \mathbb{R}^{d}. \tag{5.53}$$

The decomposition of the matrices C and B is done in a natural way as

$$C = \begin{bmatrix} \tilde{C} \\ \hat{C} \end{bmatrix} \quad \text{with } \tilde{C} \in \mathbb{R}^{d\times n} \quad \text{and} \quad B = [\, \tilde{B} \quad \hat{B} \,] \quad \text{with } \tilde{B} \in \mathbb{R}^{n\times d}. \tag{5.54}$$

The dynamics can be decomposed in

$$\begin{cases} dx = Ax(t)dt + u(t)dt + \tilde{B}\tilde{\lambda}(t)dt + \hat{B}\hat{di}, \\ \tilde{y}(t) = \tilde{C}x(t) + \tilde{a}(t) + \tilde{D}\tilde{\lambda}(t), \\ \hat{y} = \hat{C}x + \hat{a}(t), \\ 0 \in \tilde{y} + N_{\tilde{K}}(\tilde{\lambda}), \\ 0 \in \hat{y} + N_{\hat{K}}(\hat{di}). \end{cases} \tag{5.55}$$

The time-stepping scheme is a merge between the time-stepping scheme for an absolutely continuous solution (5.28) and the time-stepping scheme for a solution of bounded variations (5.46), yielding:

$$\begin{cases} x_{k+1} - x_k = h(Ax_{k+1} + u_{k+1}) + r_{k+1}, \\ \tilde{y}_{k+1} = \tilde{C}x_{k+1} + \tilde{a}_{k+1} + D\tilde{\lambda}_{k+1}, \\ \hat{y}_{k+1} = \hat{C}x_{k+1} + \hat{a}_{k+1}, \\ r_{k+1} = h\tilde{B}\lambda_{k+1} + \hat{B}\hat{\sigma}_{k+1}, \\ 0 \in \tilde{y}_{k+1} + N_K(\tilde{\lambda}_{k+1}), \\ 0 \in \hat{y}_{k+1} + N_K(\hat{\sigma}_{k+1}). \end{cases} \tag{5.56}$$

In the measure dynamics (5.55), a clear distinction is made between the components of y that are expected to jump ($\hat{y}$) and those ($\tilde{y}$) that are expected to be absolutely continuous. In the time-stepping scheme (5.56), their numerical counterparts $\hat{y}_{k+1}$ and $\tilde{y}$ are also treated in a distinct manner. The OSNSP that we have to solve is given by

$$\begin{cases} w_{k+1} = Mz_{k+1} + q, \\ 0 \in w_{k+1} + N_K(z_{k+1}), \end{cases} \tag{5.57}$$

with

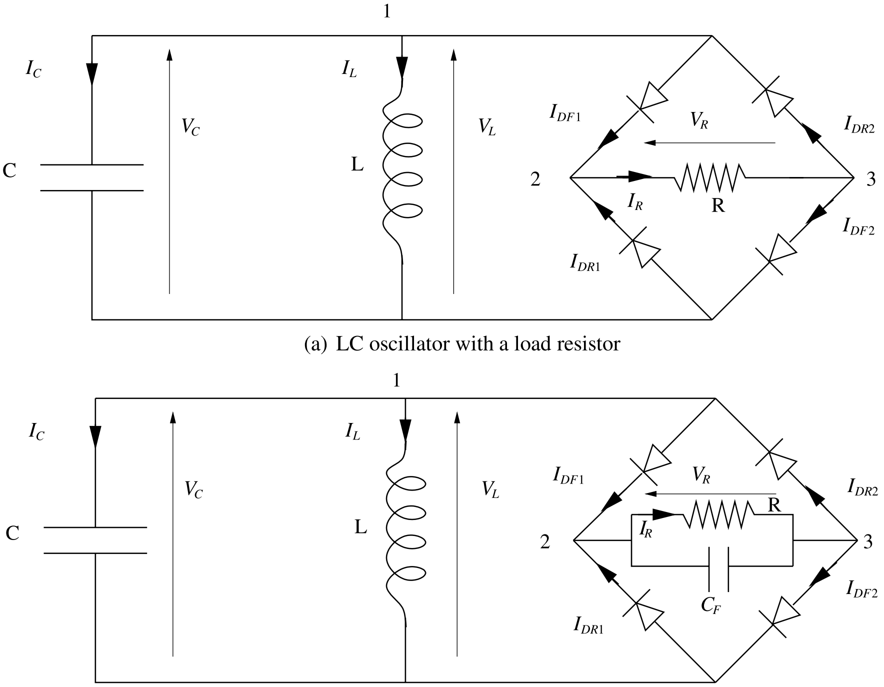
$$w_{k+1} = \begin{bmatrix} \tilde{y}_{k+1} \\ \hat{y}_{k+1} \end{bmatrix}, \qquad z_{k+1} = \begin{bmatrix} \tilde{\lambda}_{k+1} \\ \hat{\sigma}_{k+1} \end{bmatrix}, \tag{5.58}$$

$$M = \begin{bmatrix} \tilde{D} + h\tilde{C}(I - hA)^{-1}\tilde{B} & \tilde{C}(I - hA)^{-1}\hat{B} \\ h\hat{C}(I - hA)^{-1}\tilde{B} & \hat{C}(I - hA)^{-1}\hat{B} \end{bmatrix} \tag{5.59}$$

and

$$q = \begin{bmatrix} a_{k+1} + \tilde{C}(I - h\theta A)^{-1}(x_k + hu_{k+1}) \\ \hat{C}(I - h\theta A)^{-1}(x_k + hu_{k+1}) \end{bmatrix}. \tag{5.60}$$

As for (5.48), we note that the matrix M in (5.59) is regular when h vanishes if $\tilde{D}$ and CB are assumed to be regular.



(a) LC oscillator with a load resistor

(b) LC oscillator with a load resistor filtered by a capacitor

Fig. 5.5 Two configurations of the 4-diode bridge rectifier

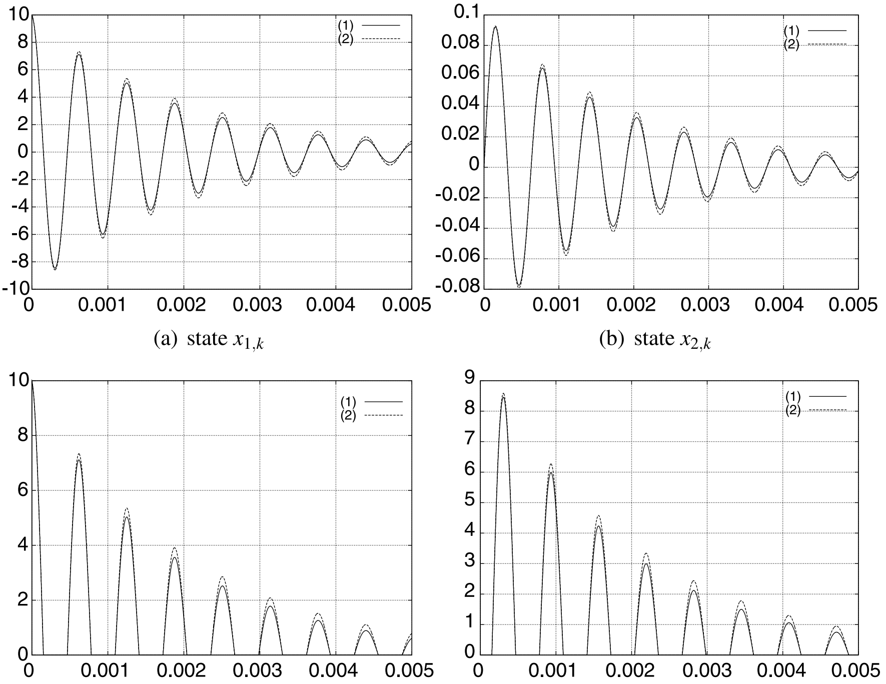
The assumptions (5.52) and (5.53) are quite strong assumptions on the structure of the dynamics which do not usually hold in the numerical practice. Nevertheless, a complete orthogonal decomposition of the matrix D can be invoked to identify a equivalent matrix D.

5.2.4 Illustrations of Wrong Discretizations

In this section, we continue to illustrate the differences between the various schemes presented in Sects. 5.2.1, 5.2.2 and 5.2.3.

Let us consider first two configurations of the 4-diode bridge illustrated in Fig. 5.5. The following values are taken for all configurations: $R = 1$ kΩ, $L = 10^{-2}$ H, $C = 1$ μF and $C_F = 300$ pF. The differences between the two configurations lie in the presence of a capacitor in the diode bridge. In Fig. 5.5(a), the resistor inside the bridges is supplied by a LC oscillator. The dynamical equations (5.18) are stated choosing:

$$x = \begin{bmatrix} V_L \\ I_L \end{bmatrix}, \quad \text{and} \quad y = \begin{bmatrix} I_{DR1} \\ I_{DF2} \\ V_2 - V_1 \\ V_1 - V_3 \end{bmatrix}, \qquad \lambda = \begin{bmatrix} V_2 \\ -V_3 \\ I_{DF1} \\ I_{DR2} \end{bmatrix}, \tag{5.61}$$



(a) state $x_{1,k}$

(b) state $x_{2,k}$

(c) variable $\lambda_{1,k}$

(d) variable $\lambda_{2,k}$

Fig. 5.6 Simulation of the configuration (Fig. 5.5(a)) with the scheme (5.22). Timestep $h = 10^{-6}$. (*1*) $\theta = 1, \gamma = 1$ (*2*) $\theta = 1/2, \gamma = 1/2$

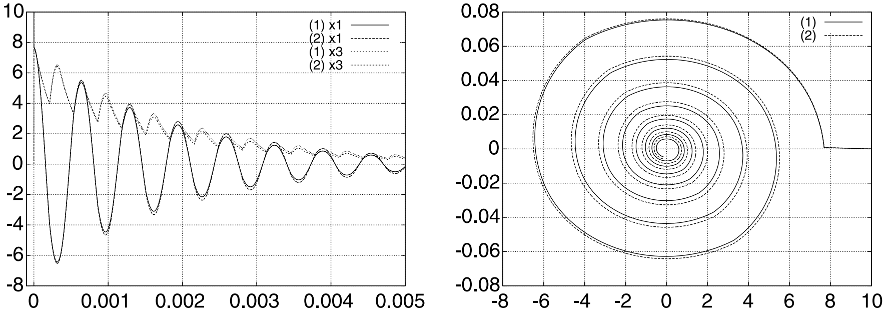
and with

$$A = \begin{bmatrix} 0 & -1/C \\ 1/L & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 & -1/C & 1/C & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \qquad u = 0,$$

$$C = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ -1 & 0 \\ 1 & 0 \end{bmatrix}, \qquad D = \begin{bmatrix} 1/R & 1/R & -1 & 0 \\ 1/R & 1/R & 0 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \qquad a = 0. \tag{5.62}$$

For this first configuration, the matrix D has full rank. The solution $x(t)$ is a solution of class C^1 and the scheme (5.22) can be used without any restrictions. The results are depicted in Fig. 5.6. As we can expect, the scheme approximates correctly the trajectory $x(t)$ and the variable $\lambda(t)$ and $y(t)$. As before, we note that the use of a midpoint rule decreases the numerical damping and improves the quality of the solution for a fixed time-step.

The other configuration depicted in Fig. 5.5(b) can be written as in (5.18) choosing:



(a) state $x_{1,k}$ and $x_{3,k}$

(b) phase portrait $x_{1,k}$ vs $x_{2,k}$

Fig. 5.7 Simulation of the configuration (Fig. 5.5(b)) with the scheme (5.28). Timestep $h = 10^{-6}$. (*1*) $\theta = 1$ (*2*) $\theta = 1/2$

$$x = \begin{bmatrix} V_L \\ I_L \\ V_R \end{bmatrix}, \qquad y = \begin{bmatrix} V_2 \\ I_{DF2} \\ V_2 - V_1 \\ V_L - V_3 \end{bmatrix}, \quad \text{and} \quad \lambda = \begin{bmatrix} I_{DR1} \\ -V_3 \\ I_{DF1} \\ I_{DR2} \end{bmatrix}, \tag{5.63}$$

and with

$$A = \begin{bmatrix} 0 & -1/C & 0 \\ 1/L & 0 & 0 \\ 0 & 0 & -1/(RC_F) \end{bmatrix},$$

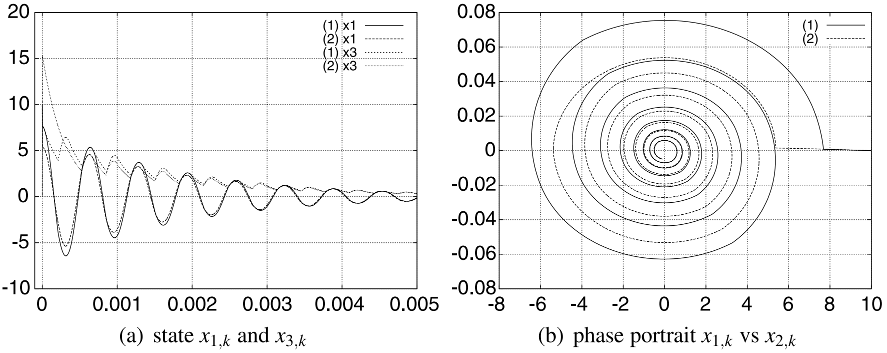
$$B = \begin{bmatrix} 0 & -1/C & 1/C & 0 \\ 0 & 0 & 0 & 0 \\ 1/C_F & 0 & 1/C_F & 0 \end{bmatrix}, \qquad u = 0, \tag{5.64}$$

$$C = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}, \qquad D = \begin{bmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 1 & -1 \\ 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \qquad a = 0.$$

For this second configuration, the matrix D does not have full rank ($\text{rank}(D) = 2$). The solution $x(t)$ is a solution of bounded variations and a jump can be encountered due to inconsistent initial conditions. Due to the fact that D is not in the simple form (5.52), it is difficult to know *a priori* the smoothness of the variables $y(t)$ and λ.

In Fig. 5.7, the state x_1 and x_3 are depicted. The simulation is performed with the scheme (5.28) and we note that the jump at the initial time is correctly approximated. In Fig. 5.8, the simulation is performed with the scheme (5.22). We note that the trajectory does not seem to be consistently approximated due to an initial error at the jump time. It is worth noting that none of the previous schemes are dedicated to simulate systems with solutions of bounded variations.

In Figs. 5.9 and 5.10, we give the details of the variables $\lambda(t)$ and $y(t)$ for the schemes (5.28) and (5.22) and various values of θ and γ. In Fig. 5.9, the



(a) state $x_{1,k}$ and $x_{3,k}$

(b) phase portrait $x_{1,k}$ vs $x_{2,k}$

Fig. 5.8 Simulation of the configuration (Fig. 5.5(b)) with the scheme (5.22). Timestep $h = 10^{-6}$. (*1*) $\theta = 1, \gamma = 1$ (*2*) $\theta = 1/2, \gamma = 1/2$

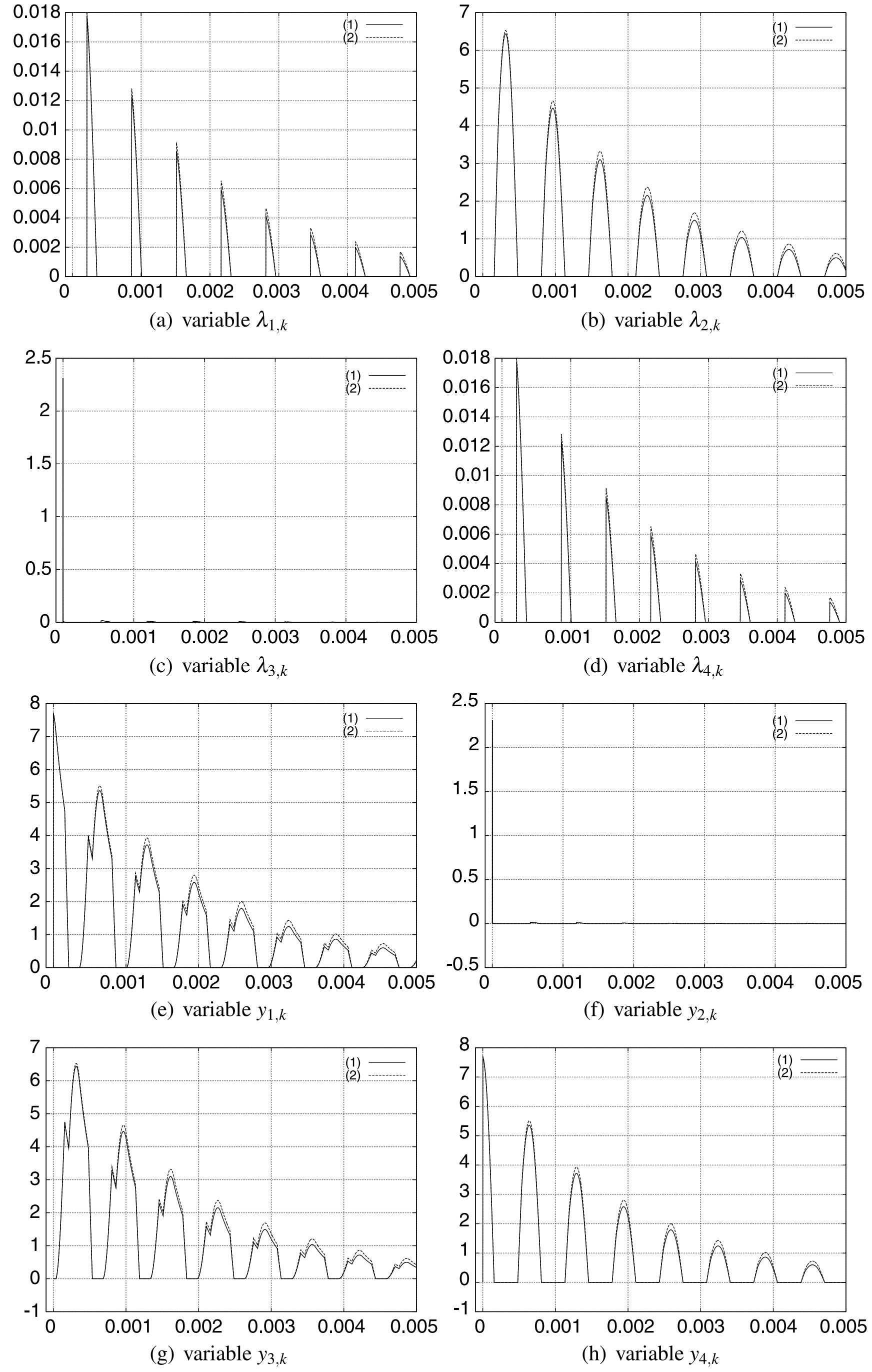
scheme (5.28) integrates the variables without instabilities. Note that the values $\lambda_{2,k}$ and $y_{3,k}$ approximate some absolutely continuous functions, while the values $\lambda_{1,k}$, $\lambda_{4,k}$ and $y_{1,k}$, y_4, k approximate some functions of bounded variations. For the values $\lambda_{3,k}$ and $y_{2,k}$, we notice in Figs. 5.9(c) and (f) that the scheme tries to evaluate at the initial point a measure. A correct estimation of these values would involve a variable σ_k in place of λ_k which is homogeneous to an impulse. Due to the fact that it is difficult to know *a priori* the nature of y and λ when D is rank-deficient, it is difficult to adapt correctly the scheme. In this situation, y should be also replaced by its associated measure and suitably integrated. In Fig. 5.10, we notice that the scheme (5.22) develops some instabilities on the values λ_k and y_k which approximate functions of bounded variations and is completely wrong in the approximation of measures.

We end this series of simulation by noting that the values of V_2 and V_3 are not everywhere uniquely defined. This fact is clearly related to the rank deficiency of the matrix D. To illustrate this behavior, we report in Fig. 5.11 the similar simulation as in Fig. 5.9 but with another OSNSP solver which is more sensible to the non-uniqueness of the solution. Some instabilities can be noticed which are the consequences of the non-uniqueness of V_2 and V_3 when the voltage V_L is negative. Indeed, the values of the node potentials belongs to a whole interval when the all the diode are in the OFF states.

5.2.5 How to Choose a Scheme in Practice?

From the above developments, one sees that the choice of a numerical scheme is not a trivial task when the mathematical nature of the solution is not known in advance. In practice, the following procedure is implemented:

1. The numerical simulation is first performed with the scheme (5.28), which is the most versatile one. By using two different time-steps, wrong evaluations of measures at jumps times can be detected.



(a) variable $\lambda_{1,k}$

(b) variable $\lambda_{2,k}$

(c) variable $\lambda_{3,k}$

(d) variable $\lambda_{4,k}$

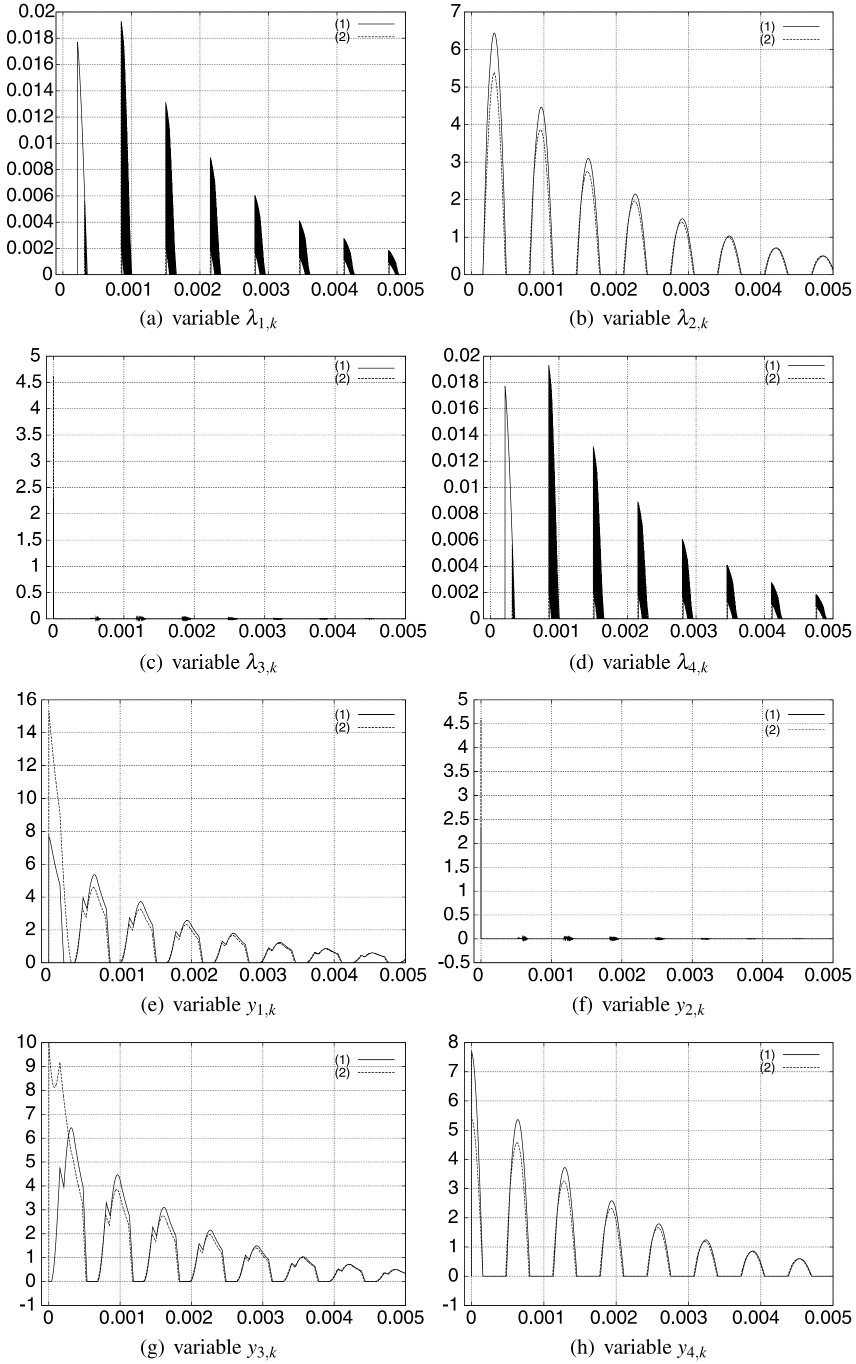
(e) variable $y_{1,k}$

(f) variable $y_{2,k}$

(g) variable $y_{3,k}$

(h) variable $y_{4,k}$

Fig. 5.9 Simulation of the configuration (Fig. 5.5(b)) with the scheme (5.28). Timestep $h = 10^{-6}$. *(1)* $\theta = 1$ *(2)* $\theta = 1/2$



(a) variable $\lambda_{1,k}$

(b) variable $\lambda_{2,k}$

(c) variable $\lambda_{3,k}$

(d) variable $\lambda_{4,k}$

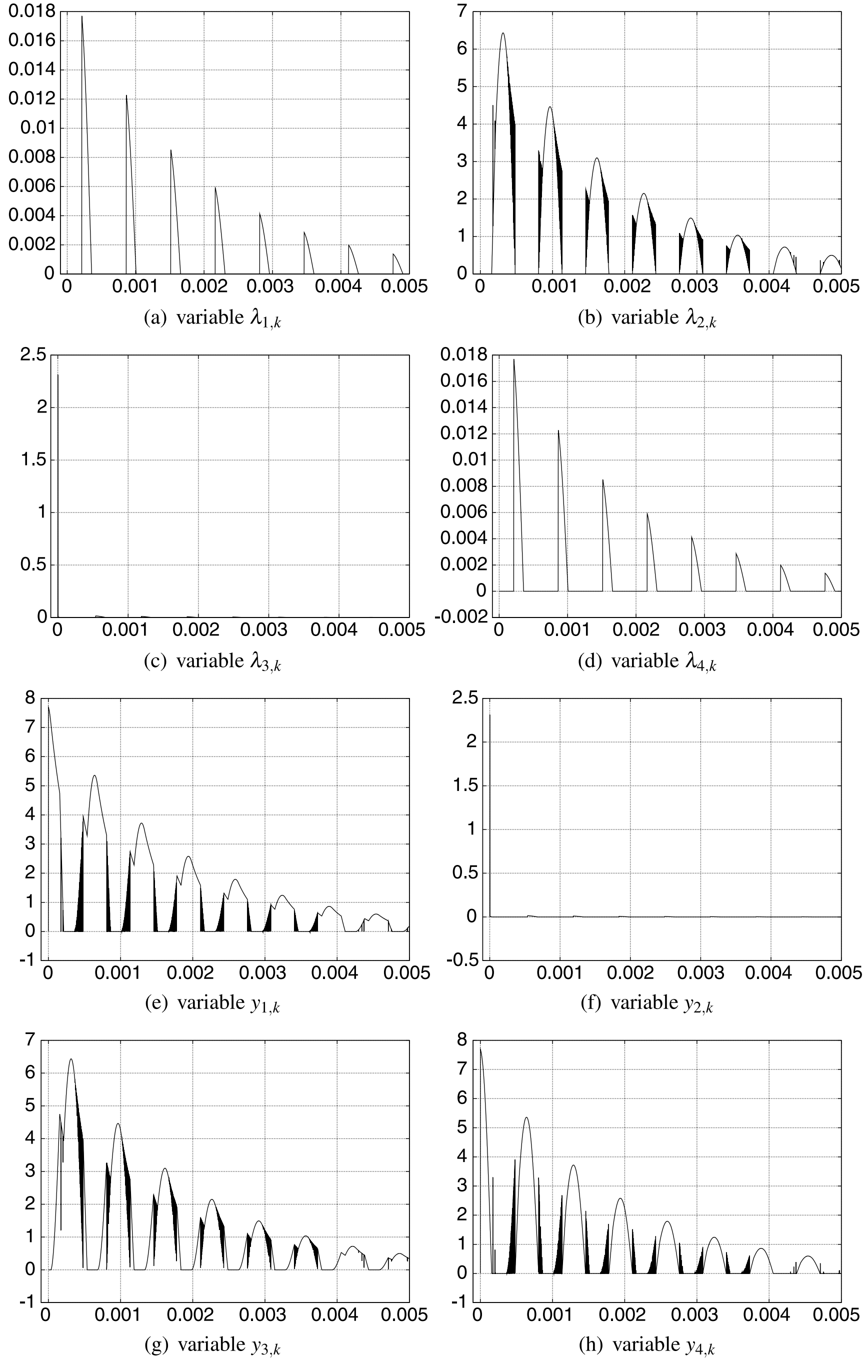
(e) variable $y_{1,k}$

(f) variable $y_{2,k}$

(g) variable $y_{3,k}$

(h) variable $y_{4,k}$

Fig. 5.10 Simulation of the configuration (Fig. 5.5(b)) with the scheme (5.22). Timestep $h = 10^{-6}$. *(1)* $\theta = 1, \gamma = 1$ *(2)* $\theta = 1/2, \gamma = 1/2$



(a) variable $\lambda_{1,k}$

(b) variable $\lambda_{2,k}$

(c) variable $\lambda_{3,k}$

(d) variable $\lambda_{4,k}$

(e) variable $y_{1,k}$

(f) variable $y_{2,k}$

(g) variable $y_{3,k}$

(h) variable $y_{4,k}$

Fig. 5.11 Simulation of the configuration (Fig. 5.5(b)) with the scheme (5.28). Timestep $h = 10^{-6}$. $\theta = 1$

2. If jumps are detected, the scheme (5.50) is used for a correct evaluation of the impulses.
3. If no jump is detected, the scheme (5.22) can be tried with $\gamma \neq 1$. If some instabilities are detected on the variables λ, we keep the value $\gamma = 1$.

5.2.6 Newton's Method for the Nonlinear Dynamics

In the nonlinear case, two choices are possible for solving the OSNSP. The first one is to use a dedicated numerical solver which is able to directly deal with the nonlinearities. The other possibility is to perform an external Newton Loop based on the linearization of the dynamics. We propose in this section to give an overview of the linearization of the scheme (5.26) that is recalled:

$$\begin{cases} x_{k+1} - x_k = h(f(x_{k+\theta}, t_{k+\theta}) + u_{k+\theta} + r_{k+\gamma}), \\ y_{k+1} = g(x_{k+1}, \lambda_{k+1}, t_{k+1}), \\ r_{k+1} = h(x_{k+1}, \lambda_{k+1}, \lambda_{k+1}), \\ 0 \in y_{k+1} + N_K(\lambda_{k+1}). \end{cases} \tag{5.65}$$

This procedure leads to a linearized problem equivalent to (5.22). The first line of the problem (5.65) can be written under the form of a residual term $\mathscr{R}$ depending only on x_{k+1} and r_{k+1} such that

$$\mathscr{R}(x_{k+1}, r_{k+1}) = 0, \tag{5.66}$$

with $\mathscr{R}(x, r) = x - x_k - h\theta f(\theta x + (1-\theta)x_k, t_{k+\theta}) - h\gamma r - h(1-\gamma)r_k$. The solution of this system of nonlinear equations is sought as a limit of the sequence $\{x^{\alpha}_{k+1}, r^{\alpha}_{k+1}\}_{\alpha \in \mathbb{N}}$ such that:

$$\begin{cases} x^0_{k+1} = x_k, \\ \mathscr{R}_L(x^{\alpha+1}_{k+1}, r^{\alpha+1}_{k+1}) = \mathscr{R}(x^{\alpha}_{k+1}, r^{\alpha}_{k+1}) + [\nabla^T_x \mathscr{R}(x^{\alpha}_{k+1}, r^{\alpha}_{k+1})](x^{\alpha+1}_{k+1} - x^{\alpha}_{k+1}) \\ \qquad\qquad + [\nabla^T_r \mathscr{R}(x^{\alpha}_{k+1}, r^{\alpha}_{k+1})](r^{\alpha+1}_{k+1} - r^{\alpha}_{k+1}) = 0. \end{cases} \tag{5.67}$$

In order to simplify the notation, we introduce the so-called "free" residual term as

$$\mathscr{R}_{\mathsf{free}}(x) = x - x_k - hf(\theta x + (1-\theta)x_k, t_{k+\theta}),$$

together with the following definitions:

$$\begin{aligned} &\mathscr{R}(x, r) = \mathscr{R}_{\mathsf{free}}(x) - h\gamma r - h(1-\gamma)r_k, \\ &\mathscr{R}^{\alpha}_{k+1} = \mathscr{R}(x^{\alpha}_{k+1}, r^{\alpha}_{k+1}) = \mathscr{R}_{\mathsf{free}}(x^{\alpha}_{k+1}, r^{\alpha}_{k+1}) - h\gamma r^{\alpha}_{k+1} - h(1-\gamma)r_k, \\ &\mathscr{R}^{\alpha}_{\mathsf{free},k+1} = \mathscr{R}_{\mathsf{free}}(x^{\alpha}_{k+1}, r^{\alpha}_{k+1}) = x^{\alpha}_{k+1} - x_k - hf(\theta x^{\alpha}_{k+1} + (1-\theta)x_k, t_{k+\theta}). \end{aligned}$$

The computation of the Jacobian of $\mathscr{R}$ with respect to x, denoted by W leads to

$$W^{\alpha}_{k+1} = \nabla^T_x \mathscr{R}(x^{\alpha}_{k+1}, r^{\alpha}_{k+1}) = I - h\theta \nabla^T_x f(x^{\alpha}_{k+1}, t_{k+1}). \tag{5.68}$$

At each time-step, we have to solve the following linearized problem:

$$\mathscr{R}^{\alpha}_{k+1} + W^{\alpha}_{k+1}(x^{\alpha+1}_{k+1} - x^{\alpha}_{k+1}) - h\gamma(r^{\alpha+1}_{k+1} - r^{\alpha}_{k+1}) = 0, \tag{5.69}$$

that is

$$h\gamma r_{k+1}^{\alpha+1} = r_c + W_{k+1}^{\alpha} x_{k+1}^{\alpha+1}, \tag{5.70}$$

with

$$\begin{aligned} r_c &= h\gamma r_{k+1}^{\alpha} - W_{k+1}^{\alpha} x_{k+1}^{\alpha} + \mathscr{R}_{k+1}^{\alpha} \\ &= -W_{k+1}^{\alpha} x_{k+1}^{\alpha} + \mathscr{R}_{\mathsf{free},k+1}^{\alpha} - h(1-\gamma) r_k. \end{aligned} \tag{5.71}$$

Note that the W is clearly non singular for small h. The same operation is performed with the second equation of (5.65) with the residual term

$$\mathscr{R}_y(x, y, \lambda) = y - g(x, \lambda, t_{k+1}) = 0, \tag{5.72}$$

leading to the following linearized equation:

$$y_{k+1}^{\alpha+1} = y_{k+1}^{\alpha} - \mathscr{R}_{y,k+1}^{\alpha} + C_{k+1}^{\alpha}(x_{k+1}^{\alpha+1} - x_{k+1}^{\alpha}) + D_{k+1}^{\alpha}(\lambda_{k+1}^{\alpha+1} - \lambda_{k+1}^{\alpha}), \tag{5.73}$$

with

$$\begin{aligned} C_{k+1}^{\alpha} &= \nabla_x^T g(t_{k+1}, x_{k+1}^{\alpha}, \lambda_{k+1}^{\alpha}), \\ D_{k+1}^{\alpha} &= \nabla_\lambda^T g(t_{k+1}, x_{k+1}^{\alpha}, \lambda_{k+1}^{\alpha}), \end{aligned} \tag{5.74}$$

and

$$\mathscr{R}_{y,k+1}^{\alpha} = y_{k+1}^{\alpha} - g(x_{k+1}^{\alpha}, \lambda_{k+1}^{\alpha}). \tag{5.75}$$

The same operation is performed with the third equation of (5.65) with the residual term

$$\mathscr{R}_r(r, x, \lambda) = r - g(x, \lambda, t_{k+1}) = 0. \tag{5.76}$$

In another notation, we obtain:

$$r_{k+1}^{\alpha+1} = r_1 + K_{k+1}^{\alpha} x_{k+1}^{\alpha+1} + B_{k+1}^{\alpha} \lambda_{k+1}^{\alpha+1}, \tag{5.77}$$

with

$$r_1 = h(x_{k+1}^{\alpha}, r_{k+1}^{\alpha}, t_{k+1}) - K_{k+1}^{\alpha} x_{k+1}^{\alpha} - B_{k+1}^{\alpha} \lambda_{k+1}^{\alpha} \tag{5.78}$$

and

$$\begin{aligned} K_{k+1}^{\alpha} &= \nabla_x^T h(x_{k+1}^{\alpha}, r_{k+1}^{\alpha}, t_{k+1}), \\ B_{k+1}^{\alpha} &= \nabla_\lambda^T h(x_{k+1}^{\alpha}, r_{k+1}^{\alpha}, t_{k+1}), \end{aligned} \tag{5.79}$$

and the residual term:

$$\mathscr{R}_{r,k+1}^{\alpha} = r_{k+1}^{\alpha} - h(x_{k+1}^{\alpha}, r_{k+1}^{\alpha}, t_{k+1}). \tag{5.80}$$

Inserting (5.77) into (5.69), we get the following linear relation between $x_{k+1}^{\alpha+1}$ and $\lambda_{k+1}^{\alpha+1}$:

$$(I - h\gamma(W_{k+1}^{\alpha})^{-1} K_{k+1}^{\alpha}) x_{k+1}^{\alpha+1} = x_p + h\gamma(W_{k+1}^{\alpha})^{-1} B_{k+1}^{\alpha} \lambda_{k+1}^{\alpha+1}, \tag{5.81}$$

with

$$\begin{aligned} x_p &= x_{\mathsf{free}} + h(W_{k+1}^{\alpha})^{-1}(g(x_{k+1}^{\alpha}, \lambda_{k+1}^{\alpha}, t_{k+1}) - B_{k+1}^{\alpha} \lambda_{k+1}^{\alpha} - K_{k+1}^{\alpha} x_{k+1}^{\alpha}), \\ x_{\mathsf{free}} &= -W_{k+1}^{\alpha,-1} R_{\mathsf{free},k+1}^{\alpha} + x_{k+1}^{\alpha}. \end{aligned} \tag{5.82}$$

Defining

$$\tilde{K}^{\alpha}_{k+1} = \big(I - h\gamma (W^{\alpha}_{k+1})^{-1} K^{\alpha}_{k+1}\big) \tag{5.83}$$

and inserting (5.81) into (5.73), we get the following linear relation between $y^{\alpha+1}_{k+1}$ and $\lambda^{\alpha+1}_{k+1}$:

$$y^{\alpha+1}_{k+1} = y_p + \Big[h\gamma C^{\alpha}_{k+1}(\tilde{K}^{\alpha}_{k+1})^{-1}(W^{\alpha}_{k+1})^{-1}B^{\alpha}_{k+1} + D^{\alpha}_{k+1}\Big]\lambda^{\alpha+1}_{k+1}, \tag{5.84}$$

with

$$y_p = y^{\alpha}_{k+1} - \mathscr{R}^{\alpha}_{yk+1} + C^{\alpha}_{k+1}((\tilde{K}^{\alpha}_{k+1})^{-1}x_p - x^{\alpha}_{k+1}) - D^{\alpha}_{k+1}\lambda^{\alpha}_{k+1}. \tag{5.85}$$

To summarize, the OSNSP we have to solve in each Newton iteration is:

$$\begin{cases} y^{\alpha+1}_{k+1} = M^{\alpha}_{k+1}\lambda^{\alpha+1}_{k+1} + q^{\alpha}_{k+1}, \\ -y^{\alpha+1}_{k+1} \in N_K(\lambda^{\alpha+1}_{k+1}), \end{cases} \tag{5.86}$$

with $M_{k+1} \in \mathbb{R}^{m\times m}$ and $q \in \mathbb{R}^m$ defined by:

$$\begin{cases} M^{\alpha}_{k+1} = h\gamma C^{\alpha}_{k+1}(\tilde{K}^{\alpha}_{k+1})^{-1}(W^{\alpha}_{k+1})^{-1}B^{\alpha}_{k+1} + D^{\alpha}_{k+1}, \\ q^{\alpha}_{k+1} = y_p. \end{cases} \tag{5.87}$$

The problem (5.86) is equivalent to the linear OSNSP that we obtained in (5.23). For the other schemes, the same procedure can be performed.

5.3 Time-Discretization of the General Cases

Let us go into details of the discretization of the general cases obtained by the adapted version of the MNA including nonsmooth electrical elements. For the most general form (DGE) in (5.1), a proposed scheme could be:

$$\begin{cases} M(X_{k+1}, t_{k+1})(X_{k+1} - X_k) = h(D(X_{k+1}, t_{k+1}) + U(t_{k+1}) + R_{k+1}), \\ y_{k+1} = G(X_{k+1}, \lambda_{k+1}, t_{k+1}), \\ R_{k+1} = H(X_{k+1}, \lambda_{k+1}, t_{k+1}), \\ 0 \in F(y_{k+1}, \lambda_{k+1}, t_{k+1}) + T(y_{k+1}, \lambda_{k+1}, t_{k+1}). \end{cases} \tag{5.88}$$

This scheme respects the first principle in the beginning of Sect. 5.2, *i.e.* a fully implicit integration of the generalized equation. In order to improve this basic scheme, further knowledge is needed. For instance, if the variable X is supposed to be absolutely continuous, a θ-method can be used. Unfortunately, at that time, it is difficult from the structure of (5.2)–(5.5) to guess *a priori* the smoothness of the solution. This is still an open issue. In the same vein, if the system encounters jumps, the dynamics and the inclusion in (5.1) have to be written in terms of measures and the numerical integration should be performed with impulses. Without any more mathematical results for the smoothness of solutions for the specific of (5.1), it is difficult to say more. In practice, the scheme is employed as a compromise between the midpoint rules developed in the previous sections and the scheme with impulses. After a first simulation, it is often possible to improve the numerical integration with a more dedicated scheme. The LTI case $(\mathsf{DGE})_{\mathsf{LTI}}$ in (5.7) is treated as well.

Concerning the semi-explicit systems (SEDGE) in (5.12), the proposed default scheme is

$$\begin{cases} x_{k+1} - x_k = hN^{-1}(x_{k+\theta}, t_{k+\theta})[f(x_{k+\theta}, z_{k+\theta}, t_{k+\theta})] + hr_{1,k+\gamma}, \\ 0 = g(x_{k+\theta}, z_{k+\theta}, t_{k+\theta}) + r_{2,k+\gamma}, \\ 0 = h_{\mathsf{NS}}(x_{k+1}, z_{k+1}, z_{\mathsf{NS},k+1}, \lambda_{k+1}, t_{k+1}), \\ y_{k+1} = g_{\mathsf{NS}}(x_{k+1}, z_{k+1}, z_{\mathsf{NS},k+1}, \lambda_{k+1}, t_{k+1}), \\ r_{k+1} = [r_{1,k+1}, r_{2,k+1}]^T = h(z_{\mathsf{NS},k+1}), \\ 0 \in F(y_{k+1}, \lambda_{k+1}, t_{k+1}) + T(y_{k+1}, \lambda_{k+1}, t_{k+1}), \end{cases} \tag{5.89}$$

with $\theta \in [0, 1]$ and $\gamma \in [0, 1]$. In this case, we assume that the trajectories $x(t)$ and $z(t)$ are at least functions of bounded variations. The default values for the parameters are $\theta = 1/2$ and $\gamma = 1$. If the solution is smooth enough, γ is chosen equal to $1/2$. In the case of jumps, the variable λ_{k+1} is substituted by the impulse σ_{k+1}.

5.4 One-Step NonSmooth Problems (OSNSP) Solvers

The difficult part in solving the various OSNSP that have been formulated in the above developments is the inclusion rule. Without any specific structure, the generalized equation (4.1) is hard to solve even if some fixed point and Newton methods can be designed. We will restrict ourselves in this section to the case of an inclusion into a normal cone to a convex set K:

$$-F(y, \lambda, t) \in N_K(\lambda). \tag{5.90}$$

For the numerical purposes, let us rewrite the problem (5.88) as a global inclusion

$$0 \in \mathsf{F}(\zeta) + N_C(\zeta), \tag{5.91}$$

where the unknown variable $\zeta = [X_{k+1}^T, y_{k+1}^T, \lambda_{k+1}^T]^T \in \mathbb{R}^{n+2m}$ and the function $\mathsf{F} : \mathbb{R}^{n+2n} \to \mathbb{R}^{n+2m}$ is defined by

$$\mathsf{F}(\zeta) = \begin{bmatrix} M(X_{k+1})(X_{k+1} - X_k) - h\left[D(X_{k+1}, t_{k+1}) + U(t_{k+1}) + H(X_{k+1}, \lambda_{k+1}, t_{k+1})\right] \\ G(X_{k+1}, \lambda_{k+1}, t_{k+1}) - y_{k+1} \\ F(y_{k+1}, \lambda_{k+1}, t_{k+1}) \end{bmatrix}. \tag{5.92}$$

The normal cone N_C is the normal cone to the following convex set

$$C = \mathbb{R}^n \times \mathbb{R}^m \times K \subset \mathbb{R}^{n+2m}. \tag{5.93}$$

We will see in the next section that the nonlinearity of $\mathsf{F}(.)$ can be directly treated by the numerical one-step solver. As it has been done in Sect. 5.2.6, another approach is to perform an outer Newton linearization of this problem by searching the solution as the limit for α of the following linearized problem

$$0 \in \nabla_\zeta^T \mathsf{F}(\zeta^\alpha)(\zeta^{\alpha+1} - \zeta^\alpha) + \mathsf{F}(\zeta^\alpha) + N_C(\zeta^{\alpha+1}), \tag{5.94}$$

for a given ζ^0. At each time-step k and at each Newton iteration α, the problem (5.94) appears to be affine in ζ.

The problem (5.91) is a VI written in the form of an inclusion into a normal cone to a convex set. The choice of the numerical solver depends mainly on the structure of the convex set K. Indeed, from a very general convex set K to a particular choice of K, the numerical solvers range from the numerical methods for VI to nonlinear equations, passing through various complementarity problem solvers. The convergence and the numerical efficiency are improved in proportion as the structure of K becomes simpler. In the sequel, major choice of K will be given leading to various classes of well-known problems in mathematical programming theory. We refer to Facchinei and Pang (2003) for a thorough presentation of the available numerical solvers and to Acary and Brogliato (2008, Chap. 12) for a comprehensive summary of the numerical algorithms.

5.4.1 K is a Finite Representable Convex Set

In practice, the convex set is finitely represented by

$$K = \{\lambda \in \mathbb{R}^m \mid h(\lambda) = 0, g(\lambda) \geqslant 0\}, \tag{5.95}$$

where the functions $h : \mathbb{R}^m \to \mathbb{R}^m$, $g : \mathbb{R}^m \to \mathbb{R}^m$ are assumed to be smooth with non vanishing Jacobians. More precisely, we assume that the following constraints qualification holds:

$$\forall \lambda \in K,\ \exists d \in \mathbb{R}^m, \quad \text{such that} \begin{cases} \nabla^T h_i(\lambda) d < 0, & i = 1 \dots m, \\ \nabla^T g_j(\lambda) d < 0, & j \in \mathscr{I}(\lambda), \end{cases} \tag{5.96}$$

where $\mathscr{I}(\lambda)$ is the set of active constraints at λ, that is

$$\mathscr{I}(\lambda) = \{j \in 1 \dots m, g_j(\lambda) = 0\}. \tag{5.97}$$

In this case, general algorithms for VI can be used. To cite a few, the minimization of the so-called regularized gap function (Fukushima 1992; Zhu and Marcotte 1993, 1994) or generalized Newton methods (Facchinei and Pang 2003, Chaps. 7 & 8) can be used. If $F(.)$ is affine (possibly after the linearization step described in (5.94)) and the functions $h(.)$ and $g(.)$ are also affine, the VI is said to be an affine VI for which the standard pivoting algorithms for LCP (Cottle et al. 1992) are extended in Cao and Ferris (1996).

5.4.2 K is a Generalized Box

Let us consider the case when K is a generalized box, *i.e.*:

$$K = \{\lambda \in \mathbb{R}^m \mid a_i \leqslant \lambda_i \leqslant b_i, a_i \in \overline{\mathbb{R}}, b_i \in \overline{\mathbb{R}}, i = 1 \dots m\}, \tag{5.98}$$

with $\overline{\mathbb{R}} = \{\mathbb{R} \cup \{+\infty, -\infty\}\}$. In this case, the problem (5.91–5.93) can be recast into an MCP by defining $p = n + m + m + m$ and the bounds l, u as $l = [0_n\ 0_m\ 0_m\ a]^T$ and $u = [0_n\ 0_m\ 0_m\ b]^T$.

The MCP can be solved by a large family of solvers based on Newton-type methods and interior-points techniques. In contrast to the interior-point methods, it is not difficult to find comparisons of numerical methods based on Newton's method for solving MCPs. We refer to Billups et al. (1997) for an impressive comparison of the major classes of algorithms for solving MCPs. If $\mathsf{F}(.)$ is affine, the MLCP is equivalent to a box-constrained affine VI. For this problem, the standard pivoting algorithm such as Lemke's method is extended in Sargent (1978). A special case of a generalized box is the positive orthant of $\mathbb{R}^m$, that is $K = \mathbb{R}^m_+$. Standard theory and most of the numerical algorithms for LCPs apply in this MCLP case.

When the circuit is simple and of low size in terms of the number of unknown variables, it is sometimes possible to write the DAE as an ODE and perform the explicit substitution of X by y and λ in the formulation (5.88). If the cone is also simply defined by a positive orthant, we arrive then at a standard LCP (Denoyelle and Acary 2006). Unfortunately, the LCP formulation is not amenable for more complicated cases where an automatic circuit equation formulation is used.

Part III
Numerical Simulations

This part presents extensive simulation results obtained with the INRIA SICONOS open-source platform (http://siconos.gforge.inria.fr/). The results show the efficiency of the NSDS approach. The examples consist first of some academic circuit cases, followed by the buck converter and the delta-sigma converter.

Chapter 6
The Automatic Circuit Equations Formulation (ACEF) Module and the Siconos Software

The Siconos Platform is a scientific computing software dedicated to the modeling, simulation, control, and analysis of nonsmooth dynamical systems (NSDS). It is developed in the Bipop team-project at INRIA[1] in Grenoble, France, and distributed under GPL GNU license.

Siconos aims at providing a general and common tool for nonsmooth problems in various scientific fields like applied mathematics, mechanics, robotics, electrical circuits, and so on. However, the platform is not supposed to re-implement the existing dedicated tools already used for the modeling of specific systems, but to integrate them.

The Automatic Circuit Equations Formulation (ACEF) module is the implementation of the automatic circuit equation extended to general nonsmooth components. From a SPICE netlist, possibly augmented by some nonsmooth components, the ACEF build a dynamical formulation that can be simulated by Siconos.

6.1 An Insight into Siconos

The present part is dedicated to a short presentation of the general writing process for a problem treated with Siconos, through a simple example. The point is to introduce the main functionalities, the main steps required to model and simulate the systems behavior, before going into more details in Sect. 6.2, where the NSDS will be described. The chosen example is a four-diode bridge wave rectifier as shown in Fig. 6.1.

An LC oscillator, initialized with a given voltage across the capacitor and a null current through the inductor, provides the energy to a load resistance through a full-wave rectifier consisting of a four ideal diodes bridge. Both waves of the oscillating voltage across the LC are provided to the resistor with current flowing always in the

[1] The French National Institute for Research in Computer Science and Control (http://bipop.inrialpes.fr).

V. Acary et al., *Nonsmooth Modeling and Simulation for Switched Circuits,*
Lecture Notes in Electrical Engineering 69,
DOI 10.1007/978-90-481-9681-4_6, © Springer Science+Business Media B.V. 2011

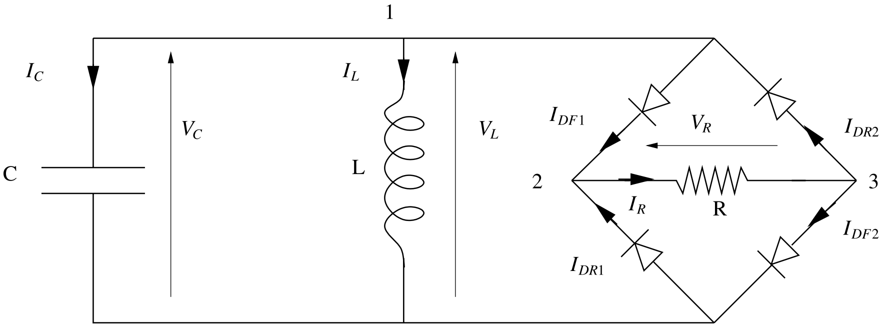


Fig. 6.1 A four-diode bridge wave rectifier

same direction. The energy is dissipated into the resistor and results in a damped oscillation.

One of the ways to define a problem with SICONOS consists in writing a C++ file. In the following, for the diode bridge example, only snippets of the C++ commands will be given, just to enlighten the main steps. It is noteworthy that one can also use an XML description or the Python interface.

6.1.1 Step 1. Building a Nonsmooth Dynamical System

In the present case, the oscillator is a time-invariant linear dynamical system, and using the Kirchhoff current and voltage laws and branch constitutive equations, its dynamics is written as (see Fig. 6.1 for the notation)

$$\begin{bmatrix} \dot{V}_L \\ \dot{I}_L \end{bmatrix} = \begin{bmatrix} 0 & -\frac{1}{C} \\ \frac{1}{L} & 0 \end{bmatrix} \cdot \begin{bmatrix} V_L \\ I_L \end{bmatrix} + \begin{bmatrix} 0 & 0 & -\frac{1}{C} & \frac{1}{C} \\ 0 & 0 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} -V_{\mathrm{DR1}} \\ -V_{\mathrm{DF2}} \\ I_{\mathrm{DF1}} \\ I_{\mathrm{DR2}} \end{bmatrix}. \tag{6.1}$$

If we denote

$$x = \begin{bmatrix} \dot{V}_L \\ \dot{I}_L \end{bmatrix}, \qquad \lambda = \begin{bmatrix} -V_{\mathrm{DR1}} \\ -V_{\mathrm{DF2}} \\ I_{\mathrm{DF1}} \\ I_{\mathrm{DR2}} \end{bmatrix}, \qquad A = \begin{bmatrix} 0 & \frac{-1}{C} \\ \frac{1}{L} & 0 \end{bmatrix},$$
$$r = \begin{bmatrix} 0 & 0 & -\frac{1}{C} & \frac{1}{C} \\ 0 & 0 & 0 & 0 \end{bmatrix} \lambda, \tag{6.2}$$

the dynamical system (6.1), (6.2) results in

$$\dot{x} = Ax + r. \tag{6.3}$$

The first step of any SICONOS problem is to define and build some `DynamicalSystemobjects` objects. The corresponding command lines to build a `FirstOrderLinearTIDS` object are:

```
// User-defined parameters
unsigned int ndof = 2;  // number of degrees of freedom of your system
double Lvalue = 1e-2;  // inductance
double Cvalue = 1e-6;  // capacitance
double Rvalue = 1e3;   // resistance
double Vinit = 10.0;   // initial voltage
// DynamicalSystem(s)
SimpleMatrix A(ndof,ndof);// All components of A are automatically  set to 0.
A(0,1) = -1.0/Cvalue;
A(1,0) = 1.0/Lvalue;
// initial conditions vector
SimpleVector x0(ndof);
x0(0) = Vinit;
// Build a First Order Linear and Time Invariant Dynamical System
//           using A matrix and x0 as initial state.
FirstOrderLinearTIDS * oscillator = new FirstOrderLinearTIDS(1,x0,A);
```

The suffix `DS` to the name of a class such as the `FirstOrderLinearTIDS` object means that this class inherits from the general class of `DynamicalSystem`.

Thereafter, it is necessary to define the way the previously defined dynamical systems will interact together. This is the role of the `Interaction` object composed of a `Relation` object, a set of algebraic equations, and of a `NonSmoothLaw` object.

The linear relations between the voltages and the currents inside the circuit are given by

$$\begin{bmatrix} I_{\mathrm{DR1}} \\ I_{\mathrm{DF2}} \\ -V_{\mathrm{DF1}} \\ -V_{\mathrm{DR2}} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ -1 & 0 \\ 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} V_L \\ I_L \end{bmatrix} + \begin{bmatrix} 1/R & 1/R & -1 & 0 \\ 1/R & 1/R & 0 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} -V_{\mathrm{DR1}} \\ -V_{\mathrm{DF2}} \\ I_{\mathrm{DF1}} \\ I_{\mathrm{DR2}} \end{bmatrix}, \tag{6.4}$$

which can be stated by the linear equation

$$y = Cx + D\lambda, \tag{6.5}$$

with

$$y = \begin{bmatrix} I_{\mathrm{DR1}} \\ I_{\mathrm{DF2}} \\ -V_{\mathrm{DF1}} \\ -V_{\mathrm{DR2}} \end{bmatrix}, \qquad D = \begin{bmatrix} 1/R & 1/R & -1 & 0 \\ 1/R & 1/R & 0 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \qquad \lambda = \begin{bmatrix} -V_{\mathrm{DR1}} \\ -V_{\mathrm{DF2}} \\ I_{\mathrm{DF1}} \\ I_{\mathrm{DR2}} \end{bmatrix}. \tag{6.6}$$

Completed with the relation between r and λ (see (6.2)) it results in a linear equation as

$$r = B\lambda. \tag{6.7}$$

This corresponds to a SICONOS `FirstOrderLinearTIR` object, *i.e.*, a linear and time-invariant coefficients relation. The corresponding code is as follows:

```
// -- Interaction --
// - Relations -
unsigned int ninter = 4;   // dimension of your Interaction
                           //  = size of y and lambda vectors
SimpleMatrix B(ndof,ninter);
B(0,2) =-1.0/Cvalue ;
B(0,3) = 1.0/Cvalue;
SimpleMatrix C(ninter,ndof);
C(2,0) = -1.0;
C(3,0) = 1.0;
// Build the Relation:
FirstOrderLinearTIR * myRelation = new FirstOrderLinearTIR(C,B);
// Add the D matrix to the relation.
SimpleMatrix D(ninter,ninter); D(0,0) = 1.0/Rvalue;
D(0,1) = 1.0/Rvalue; D(0,2) = -1.0; D(1,0) = 1.0/Rvalue;
D(1,1) = 1.0/Rvalue; D(1,3) = -1.0; (2,0) = 1.0; D(3,1)=1.0;
myRelation->setD(D);
```

To complete the `Interaction` object, a nonsmooth law is needed to define what the behavior will be when a nonsmooth event occurs.

Thus the behavior of each diode of the bridge, supposed to be ideal, can be described with a complementarity condition between the current and the reverse voltage (variables (y, λ)). Depending on the diode position in the bridge, y stands for the reverse voltage across the diode or for the diode current. Then, the complementarity conditions, as the results of the ideal diodes characteristics, are given by

$$\begin{array}{l} 0 \leqslant -V_{\mathrm{DR1}} \perp I_{\mathrm{DR1}} \geqslant 0 \\ 0 \leqslant -V_{\mathrm{DF2}} \perp I_{\mathrm{DF2}} \geqslant 0 \\ 0 \leqslant I_{\mathrm{DF1}} \perp -V_{\mathrm{DF1}} \geqslant 0 \\ 0 \leqslant I_{\mathrm{DR2}} \perp -V_{\mathrm{DR2}} \geqslant 0 \end{array} \quad \Longleftrightarrow \quad 0 \leqslant y \perp \lambda \geqslant 0, \tag{6.8}$$

which correspond to a `ComplementarityConditionNSL` object which is an inherited class form of the `NonSmoothLaw` class. The SICONOS code is as follows:

```
   // NonSmoothLaw definition
   unsigned int nslawSize = 4;
   NonSmoothLaw * myNslaw = new ComplementarityConditionNSL(nslawSize) ;
```

The `Interaction` is built using the concerned `DynamicalSystem`, the `Relation`, and the `NonSmoothLaw` defined above:

```
   // A name and a id-number for the Interaction
   string nameInter = "InterDiodeBridge";
   unsigned int numInter = 1;
   unsigned int ninter = 4; // ninter is the size of y
   Interaction* myInteraction = new Interaction(ninter, myNslaw, myRelation);
```

This is the end of the first step, you have now a `DynamicalSystem` and an `Interaction`. Before dealing with the Simulation, we first create the `Model` object that handles the `NonSmoothDynamicalSystem` and the `Simulation`.

```
// Model
double t0 = 0; // Initial time
double T = 10; // Total simulation time
Model * diodeBridge = new Model(t0,T);
diodebridge->nonSmoothDynamicalSystem()->insertDynamicalSystem(oscillator);
diodebridge->nonSmoothDynamicalSystem()->link(myInteraction,oscillator);
```

From this point, the diode bridge system is completely defined by the `NonSmoothDynamicalSystem` and handled by the `Model` object `DiodeBridge`. In the next section, a strategy of simulation will be defined and applied to this model.

6.1.2 Step 2. Simulation Strategy Definition

It is now necessary to define the way the dynamical behavior of the `NonSmoothDynamicalSystem` will be computed. This is the role of `Simulation` class. In SICONOS, two different strategies of simulation are available: the time-stepping schemes or the event-driven algorithms. To be complete, a `Simulation` object requires:

- a discretization of the considered time interval of study,
- a time-integration method for the dynamics,
- a way to formalize and solve the possibly nonsmooth problems.

For the diode bridge example, the Moreau's time-stepping scheme is used (Sect. 9.4), where the integration of the equations over the time steps is based on a θ-method. The nonsmooth problem is written as an LCP and solved with a projected Gauss–Seidel algorithm (Sect. 12.4.6). The resulting code in SICONOS is

```
double h =  1.0e-6;  // Time step
// The time discretisation, linked to the Model.
TimeDiscretisation * td = new TimeDiscretisation(h,diodeBridge);
// Moreau Integrator for the dynamics:
double theta = 0.5;
Moreau* myIntegrator = new Moreau(oscillator,theta);
// One Step nonsmooth problem
OneStepNSProblem* myLCP = new LCP("Lemke");
\\ max number of iteration
myLCP->numericsSolverOptions()->iparam[0]=101;
\\ tolerance
myLCP->numericsSolverOptions()->dparam[0]=1e-4;
// Build the Simulation Object
Simulation* s=new TimeStepping(DiodeBridge, myIntegrator, myLCP);
```

The last step is the simulation process with first the initialization and then the time-loop:

```
diodeBridge->initialize(s);
// Simulation process
s->run()
```

For a more detailed access to the simulation values inside a step, a time loop can be written explicitly:

```
  int k = 0; // Current step
  int N = td->getNSteps(); // Number of time steps
  for(k = 1 ; k < N ; ++k)
  {
    s->computeOneStep();
    s->nextStep();
  }
```

6.2 Siconos Software

6.2.1 General Principles of Modeling and Simulation

The Siconos software is mostly written in C++ and thus entirely relies on the object-oriented paradigm. In this first section we will not go into details on how to build these objects,[2] but rather on what they are and what they are used for.

As explained in Sect. 6.1, the central object is the `Model`. The model is the overall object composed of a nonsmooth dynamical system and a simulation object. The nonsmooth dynamical system object contains all the informations to describe the system and the simulation object contains all the informations to simulate it. The compulsory process to handle a problem with Siconos is first to build a nonsmooth dynamical system and then to describe a simulation strategy, see Sect. 6.2.1.2. Additionally, a control of the `Model` object can possibly be defined, see Sect. 6.2.1.3.

The way the software is written relies also on this “cutting-out” with clearly separated modeling and simulation components as explained in Sect. 6.2.4.

6.2.1.1 NSDS Modeling in Siconos Software

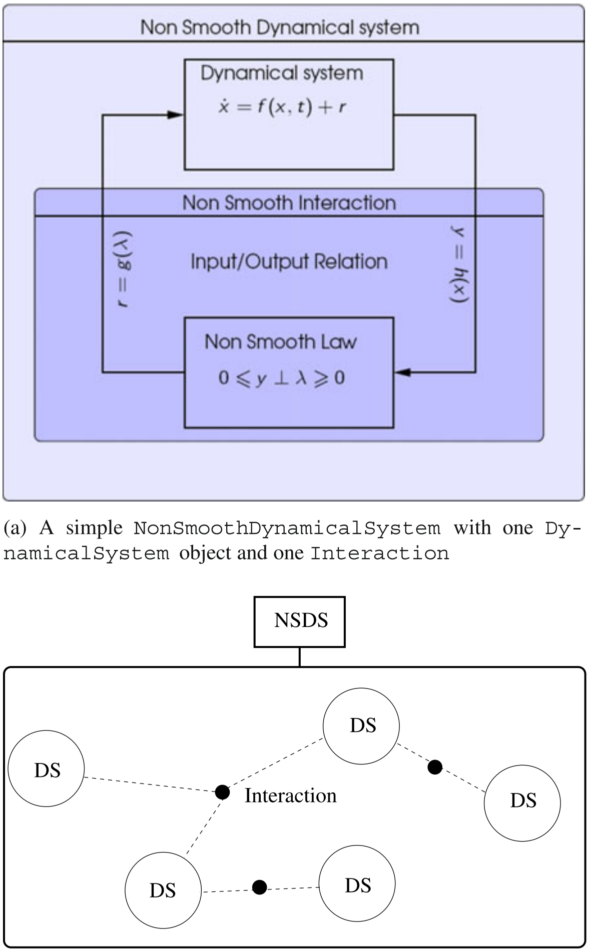
An NSDS can be viewed as a set of dynamical systems that may interact in a nonsmooth way through interactions. The modeling approach in the Siconos platform consists in considering the NSDS as a graph with dynamical systems as nodes and nonsmooth interactions as branches. Thus, to describe each element of this graph in Siconos, one needs to define a `NonSmoothDynamicalSystem` object composed of a set of `DynamicalSystem` objects and a set of `Interaction` objects.

A `DynamicalSystem` object is just a set of equations to describe the behavior of a single dynamical system, with some specific operators, initial conditions, and so on. A complete review of the dynamical systems available in Siconos is given in Sect. 6.2.2.1.

An `Interaction` object describes the way one or more dynamical systems are linked or may interact. For instance, if one considers a set of rigid bodies, the `Interaction` objects define and describe what happens at contact. The `Inter-`

[2] This is the role of the tutorial, users, guide or others manuals that may be found at http://siconos.gforge.inria.fr/.

Fig. 6.2 SICONOS nonsmooth dynamical system modeling principle



(a) A simple `NonSmoothDynamicalSystem` with one `DynamicalSystem` object and one `Interaction`

(b) The graph structure of a complex NSDS with `DynamicalSystem` objects as nodes and nonsmooth `Interaction` object as branches

`action` object is characterized by some "local" variables, y (also called output), and λ (input) and is composed of

- a `NonSmoothLaw` object that describes the mapping between y and λ,
- a `Relation` object that describes the equations between the local variables (y, λ) and the global ones (those of the `DynamicalSystem` object).

One can find a review of the various possibilities for the `Relation` and the `NonSmoothLaw` objects in Sects. 6.2.2.2 and 6.2.2.3. As summarized in Fig. 6.2, building a problem in SICONOS relies on the proper identification and construction of some `DynamicalSystems` and of all the potential interactions.

6.2.1.2 Simulation Strategies for the NSDS Behavior

Once an NSDS has been fully designed and described thanks to the objects detailed above, it is necessary to build a `Simulation` object, namely to define the way the nonsmooth response of the NSDS will be computed.

First of all, let us introduce the `Event` object, which is characterized by a type and a time of occurrence. Each event has also a `process` method which defines a list of actions that are executed when this event occurs. These actions depend on the object type. For the objects related to nonsmooth time events, namely `NonSmoothEvent`, an action is performed only if an event-driven strategy is chosen. For the `SensorsEvents` and `ActuatorEvent` related to control tools (see Sect. 6.2.1.3), an action is performed for both time-stepping and event-driven strategies at the times defined by the control law. Finally, thanks to a registration mechanism, user-defined events can be added.

To build the `Simulation` object, we first define a discretization, using a `TimeDiscretisation` object, to set the number of time steps and their respective size. Note that the initial and final time values are part of the `Model`. The time instants of this discretization define `TimeDiscretisationEvent` objects used to initialize an `EventsManager` object, which contains the list of `Event` objects and their related methods. The `EventsManager` object belongs to the simulation and will lead the simulation process: the system integration is always done between a "current" and a "next" event. Then, during the simulation, events of different types may be added or removed, for example when the user creates a sensor or when an impact is detected.

Thereafter, to complete the `Simulation` object, we need:

- some instructions on how to integrate the smooth dynamics over a time step, which is the role of the `OneStepIntegrator` objects,
- some details on how to formalize and solve the nonsmooth problems when they occur, this is done with the `OneStepNSProblem` objects.

To summarize, a `Simulation` object is composed of a `TimeDiscretisation`, a set of `OneStepIntegrator` plus a set of `OneStepNSProblem` and belongs to a `Model` object. The whole simulation process is led by the chosen type of strategy, either time-stepping or event-driven. To proceed, one needs to instantiate one of the classes that inherits from `Simulation` object: `TimeStepping` or `EventDriven`.

6.2.1.3 Control Tools

In SICONOS, some control can be applied on an NSDS. The principle is to get information from the systems thanks to some `Sensor` objects, used by some `Actuator` objects to act on the NSDS components. Each `Sensor` or `Actuator` object has its own `TimeDiscretisation` object, a list of time instants where data are to be captured for sensors or where action occurs for actuators. Those instants are scheduled as events into the simulation's `EventsManager` object and thus processed when necessary.

The whole control process is handled with a `ControlManager` object, which is composed of a set of `Sensor` objects and another set of `Actuator` objects. The `ControlManager` object "knows" the `Model` object and thus all its components.

Each `DynamicalSystem` object has a specific variable, named z, which is a vector of discrete parameters (see Sect. 6.2.2.1). To control the systems with a sampled control law, the `Actuator` object sets the values of z components according to the user instructions.

6.2.2 NSDS Related Components

In the following paragraphs, we turn our attention to the specific types of systems, relations, and laws available in the platform.

6.2.2.1 Dynamical Systems

The most general way to write dynamical systems in SICONOS is

$$g(\dot{x}, x, t, z) = 0,$$

which is an n-dimensional set of equations where

- t is the time,
- $x \in \mathbb{R}^n$ is the state,[3]
- the vector of algebraic variables $z \in \mathbb{R}^s$ is a set of discrete states, which evolves only at user-specified events. The vector z may be used to set some perturbation parameters or to stabilize the system with a sampled control law.

Under some specific conditions, we can rewrite this as

$$\dot{x} = \mathrm{rhs}(x, t, z),$$

where "rhs" means right-hand side. Note that in this case $\nabla_{\dot{x}} g(\cdot, \cdot, \cdot, \cdot)$ must be invertible. From this generic interface, some specific dynamical systems are derived, to fit with different application fields. They are separated into two categories: first- and second-order (Lagrangian) systems, and then specialized according to the type of their operators (linear or not, time invariant, *etc.*).

The following list reviews the dynamical systems implemented in SICONOS which inherit from the `DynamicalSystem` class:

[3] The typical dimension of the state vector can range between a few degrees of freedom and more than several hundred thousands, for example for mechanical or electrical systems. The implementation of the software has been done to deal either with small- or large-scale problems.

- `FirstOrderNonLinearDS` class, which describes the nonlinear dynamical systems of first order in the form
$$\begin{cases} M\dot{x}(t) = f(t, x(t), z) + r, \\ x(t_0) = x_0 \end{cases} \tag{6.9}$$
with M a $n \times n$ matrix, $f(x,t,z)$ the vector field, and r the input due to the nonsmooth behavior.
- `FirstOrderLinearDS` class, which describes the linear dynamical systems of first order in the form (coefficients may be time invariant or not)
$$\begin{cases} \dot{x}(t) = A(t,z)x(t) + b(t,z) + r, \\ x(t_0) = x_0. \end{cases} \tag{6.10}$$
Simple Electrical circuits for instance fit into this formalism, as shown in the diode bridge example in Sect. 6.1.
- `LagrangianDS` and also `LagrangianLinearTIDS` and class, which describes the Lagrangian nonlinear and linear dynamical systems are also implemented.

6.2.2.2 Relations

As explained above, some relations between local(y, λ), and global variables (x, r), have to be set to describe the interactions between systems. The general form of these algebraic equations is
$$\begin{cases} y = \text{output}(x, t, z, \ldots), \\ r = \text{input}(\lambda, t, z, \ldots), \end{cases} \tag{6.11}$$
and is contained in the abstract `Relation` class. Any other `Relation` objects are derived from this one.

As for `DynamicalSystems` they are separated in first- and second-order relations and specified according to the type and number of variables, the linearity of the operators, *etc*. The possible cases inherit from the `Relation` class as follows:

- `FirstOrderR` class, which describes the nonlinear relations of first order as
$$\begin{cases} y = h(X, t, Z), \\ R = g(\lambda, t, Z). \end{cases} \tag{6.12}$$
Note that we use upper case for all variables related to `DynamicalSystem` objects. Remember that a `Relation` object applies through the `Interaction` object to a set of dynamical systems, and thus, X, $Z, \ldots$ are concatenation of x, $z, \ldots$ of the `DynamicalSystem` objects involved in the relation.
- FirstOrderLinearTIR class, which describes the first-order linear and time-invariant relations:
$$\begin{cases} y = CX + FZ + D\lambda + e, \\ R = B\lambda. \end{cases} \tag{6.13}$$
Once again, see for instance the diode-bridge example in Sect. 6.1.

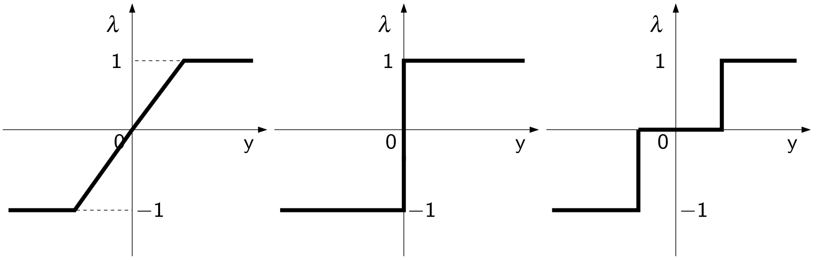


Fig. 6.3 Some multivalued piecewise-linear laws: saturation, relay, relay with dead zone

- `LagrangianScleronoumousR`, `LagrangianRheonomousR`, `LagrangianCompliantR` and `LagrangianLinearR` class are also implemented for mechanical applications.

6.2.2.3 Nonsmooth Laws

The `NonSmoothLaw` object is the last required object to complete the `Interaction` object. We present here a list of the existing laws in SICONOS which inherit from the generic `NonmSmoothLaw` class:

- `ComplementarityConditionNSL` class which models a complementarity condition as

$$0 \leqslant y \perp \lambda \geqslant 0. \tag{6.14}$$

- `RelayNSL` class which models the simple relay mapping as

$$\begin{cases} \dot{y} = 0 : |\lambda| \leqslant 1, \\ \dot{y} \neq 0 : \lambda = \text{sign}(y). \end{cases} \tag{6.15}$$

- `PiecewiseLinearNSL` class which models 1D piecewise-linear set-valued mapping with fill-in graphs as depicted in Fig. 6.3.

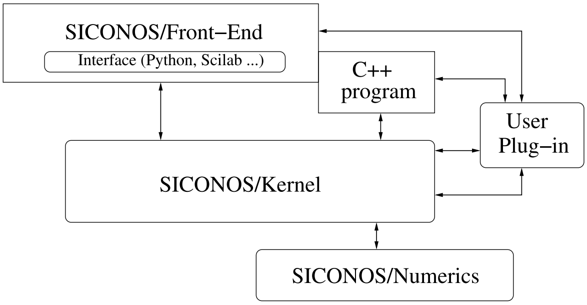
6.2.3 Simulation-Related Components

6.2.3.1 Integration of the Dynamics

To integrate the dynamics over a time step or between two events, `OneStepIntegrator` objects have to be defined. Two types of integrators are available at the time in the platform, listed below:

- `Moreau` class for Moreau's time-stepping scheme, based on a θ-method,
- `Lsodar` class for the event-driven strategy; this class is an interface for LSODAR, odepack integrator (see http://www.netlib.org/alliant/ode/doc).

Fig. 6.4 General design of SICONOS software



6.2.3.2 Formalization and Solving of the Nonsmooth Problems

Depending on the encountered situation, various formalizations for the nonsmooth problem are available:

- `LCP` class which describes the linear complementarity problem

$$\begin{cases} w = Mz + q, \\ 0 \leqslant w \perp z \geqslant 0, \end{cases}$$

- `FrictionContact2D(3D)` class, for two(three)-dimensional contact and friction problems,
- `QP` class for the quadratic programming problem,
- `Relay` class for the relay problem.

From a practical point of view, the solving of nonsmooth problems relies on low-level algorithms (from the SICONOS/Numerics package).

6.2.4 SICONOS *software design*

6.2.4.1 Overview

SICONOS is composed of three main parts: Numerics, Kernel and Front-End, as represented in Fig. 6.4.

The SICONOS/**Kernel** is the core of the software, providing high-level description of the studied systems and numerical solving strategies. It is fully written in C++, using extensively the STL utilities. A complete description of the Kernel is given in Acary and Brogliato (2008, Sect. 14.3.4.2).

The SICONOS/**Numerics** part holds all low-level algorithms, to compute basic well-identified problems (ordinary differential equations, LCP, QP, etc.).

The last component, SICONOS/**Front-End**, provides interfaces with some specific command-languages such as Python or SCILAB. This to supply more pleasant and easy-access tools for users, during pre/post-treatment. Front-End is only an optional pack, while the Kernel cannot work without Numerics.

6.2.4.2 The SICONOS/Numerics library

The SICONOS/Numerics library which is a stand-alone library, contains a collection of low-level numerical routines in C and F77 to solve linear algebra problems and OSNSP. It is based on well-known netlib libraries such as BLAS/LAPACK, ATLAS, Templates. Numerical integration of ODE is also provided thanks to ODEPACK (LSODE solver). At the present time, the following OSNSP solvers are implemented:

- LCP solvers:
  - Splitting based methods (PSOR, PGS, RPSOR, RPGS) of Acary and Brogliato (2008, Sect. 12.4.6).
  - Lemke's algorithm of Acary and Brogliato (2008, Sect. 12.4.7).
  - Newton's method of Acary and Brogliato (2008, Sect. 12.5.4).
- MLCP solvers:
  - Splitting based methods of Acary and Brogliato (2008, Sect. 12.4.6).
- NCP solvers.
  - Newton's method based on the Fischer–Burmeister function
  - Interface to the PATH solver described in Acary and Brogliato (2008, Sect. 13.5.3).
  - QP solver based on QLD due to Prof. K. Schittkowski of the University of Bayreuth, Germany (modification of routines due to Prof. M.J.D. Powell at the University of Cambridge).
- Frictional contact solvers:
  - Projection-type methods of Acary and Brogliato (2008, Sect. 13.7.2).
  - NSGS splitting based method of Acary and Brogliato (2008, Sect. 13.7.4).
  - Alart–Curnier's method of Acary and Brogliato (2008, Sect. 13.6.1).
  - NCP reformulation method of Acary and Brogliato (2008, Sect. 13.4.3).

6.2.4.3 SICONOS Kernel Components

As previously said, Kernel is the central and main part of the software. The whole dependencies among Kernel parts are fully depicted in Fig. 6.5. All the Kernel implementation is based on the principle we gave in Sect. 6.2.1. It is mainly composed of two rather distinct parts, modeling and simulation, that handle all the objects used, respectively, in the NSDS modeling (see Sect. 6.2.1.1) and the Simulation description (see Sect. 6.2.1.2).

The Utils module contains tools, mainly to handle classical objects such as matrices or vectors and is based on the Boost library,[4] especially, uBLAS,[5] a C++ library that provides BLAS functionalities for vectors, dense and sparse matrices.

[4] http://www.boost.org.

[5] http://www.boost.org/libs/numeric/ublas/doc/index.htm.

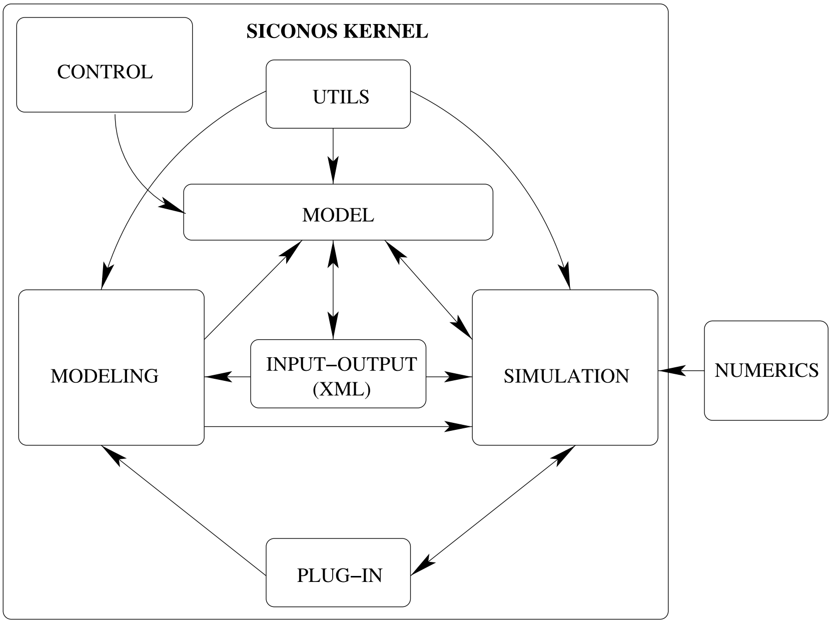


Fig. 6.5 Kernel components dependencies

The Input–Output module concerns objects for data management in XML format, thanks to the libxml2 see footnote[6] library. More precisely, all the description of the Model, NSDS and Simulation, can be done thanks to an XML input file.

Control package provides objects like `Sensor` and `Actuator`, to add control of the dynamical systems through the `Model` object, as explained in Sect. 6.2.1.3.

A plug-in system is available, mainly to allow the user to provide one's own computation methods for some specific functions (vector field of a dynamical system, mass, *etc.*), this without having to recompile the whole platform. Moreover, the platform is designed in a way that allows user to add dedicated modules through object registration and object factories mechanisms (for example to add a specific nonsmooth law, a user-defined sensor, *etc.*).

To conclude, class diagrams for modeling and simulation components are given in Figs. 6.6 and 6.7, which make clearer the various links between all the objects presented before.

6.3 The ACEF Module and Algorithms

The most technical part in the automatic circuit equation formulation is protected under patent Acary et al. (2009). This section presents the algorithms used by the automatic circuit equation formulation.

[6] http://xmlsoft.org/.

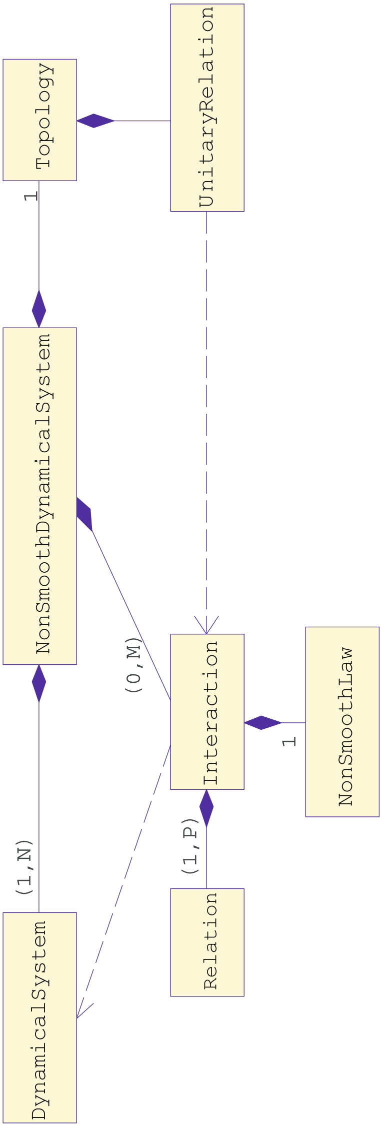


Fig. 6.6 Simplified class diagram for Kernel modeling part

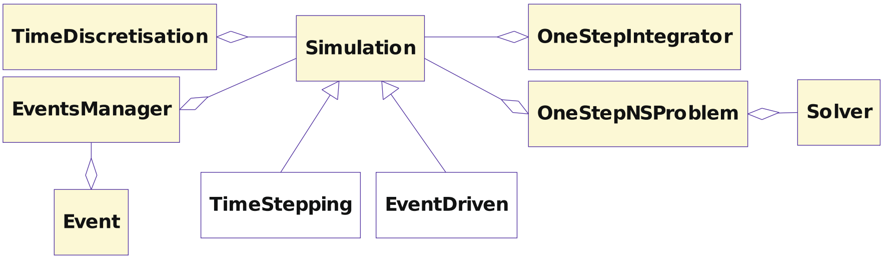


Fig. 6.7 Simplified class diagram for Kernel simulation part

6.3.1 A Module Able to Read a Circuit File: A Parser

The first step consists in reading and storing in memory a circuit from a textual description, a netlist SPICE file. The Application Programming Interface (API) of this module is briefly shown in the sequel:

(A) File read and store function:

- `int ParserReadFile(char *file)`
  Read and store in the local memory a netlist file

(B) Topological exploration functions:

- `int ParserInitComponentList(char *type)`
  Specify the type of component.
- `int ParserNextComponent(void * data)`
  Get data about the current component.
- `int ParserGetNbElementsOfType(char *type)`
  Get the number of component of a specified type.

(C) Source values functions:

- `int ParserComputeSourcesValues(double time)`
  Set time
- `int ParserGetSourceValue(char *type,void* id,double* value)`
  Get value of the corresponding source

(D) Initial values and simulation parameters function:

- `int ParserGetICvalue(int * numNode,int * icGiven, double * icValue)`
  Get initial values.
- `int ParserGetTransValues(double * step, double * stop, double * start)`
  Get simulation parameters

Algorithm 1 Build I_{NS}

```
Ensure: $\mathsf{index_{ins}}$ dimension of $\mathsf{I_{NS}}$.
Ensure: $\mathsf{index}_\lambda$ dimension of $\lambda$.
  $\mathsf{index_{ins}} \leftarrow 0$
  $\mathsf{index}_\lambda \leftarrow 0$
  for all nonsmooth component k do
    $\mathsf{COMP(k).indexStartIns} \leftarrow \mathsf{index_{ins}}$  // index concerning component k in $\mathsf{I_{NS}}$
    $\mathsf{index_{ins}} \leftarrow \mathsf{index_{ins}} + \mathsf{sizeCompIns}$  // Where sizeCompIns is the dimension of $\mathsf{I_{NS_k}}$
    $\mathsf{COMP(k).indexStart}\lambda \leftarrow \mathsf{index}_\lambda$  // index concerning component k in $\lambda$
    $\mathsf{index}_\lambda \leftarrow \mathsf{index}_\lambda + \mathsf{sizeComp}\lambda$  // Where sizeComp$\lambda$ is the dimension of $\lambda_\mathsf{k}$
    if nonsmooth component k is nonlinear then
      $\mathsf{COMP(k).geval} \leftarrow \mathsf{geval}$  // Plug functions $g_{\mathsf{NS}}$ and $h_{\mathsf{NS}}$. Used by the Newton linearization.
      $\mathsf{COMP(k).heval} \leftarrow \mathsf{heval}$
    end if
  end for
```

6.3.2 Build the Vector of Unknowns I_{NS}

With the notation of the system (4.25), Algorithm 1 builds the I_{NS} composed of all the I_{NS_k}. For each nonsmooth component, a location in the system h and g is reserved. At the end, the stamp algorithm of a nonsmooth component will consist in writing in this location its contribution.

6.3.3 An Algorithm to Choose the Unknowns

The algorithm described in this section builds the vectors of unknowns x and z of Sect. 3.6. It is based on the spanning tree algorithm. It also ensures the building of the subset C_{F} and C_{L}.

Following the development in Sect. 3.6.3, Algorithm 2 cuts the node indices in two complementary subsets:

- $\widetilde{N}$: The set of node indices whose KCL is written to build the system $N(x,t)x' = f(x,z,t)$.
- $\widehat{N}$: The complementary set of the previous set.

It also defines two complementary subsets of C such that

$$\mathsf{C} = \mathsf{C}_{\mathrm{F}} \cup \mathsf{C}_{\mathrm{L}}.$$

For each branch with index in C_{L}, an unknown has been added in z. For each branch with index in C_{F}, a KCL has been selected. This relation is stored in the list M. M is a list of couples used in Algorithm 3. Each couple of M contains an index of a capacitive branch and an index of a node. More precisely, a couple (c,n) means that the KCL of the node n will be used to describe the dynamic of the branch c. In the meantime, Algorithm 2 builds the following vectors of unknowns: $x = [I_{\mathsf{L}}, U_{C_{\mathrm{F}}}, U_{C_{\mathrm{L}}}]^T$ and $z = [V, I_{\mathsf{V}}, I_{\mathsf{NS}}, I_{C_{\mathrm{L}}}]^T$.

Algorithm 2 Build x and z

```
Require: Algorithm 1
Require: Spanning Tree (ST) of the capacitor branches indexed by C
Require: Init_ST: prepare an transversal (or depth-transversal) tree exploration.
Require: Next_branch_in_ST: return the branch indexes of an ST's edge.
Require: Next_branch_not_in_ST: return the branch indexes not included in the ST.
Require: index k: index of the current branch.
Require: index $i_k$ $j_k$: Nodes index of the current branch.
Require: index l: a node index.
Ensure: Build x and z, the vectors of unknowns.
Ensure: subset $\widetilde{N}$, $C_F$, $C_L$.
Ensure: A list of couple M.
  // Initialize x with $I_L$ and $U_C$.
  $x \leftarrow [I_L, U_C]^T$
  // Initialize z with $I_V$ .
  $z \leftarrow [V, I_V, I_{NS}]^T$
  // Initialize $\widetilde{N}$.
  $\widetilde{N} \leftarrow \emptyset$
  // Initialize M.
  $M \leftarrow \emptyset$
  Init_ST()
  k ← Next_branch_in_ST()
  while $k$ do
    $l \leftarrow i_k$ or $j_k$ with $l \notin \widetilde{N}$.
    add l in $\widetilde{N}$
    add k in $C_F$
    add (k, l) in M
    k ← Next_branch_in_ST ()
  end while
  k ←Next_branch_not_in_ST()
  while k do
    Add an unknown $I_k$ in z
    add k in $C_L$
    k ←Next_branch_not_in_ST()
  end while
```

6.3.4 Building the System $N(x, t)\dot{x} = f(x, z, t)$ of (3.70)

The first step consists in building a system (3.70), *i.e.* $N(x, t)\dot{x} = f(x, z, t)$ with a regular matrix $N(x, t)$. This algorithm specifies which physical law will be used in each line of the system. The writing in memory is ensured by the stamp method of each component. For each component and equation, an amount of memory has been allocated to store all the necessary parameters, that specify where and how

Algorithm 3 How build the system $\mathsf{N(x,t)x' = f(x,z,t)}$ with $\mathsf{N(x,t)}$ a regular matrix?

```
Require: x and z built by Algorithm 2.
Require: index k: index of the current branch.
Require: index $i_k j_k$: nodes index of the current branch.
Require: index l: a node index.
Require: integer i: current line number of the system, initialized to 0.
Require: memoryAlloc: a system function that reserved some memory.
Ensure: Stamp method to build the system $N(x,t)x' = f(x,,z,t)$.

  for all Inductive branches do
    // The law $Li_{L_k} = V_{i_k} - V_{j_k}$ will be used to fill the line i of the system.
    BCE(k).line $\leftarrow$ i
    i $\leftarrow$ i + 1
  end for
  // Difference with the standard MNA start here
  for all couple $(k,l) \in M$ do
    // The law KCL(l) will be used to fill the line i of the system with the, using $I_l = C_k \dot{U}_k$.
    KCL(l).line $\leftarrow$ i
    i $\leftarrow$ i + 1
  end for
  for all $k \in C_L$ do
    // Allocate memory for this BCE capacitor branch
    BCE(k)=memoryAlloc(sizeForCapacitorBranch)
    // The law $I_k = C_k \dot{U}_k$ will be used to fill the line number i of the system.
    BCE(k).line $\leftarrow$ i
  end for
```

to write the component contribution in the equation system. Moreover, the contribution of a component is written when the named function ‘stamp’ is called. In other words, memory has been allocated to customize the stamp algorithm. The stamp algorithm consists in analyzing these parameters to fill the system of equation. Algorithm 3 describes more precisely the process of building the matrix N of (3.105).

Remark 6.1 Our implementation consists in using some object oriented structures. We use some component and equation classes. For each component and equation, a corresponding instance of object is built.

6.3.5 Building the Relation $\mathbf{0} = g(x,z,t)$ *of (3.70)*

The previous section describes an algorithm to build a part of the dynamical system (3.70). This section presents an algorithm that specifies which law will be used to build the remaining part of the system, $0 = g(x,z,t)$.

Algorithm 4 How to build the system $0 = g(x, z, t)$?

```
Require: x and z built by the Algorithm 2.
Require: index k: index of the current branch.
Require: index $i_k j_k$:Nodes index of the current branch.
Require: indexl: a node index.
Require: integer i: current line number of the system, initialized to dim(x).
Ensure: It customizes the stamp methods to build the system $0 = g(x, z, t)$
  // KCL not written
  for all $l \in \widehat{N}$ do
    // The KCL(l) is recorded as the law of the line i of the system.
    $KCL(l).line \leftarrow i$
    $i \leftarrow i + 1$
  end for
  // BCE of voltage defined branches
  for all $k \in \widehat{B}$ and the branch k is voltage defined do
    // the law BCE: $V_{i_k} - V_{j_k} = u_k(I_L, I_V, A^T V, \dot{I}_L, \dot{I}_V, A^T \dot{V}, t)$ will be used to fill the line i of the system with.
    $BCE(k).line \leftarrow i$
    $i \leftarrow i + 1$
  end for
  // KVL of the capacitor branches
  for all $k \in C$ do
    // The law KVL of the capacitor branch k: $U_{C_i} = V_{i_k} - V_{j_k}$ will be used to fill the line i of the system with.
    $KVL(k).line \leftarrow i$
    $i \leftarrow i + 1$
  end for
```

Note that it consists in adding all physical equations of the complete circuit that are not written in Algorithm 3:

- KCL not used in the Algorithm 3. Indeed, for each node index in $\widehat{N}$, the corresponding KCL has not been written.
- Branch Constitutive Equation of voltage branch defined, not used in the Algorithm 3.
- KVL of the capacitor branches.

6.3.6 The Stamp Method for Nonsmooth Components

Algorithms 4 and 3 specify which laws are used to fill each line of the system (3.70). These algorithms are run only once. The stamp methods have to update the system following the rules imposed by the previous algorithms. It could be said that the Algorithms 4, 3 and 2 customize the stamp methods. Algorithm 12 shows how, in the nonlinear case, the stamp methods are called at each step of the simulation. Algorithms 9, 10 and 11 describe the stamp methods for the diode and the MOS transistor. Moreover, Algorithms 5, 6, 7 and 8 show how the stamp methods set the content of the memory. As it has been underlined Sect. 6.3.6, it depends strongly on Algorithm 2.

Algorithm 5 Resistor stamp algorithm

```
Require: k: an index of resistor.
Require: R: value from the circuit description.
Ensure: Fill the contribution of the component in the system
  // It concerns only the KCL(node1) and KCL(node2)
  MEM_F[KCL(node1).line, node1] $\leftarrow +\frac{1}{R}$ // It means $\frac{1}{R}$ is added
  MEM_F[KCL(node1).line, node2] $\leftarrow -\frac{1}{R}$ // It means $\frac{1}{R}$ is subtracted
  MEM_F[KCL(node2).line, node1] $\leftarrow -\frac{1}{R}$
  MEM_F[KCL(node2).line, node2] $\leftarrow +\frac{1}{R}$
```

Algorithm 6 Inductor stamp algorithm

```
Require: k: an index of inductor.
Require: $L_k$: is the index of $I_L$ in the vector x.
Require: L: value from the circuit description.
Ensure: Fill the contribution of the component in the system
  // It concerns the KCL(node1), KCL(node2) and BCE(k)
  MEM_N[BCE(k).line, $L_k$] $\leftarrow +L$
  MEM_F[BCE(k).line, node1] $\leftarrow +1$
  MEM_F[BCE(k).line, node2] $\leftarrow -1$ // $L_k$ is the index of$I_L$ in the vector x
  MEM_F[KCL(node1).line, $L_k$] $\leftarrow +1$
  MEM_F[KCL(node2).line, $L_k$] $\leftarrow -1$
```

6.3.7 Some Stamp Examples

For the sake of simplicity, this section provides some examples of the stamp algorithm of the form (5.12).

1. Diode stamp in H

$$\left(\begin{array}{cc} & I_{\mathrm{NS}_j} \\ \hline KCL(i) & +1 \\ KCL(j) & -1 \end{array}\right) \tag{6.16}$$

2. Diode stamp in g_{NS}

$$\left(\begin{array}{ccc} & V_j & V_i \\ \hline (\text{line in } g_{\mathrm{NS}}) & +1 & -1 \end{array}\right) \tag{6.17}$$

3. Diode stamp in h_{NS}

$$\left(\begin{array}{ccc} & I_{ns_j} & \lambda_j \\ \hline (\text{line in } h_{\mathrm{NS}}) & +1 & -1 \end{array}\right) \tag{6.18}$$

Algorithm 7 Capacitor stamp algorithm

```
Require: k: an index of capacitor.
Require: $UC_k$: index of the unknown $U_k$ in x
Require: $IC_k$: index of the unknown $I_k$ in z
Require: C: value from the circuit description.
Ensure: Fill the contribution of the component in the system
  // It concerns the KCL(node1), KCL(node2), KVL(k), and may be BCE(k)
  // About KVL
  MEM_F[KVL(k).line, $UC_k$] $\leftarrow -1$
  MEM_F[KVL(k).line, node1] $\leftarrow +1$
  MEM_F[KVL(k).line, node2] $\leftarrow -1$
  if node1 $\in \tilde{N}$ then
    if $k \in C_L$ then
      MEN_F[KCL(node1).line, ICk] $\leftarrow -1$
    else
      MEM_N[KCL(node1).line, UCk] $\leftarrow +C$
    end if
  end if
  if node2 $\in \tilde{N}$ then
    if $k \in C_L$ then
      MEN_F[KCL(node2).line, ICk] $\leftarrow +1$
    else
      MEM_N[KCL(node2).line, UCk] $\leftarrow -C$
    end if
  end if
  if $k \in C_L$ then
    // Write the BCE
    MEM_N[BCE(k).line, $UC_k$] $\leftarrow +C$
    MEM_F[BCE(k).line, $IC_k$] $\leftarrow +1$
  end if
```

6.3.8 The ACEF Global Execution Algorithm

In this section, it is shown how the algorithms of the previous sections are used to get the system of equations.

6.3.9 An Example of the Stamp Method with Nonsmooth Component

In Sect. 6.3.9.1, the standard MNA is applied on a smooth circuit and leads to a full implicit DAE. In Sect. 6.3.9.2, the previous algorithms are applied on a cir-

Algorithm 8 Capacitor stamp after inversion algorithm

```
Require: k: an index of capacitor.
Ensure: Fill the contribution of the component in the system
  // It could concern the KCL(node1) or KCL(node1)
  if $KCL(\text{node1}).line \geqslant \dim(x)$ then
    // get the current expression from the system $\dot{x} = ..$ and add it in the line KCL(node1).line of the system f2
  end if
  if $KCL(\text{node2}).line \geqslant \dim(x)$ then
    // get the current expression from the system $\dot{x} = ..$ and subtract it in the line KCL(node2).line of the system f2
  end if
```

Algorithm 9 Diode stamp algorithm

```
Require: k: an index of diode.
Require: InsIndex: index of the unknown $I_{NS_k}$ in z.
Ensure: Fill the contribution of the component in the system
  MEM_F[KCL(node1).line, dim(x) + InsIndex] $\leftarrow +1$
  MEM_F[KCL(node2).line, dim(x) + InsIndex] $\leftarrow -1$
  // About the equation h: $0 = -I_{NS_k} + \lambda_k$
  MEM_H[COMP(k).indexStartIns, dim(x) + InsIndex] $\leftarrow -1$
  MEM_H[COMP(k).indexStartIns, dim(x) + dim(z) + COMP(k).indexStart$\lambda$]
     $\leftarrow +1$
  // About the equation g: $y_k = V_1 - V_2$
  MEM_G[COMP(k).indexStart$\lambda$, dim(x) + node1] $\leftarrow +1$
  MEM_G[COMP(k).indexStart$\lambda$, dim(x) + node2] $\leftarrow -1$
```

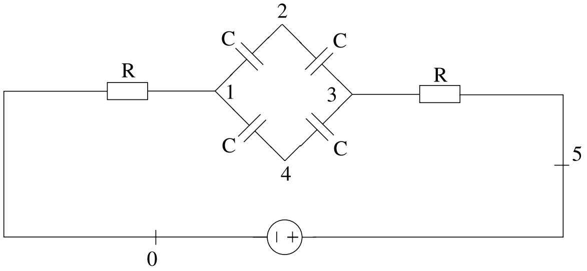


Fig. 6.8 Circuit containing a loop of capacitors

cuit with a nonsmooth and nonlinear MOS transistor and leads to a semi-explicit system.

6.3.9.1 Standard MNA Algorithm

The circuit is depicted in Fig. 6.8. The vectors of unknowns are $x = (U_{12}, U_{23}, U_{34}, U_{41})^T$ and $z = (V_1, V_2, V_3, V_4, V_5, I_{50})^T$.

Algorithm 10 Linear and nonsmooth MOS stamp algorithm

```
Require: k: index of the mos component.
Require: node$_g$: index of the gate node.
Require: node$_s$: index of the source node.
Require: node$_d$: index of the drain node.
Require: InsIndex: index of the unknown $I_{NS_k}$ in z (ie $I_{ds}$).
Ensure: Fill the contribution of the component in the system
  $N_{hyp} = COMP(k).N_{hyp}$
  // KCL contribution
  MEM_F[KCL(node$_d$).line, dim(x) + InsIndex] $\leftarrow +1$
  MEM_F[KCL(node$_s$).line, dim(x) + InsIndex] $\leftarrow -1$
  // About the equation h:$0 = -I_{NS_k} + C\lambda$
  // $-I_{NS_k}$
  MEM_H[COMP(k).indexStartIns, dim(x) + InsIndex] $\leftarrow -1$
  for all $0 \leqslant n < N_{hyp}$ do

    MEM_H[COMP(k).indexStartIns, dim(x) + dim(z) + COMP(k).indexStart$\lambda$ + n]
      $\leftarrow$ COMP[k].mCoefs[n]
    MEM_H[COMP(k).indexStartIns, dim(x) + dim(z) + COMP(k).indexStart$\lambda$
      + $N_{hyp}$ + n] $\leftarrow$ −COMP[k].mCoefs[n]
  end for
  // About the equation Y=g().
  for all $0 \leqslant n < N_{hyp}$ do
    MEM_G[COMP(k).indexStart$\lambda$ + n, dim(x) + node$_g$] $\leftarrow 1$
    MEM_G[COMP(k).indexStart$\lambda$ + n + $N_{hyp}$, dim(x) + node$_g$] $\leftarrow 1$
    MEM_G[COMP(k).indexStart$\lambda$ + n, dim(x) + node$_s$] $\leftarrow -1$
    MEM_G[COMP(k).indexStart$\lambda$ + n + $N_{hyp}$, dim(x) + node$_d$] $\leftarrow -1$
    MEM_G[COMP(k).indexStart$\lambda$ + n, dim(x) + dim(z) + COMP(k).indexStart$\lambda$
      + n] $\leftarrow 1$
    MEM_G[COMP(k).indexStart$\lambda$ + n + $N_{hyp}$, dim(x) + dim(z)
      + COMP(k).indexStart$\lambda$ + n + $N_{hyp}$] $\leftarrow 1$
    MEM_G[COMP(k).indexStart$\lambda$ + n, dim(x) + dim(z) + dim($\lambda$) + 1]
      $\leftarrow$ COMP[k].mh[n]
    MEM_G[COMP(k).indexStart$\lambda$ + n + $N_{hyp}$, dim(x) + dim(z) + dim($\lambda$) + 1]
      $\leftarrow$ COMP[k].mh[n]
  end for
```

Let us start to write M in (3.30). This yields

$$\begin{pmatrix} KCL(1) \\ KCL(2) \\ KCL(3) \\ KCL(4) \end{pmatrix} \left(\begin{array}{cccc} \dot{U}_{12} & \dot{U}_{23} & \dot{U}_{34} & \dot{U}_{41} \\ \hline C & 0 & 0 & -C \\ -C & C & 0 & 0 \\ 0 & -C & C & 0 \\ 0 & 0 & -C & C \end{array} \right) \dot{x} = RHS.$$

Algorithm 11 Nonlinear and nonsmooth MOS stamp algorithm

```
Require: k: index of the mos component.
Require: node$_g$: index of the gate node.
Require: node$_s$: index of the source node.
Require: node$_d$: index of the drain node.
Require: InsIndex: index of the unknown $I_{NS_k}$ in z (ieI$_{ds}$).
Ensure: Fill the contribution of the component in the system
  $K = COMP(k).K$
  $V_T = COMP(k).V_T$
  // KCL contribution
  MEM_F[KCL(node$_d$).line, dim(x) + InsIndex] $\leftarrow +1$
  MEM_F[KCL(node$_s$).line, dim(x) + InsIndex] $\leftarrow -1$
  // About the equation h: $0 = -I_{NS_k} + \frac{K}{2}(\lambda_4(V_{gs} - V_T)^2 - \lambda_2(V_{gd} - V_T)^2)$
  // $-I_{NS_k}$
  MEM_H[COMP(k).indexStartIns, dim(x) + InsIndex] $\leftarrow -1$
  // $\frac{dh}{dV_g}$
  MEM_$\nabla_z$H[COMP(k).indexStartIns, node$_g$] $\leftarrow K(\lambda_4 * (V_g - V_s - V_T)$
      $- \lambda_2(V_g - V_d - V_T))$
  // $\frac{dh}{dV_s}$
  MEM_$\nabla_z$H[COMP(k).indexStartIns, node$_s$] $\leftarrow -K(\lambda_4 * (V_g - V_s - V_T))$
  // $\frac{dh}{dV_d}$
  MEM_$\nabla_z$H[COMP(k).indexStartIns, node$_d$] $\leftarrow K(\lambda_2 * (V_g - V_d - V_T))$
  // $\frac{dh}{d\lambda_2}$
  MEM_$\nabla_\lambda$H[COMP(k).indexStartIns, COMP(k).indexStart$\lambda$ + 1]
      $\leftarrow -\frac{K}{2}K((V_g - V_d - V_T)^2)$
  // $\frac{dh}{d\lambda_4}$
  MEM_$\nabla_\lambda$H[COMP(k).indexStartIns, COMP(k).indexStart$\lambda$ + 3]
      $\leftarrow \frac{K}{2}K((V_g - V_s - V_T)^2)$
  // About the equation g:
  // $\frac{dg}{dV_g}$
  MEM_$\nabla_z$G[COMP(k).indexStart$\lambda$ + 1, node$_g$] $\leftarrow -1$
  MEM_$\nabla_z$G[COMP(k).indexStart$\lambda$ + 3, node$_g$] $\leftarrow -1$
  // $\frac{dg}{dV_d}$
  MEM_$\nabla_z$G[COMP(k).indexStart$\lambda$ + 1, node$_d$] $\leftarrow +1$
  // $\frac{dg}{dV_s}$
  MEM_$\nabla_z$G[COMP(k).indexStart$\lambda$ + 3, node$_s$] $\leftarrow +1$
  // $\frac{dg}{d\lambda_1}$
  MEM_$\nabla_\lambda$G[COMP(k).indexStart$\lambda$ + 1, COMP(k).indexStart$\lambda$] $\leftarrow +1$
  // $\frac{dg}{d\lambda_2}$
  MEM_$\nabla_\lambda$G[COMP(k).indexStart$\lambda$, COMP(k).indexStart$\lambda$ + 1] $\leftarrow -1$
  // $\frac{dg}{d\lambda_3}$
  MEM_$\nabla_\lambda$G[COMP(k).indexStart$\lambda$ + 3, COMP(k).indexStart$\lambda$ + 2] $\leftarrow +1$
  // $\frac{dg}{d\lambda_4}$
  MEM_$\nabla_\lambda$G[COMP(k).indexStart$\lambda$ + 2, COMP(k).indexStart$\lambda$ + 3] $\leftarrow -1$
```

Algorithm 12 ACEF global algorithm

```
Require: circuit.cir: a netlist file
Ensure: The simulation result is stored in memory(not yet describe in this version)
  // read and store the circuit
  ParserReadFile("circuit.cir")
  // Build $I_n$s
  Perform Algorithm 1
  // Build x and z using the parser interface
  Perform Algorithm 2
  // Prepare the stamp Algorithms
  Perform Algorithms 3 and 4
  t ← $T_{init}$ // Initialize t to initial time
  while $t < T_{end}$ do
    // We are now ready to stamp the system $N(x, t) = f(x, z, t)$
    for all component k do
      COMP(k).stamp()
    end for
    // We are now ready to stamp the system $0 = g(x, z, t)$
    for all capacitor k do
      COMP(k).stampAfterInversion()
    end for
    // Solve the complementarity problem to get current point of the simulation
    ...
  end while
```

The obtained matrix is not regular because of the cycle $\{1\text{-}2, 2\text{-}3, 3\text{-}4, 4\text{-}1\}$, so it leads to a implicit DAE. The Right-hand-side RHS is not detailed here.

The previous algorithms are now applied. The chosen solution is to use a Spanning Tree $\{1\text{-}2, 2\text{-}3, 3\text{-}4\}$ to write the KCL. For the capacitive branch {4, 1}, I_{41} is added in the vector of unknowns z. Algorithm 3 provides the following semi-explicit system:

$$\begin{pmatrix} KCL(1) \\ KCL(2) \\ KCL(3) \\ I_{41} \end{pmatrix} \left(\begin{array}{cccc} \dot{U}_{12} & \dot{U}_{23} & \dot{U}_{34} & \dot{U}_{41} \\ \hline C & 0 & 0 & -C \\ -C & C & 0 & 0 \\ 0 & -C & C & 0 \\ 0 & 0 & 0 & C \end{array}\right) \dot{x}$$

$$= 0x + \left(\begin{array}{ccccccc} V_1 & V_2 & V_3 & V_4 & V_5 & I_{50} & I_{41} \\ \hline -\frac{1}{R} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{R} & -\frac{1}{R} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right) z.$$

Multiplying by N^{-1}, all the currents are then expressed as a linear combination of the unknowns:

$$\begin{aligned} &I_{41} \quad \text{unknown variable,}\\ &I_{12} = I_{41} - \frac{V_1}{R},\\ &I_{23} = I_{41} - \frac{V_1}{R},\\ &I_{43} = I_{41} - \frac{V_1}{R} - \frac{V_3 - V_5}{R}. \end{aligned} \tag{6.19}$$

The last step consists in writing the missing equations with the relation (6.19):

$$\begin{pmatrix} KCL(4)\\ KCL(5)\\ U_{12}\\ U_{23}\\ U_{34}\\ U_{41}\\ VD_{50} \end{pmatrix} \left(\begin{array}{cccc} U_{12} & U_{23} & U_{34} & U_{41}\\ \hline 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ 1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\\ 0 & 0 & 0 & 0 \end{array}\right) x$$

$$+ \left(\begin{array}{ccccccc} V_1 & V_2 & V_3 & V_4 & V_5 & I_{50} & I_{41}\\ \hline \frac{1}{R} & 0 & -\frac{1}{R} & 0 & \frac{1}{R} & 0 & \underline{-1+1}\\ 0 & 0 & \frac{1}{R} & 0 & -\frac{1}{R} & -1 & 0\\ -1 & 1 & 0 & 0 & 0 & 0 & 0\\ 0 & -1 & 1 & 0 & 0 & 0 & 0\\ 0 & 0 & -1 & 1 & 0 & 0 & 0\\ 0 & 0 & 0 & -1 & 1 & 0 & 0\\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{array}\right) z = \begin{pmatrix} 0\\ 0\\ 0\\ 0\\ 0\\ 0\\ E \end{pmatrix},$$

$N_I = N_E = 11.$

Note that the first line $KCL(1)$ expresses that the current which goes in the capacitor cycle is equals to the current that goes out.

6.3.9.2 A Nonsmooth Nonlinear Example

This section provides an example composed of a MOS component with nonlinear model described by (4.53). Let us consider the following netlist file describing our example:

```
exampleNSNL
.model nmos nmos level=1 tox=100e-09 vto=0.2 kp=10.0
VIN 1 0 AC 1 SIN(0 1 10)
C1 1 2 1mF
C2 2 4 1mF
C3 1 3 1mF
C4 3 4 1mf
R2 0 4 500
mosnswitch 5 4 0 0 nmos w=0.0001 l=0.0001
L1 5 6 10u
R1 0 6 500
.tran 0.001 1.25
.print tran v(1) V(2)-V(1)
.END
```

The circuit is depicted in Fig. 6.9.

The MOS model is described by (4.53), that leads to:

$$\begin{cases} I_{DS} = I_{50} = \frac{K}{2}(\lambda_4(V_4 - V_T)^2 - \lambda_2(V_5 - V_4 - V_T)^2), \\ \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 1-\lambda_2 \\ V_T - V_5 + V_4 + \lambda_1 \\ 1-\lambda_4 \\ V_T - V_4 + \lambda_3 \end{pmatrix}, \\ 0 \leqslant \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \end{pmatrix} \perp \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} \geqslant 0. \end{cases} \tag{6.20}$$

An ideal transistor adds only I_{DS} in the vector I_{NS}, in our case it is:

$$I_{\mathsf{NS}} = (I_{50}).$$

The vector X is defined by:

$$X = (I_L, U_{C_1}, U_{C_2}, U_{C_3}, U_{C_4})^T.$$

Concerning Z, note that I_{34} has been added by Algorithm 2

$$Z = (V_1, V_2, V_3, V_4, V_5, V_6, I_{01}, I_{50}, I_{34}).$$

The spanning tree that is used is composed of the edges: (2,1), (2,4) and (1,3).

$$\begin{aligned} &M = \{(C_1, 2), (C_2, 4), (C_3, 1)\}, \\ &C_{\mathrm{F}} = \{C_1, C_2, C_3\}, \\ &C_{\mathrm{L}} = \{C_4\}, \\ &\widetilde{N} = \{2, 4, 1\}, \\ &\widehat{N} = \{3, 5, 6\} \end{aligned}$$

and the index concerning the nonsmooth law:

$$index_{ins} = 1 \qquad index_{\lambda} = 4.$$

The content of the MOS component memory is:

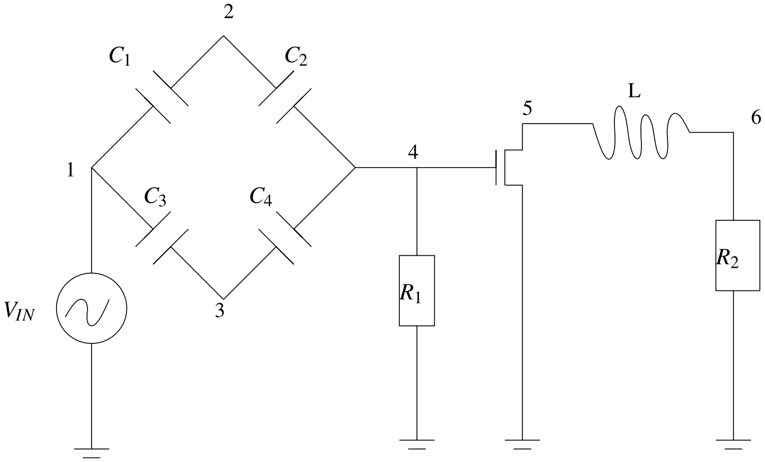


Fig. 6.9 Circuit with a loop of capacitors and a nonsmooth electrical element

- $COMP(k_{mos}).indexStart = 0$
- $COMP(k_{mos}).indexStart = 0$
- and the functions to compute g, h and the gradient ∇ has been plugged.

The aim of these algorithms is to fill the memory of the law:

- $BCE(k_L).line = 0$
- $KCL(2).line = 1$
- $KCL(4).line = 2$
- $KCL(1).line = 3$
- $BCE(C_4).line = 4$
- $KCL(3).line = 5$
- $KCL(5).line = 6$
- $KCL(6).line = 7$
- $BCE(V_{IN}).line = 8$
- $KVL(C_1).line = 9$
- $KVL(C_2).line = 10$
- $KVL(C_3).line = 11$
- $KVL(C_4).line = 12$

We are now ready to build the system using the stamp algorithm of each component.

About MEM_N:

L				
	C_1	$-C_2$		
		C_2		
	$-C_1$		$-C_3$	
				C_4

About MEM_F:

I_L	U_{c1}	U_{c2}	U_{c3}	U_{c4}	V_1	V_2	V_3	V_4	V_5	V_6	I_{01}	I_{50}	I_{43}
									-1	1			
								$-\frac{1}{R_2}$					-1
											-1		
													1
-1												1	
1										$\frac{1}{R_1}$			
					1								
	-1				-1	1							
		1				-1		1					
			-1		-1		1						
				-1			-1	1					

After the inversion of N, we get the missing current through the node 3:

$$C_3\dot{U}_{C_3} = I_{34} + \frac{V_4}{R_2} + I_{01}.$$

The second stamp after the inversion of N change only the line 6 about the $KCL(3)$:

$$\begin{array}{|c|c|c|c|c|c|c|c|c|c|c|c|c|c|}\hline &&&&&&&& \frac{1}{R_2} &&& 1 && \\ \hline\end{array},$$

We note that $\nabla_x h$ and $\nabla_x g$ are null.

$$\nabla_z h = \begin{array}{|c|c|c|c|c|c|c|c|c|}\hline &&& K(\lambda_4(V_4+V_T)+\lambda_2(V_5-V_4-V_T)) & -K\lambda_2(V_5-V_4-V_T) &&&& \\ \hline\end{array},$$

$$\nabla_\lambda h = \begin{array}{|c|c|c|c|}\hline & -K(V_5-V_4-V_T)^2/2 & & K(-V_4-V_T)^2/2 \\ \hline\end{array},$$

$$\nabla_z g = \begin{array}{|c|c|c|c|c|c|c|c|c|}\hline &&&&&&&& \\ \hline &&& 1 & -1 &&&& \\ \hline &&&&&&&& \\ \hline &&& 1 & &&&& \\ \hline\end{array} \quad \text{and} \quad \nabla_\lambda g = \begin{array}{|c|c|c|c|}\hline & -1 & & \\ \hline 1 & & & \\ \hline & & & -1 \\ \hline & & 1 & \\ \hline\end{array}.$$

Chapter 7
Simple Circuits

This chapter is devoted to present numerical simulation results obtained with the SICONOS platform, on several simple circuits: the first circuit has been built to show that conventional analog simulators fail to converge; the other circuits are classical diode-bridge wave rectifiers, and the last one is a circuit that exhibits a sliding mode. In this chapter and in Chap. 8, five simulation software packages were used:

SICONOS: the platform developed at INRIA Grenoble Rhône-Alpes (France) dealing with nonsmooth dynamical systems with dedicated time integrators and algorithms to solve sets of equations and inequalities (for instance LCP: linear complementarity problems).

NGSPICE: an open-source version of the original SPICE3F5 software developed by Berkeley university. Even if this version may differ from existing commercial ones, it shares with them a common set of models and the solving algorithms belong also to the same class that deals with regular functions.

SMASH: a commercial version of SPICE developed by Dolphin Integration (see http://www.dolphin.fr).

ELDO: a commercial version of SPICE with Newton-Raphson and OSR (one step relaxation) algorithms developed by Mentor Graphics (http://www.mentor.com).

PLECS: a SIMULINK/MATLAB toolbox dedicated to the simulation of power electronics circuits (see http://www.plexim.com). The models and algorithms come from the hybrid approach. In our work the freely available demonstration version of PLECS has been used.

7.1 Maffezzoni's Example

This section is devoted to the modeling and the simulation of the circuit in Fig. 7.1. In Maffezzoni et al. (2006) it is shown that Newton-Raphson based methods fail to converge on such a circuit, with the switch model as in (1.46). The diode model is the equivalent resistor model in Fig. 1.2(d). On the contrary the OSNSP solver correctly behaves on the same model, as demonstrated next.

V. Acary et al., *Nonsmooth Modeling and Simulation for Switched Circuits*,
Lecture Notes in Electrical Engineering 69,
DOI 10.1007/978-90-481-9681-4_7, © Springer Science+Business Media B.V. 2011

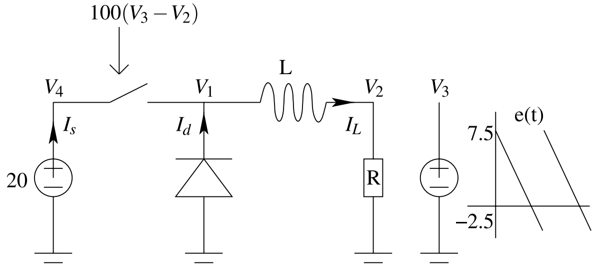


Fig. 7.1 A simple switched circuit

7.1.1 The Dynamical Model

The dynamics of the circuit in Fig. 7.1 is obtained using the algorithm of automatic circuit equation formulation of Chap. 6. In a first step, the vector of unknown variables is built, in a second step, the dynamical system is written, and in a last step, the nonsmooth laws are added. Applying the automatic equations generation algorithm leads to the following 9-dimensional unknown (dynamic and algebraic unknown variables) vector: $X = (V_1\ V_2\ V_3\ V_4\ I_L\ I_{03}\ I_{04}\ I_s\ I_d)^T$ in the system (5.1) or $x = (I_L)$ and $z = (V_1\ V_2\ V_3\ V_4\ I_{03}\ I_{04}\ I_s\ I_d)^T$ in the system (5.14), where the potentials and the currents are depicted in Fig. 7.1. Building the dynamical equations consists in writing the Kirchhoff current laws at each node, the constitutive equation of the smooth branch, and the nonsmooth law of the other branches. The two nonsmooth devices are the diode and the switch. It yields the following system, that fits within the general framework in (3.32). For the semi-explicit DAE, we obtain:

$$\begin{cases} L\frac{dI_L}{dt}(t) = V_1(t) - V_2(t), \\ I_d(t) + I_s(t) - I_L(t) = 0, \quad I_L(t) - \frac{V_2(t)}{R} = 0, \\ I_{03}(t) = 0, \qquad I_{04}(t) - I_s(t) = 0, \\ V_4(t) = 20, \qquad V_3 = e(t). \end{cases} \tag{7.1}$$

For the input/output relations of the nonsmooth components, we get:

$$\begin{cases} V_1(t) = \frac{1}{2}(\tau_1(t) - 1)R_{\mathsf{off}} I_d(t) - \frac{1}{2}(\tau_1(t) + 1)R_{\mathsf{on}} I_d(t), \\ 2(V_4(t) - V_1(t)) = [(1 + \tau_2(t))R_{\mathsf{off}} + (1 - \tau_2(t))R_{\mathsf{on}}]I_s(t). \end{cases} \tag{7.2}$$

Finally, the inclusion rule is written as:

$$\begin{cases} V_1(t) \in -\mathbb{N}_{[-1,1]}(\tau_1(t)) \\ 100(V_3(t) - V_2(t)) \in -\mathbb{N}_{[-1,1]}(\tau_2(t)). \end{cases} \tag{7.3}$$

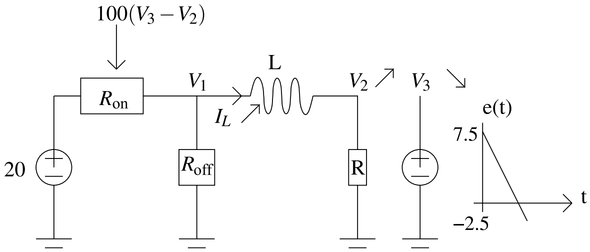


Fig. 7.2 Equivalent linear circuit

On this example, the fully implicit ($\theta = 1$) Moreau's time-stepping scheme reads as:

$$\begin{cases} L(I_{L,k+1} - I_L) = h(V_{1,k+1} - V_{2,k+1}), \\ I_{d,k+1} + I_{s,k+1} - I_{L,k+1} = 0, \quad I_{L,k+1} - \frac{1}{R} V_{2,k+1} = 0, \\ I_{03,k+1} = 0, \quad I_{04,k+1} - I_{s,k+1} = 0, \\ V_{4,k+1} = 20, \quad V_{3,k+1} = e(t_{k+1}), \\ 2V_{1,k+1} = (\tau_{1,k+1} - 1)R_{\mathsf{off}} I_{d,k+1} - (\tau_{1,k+1} + 1)R_{\mathsf{on}} I_{d,k+1}, \\ 2(V_{4,k+1} - V_{1,k+1}) = [(1 + \tau_{2,k+1})R_{\mathsf{off}} + (1 - \tau_{2,k+1})R_{\mathsf{on}}] I_{s,k+1}, \\ V_{1,k+1} \in -\mathbb{N}_{[-1,1]}(\tau_{1,k+1}), \\ 100(V_{3,k+1} - V_{2,k+1}) \in -\mathbb{N}_{[-1,1]}(\tau_{2,k+1}). \end{cases} \tag{7.4}$$

7.1.2 Simulation Results: Failure of the Newton-Raphson Algorithm

The simulation consists of two phases: a linear behavior followed by a change in the state of the switch.

7.1.2.1 A Linear Behavior

The time step has been fixed to 0.1 µs and the initial state is $\{0, 7.5, 0\}$. While the value of $V_3 - V_2$ is positive, the Newton-Raphson algorithm converges in one iteration. It is a linear circuit shown in Fig. 7.2. During this period, I_L and V_2 are increasing and V_3 is decreasing. When V_2 becomes equals to V_3, the switch will change its state.

7.1.2.2 The Switch Change of State

We denote N the integer such that the switch will change its state on $[t_N, t_{N+1}]$. Figure 7.3 depicts the initial linearized circuit used by the Newton-Raphson iterations to compute the state at t_{N+1}. Setting V_3 to the value $e(t_{N+1})$, that is 1.20 V, leads to change the switch's state. Figure 7.4 shows the equivalent circuit after the first

Fig. 7.3 Circuit state at $t = t_N$

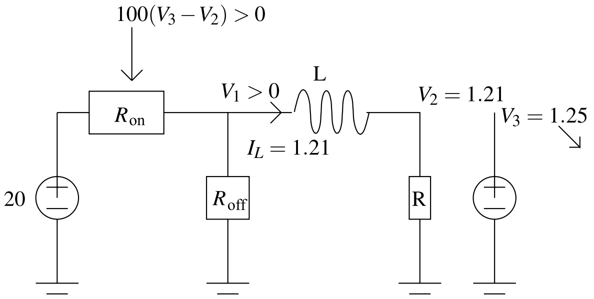


Fig. 7.4 Equivalent linear model, first step

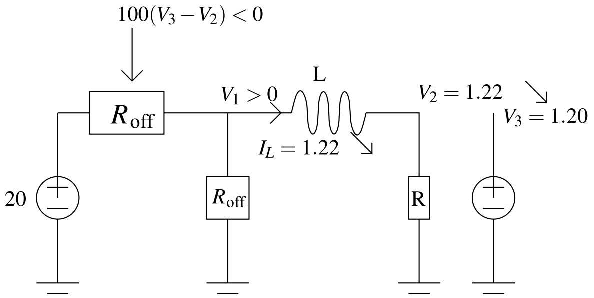
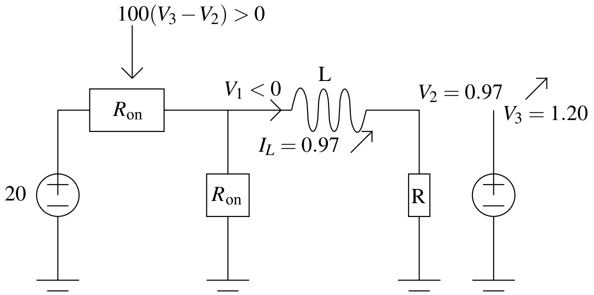


Fig. 7.5 Equivalent linear model, second step



Newton-Raphson iteration. We note that the switch is now OFF, due to the negative value of $V_2 - V_3$. The second Newton-Raphson iteration causes the decreasing of V_2 resulting in changing the states of the diode and of the switch. Figure 7.5 points out the new state of the circuit. The third iteration results in a new setting of both the switch and diode components depicted in Fig. 7.6. On this example, the linearization performed at each Newton-Raphson iteration leads to an oscillation between two incorrect states and never converges to the correct one. The Newton-Raphson iterations enter into an infinite loop without converging.

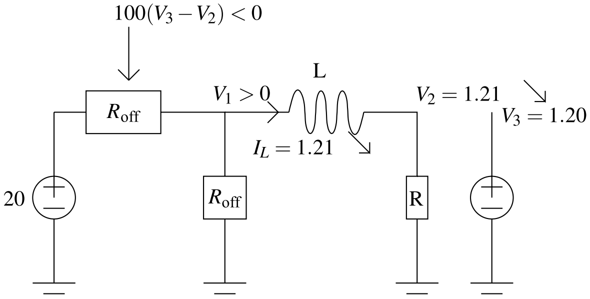


Fig. 7.6 Equivalent linear model, third step

7.1.2.3 The Newton-Raphson Iterations at $t = t_N$

The next table summarizes the oscillation between two incorrect states:

	$k=0$	$k=1$	$k=2$	$k=3$	$k=4$	...	Solution
S	*ON*	*OFF*	*ON*	*OFF*	*ON*	...	*OFF*
D	*OFF*	*OFF*	*ON*	*OFF*	*ON*	...	*ON*

(7.5)

7.1.3 Numerical Results with Siconos

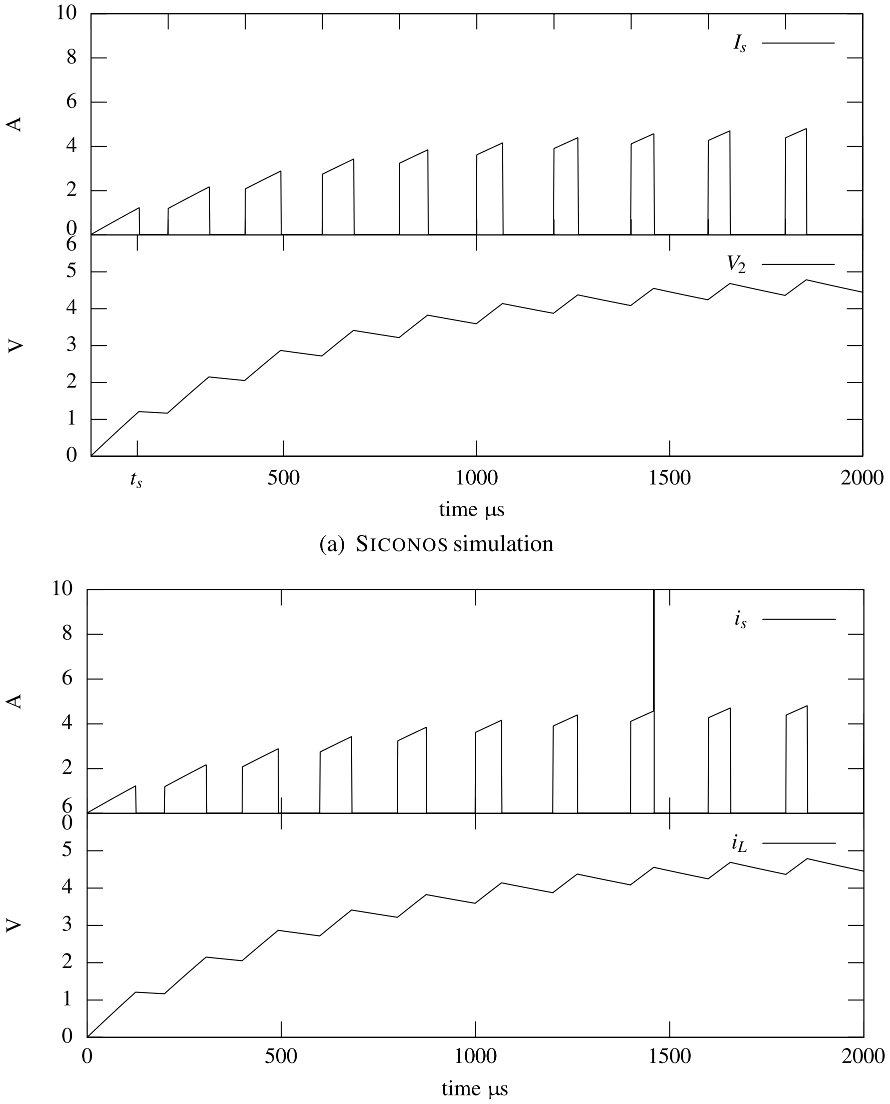
The time step has been fixed to 0.1 μs, the values of the parameters are $R = 1\ \Omega$, $R_{\mathsf{on}} = 0.001\ \Omega$, $R_{\mathsf{off}} = 1000\ \Omega$, $L = 2.10^{-4}$ H and the initial condition is $I_L(0) = 0$ A. Figure 7.7(a) depicts the current evolution through the inductor L. Using the NSDS approach the OSNSP solver converges and computes the correct state. For such a simple system, any OSNSP solver gives a correct solution. We have used indifferently PATH and a SEMISMOOTH Newton method.

Remark 7.1 In Maffezzoni et al. (2006) an event-driven numerical method is proposed to solve the non convergence issue. However it is reliable only if the switching times can be precisely estimated, a shortcoming not encountered with the NSDS and the Moreau's time-stepping method.

7.1.4 Numerical Results with Eldo

Eldo does not provide any nonsmooth switch model. But it furnishes the 'VSWITCH' one described in (7.6), where R_S is the controlled resistor value of the switch, and V_c the voltage control yielding to the model:

$$R_S(t) = \begin{cases} R_{\mathsf{on}} & \text{if } V_c(t) \geqslant V_{\mathsf{on}}, \\ R_{\mathsf{off}} & \text{if } V_c(t) \leqslant V_{\mathsf{off}}, \\ (V_c(t)(R_{\mathsf{off}} - R_{\mathsf{on}}) + R_{\mathsf{on}}\, V_{\mathsf{off}} & \\ \quad - R_{\mathsf{off}}\, V_{\mathsf{on}})/(V_{\mathsf{off}} - V_{\mathsf{on}}) & \text{otherwise.} \end{cases} \tag{7.6}$$



(a) SICONOS simulation

(b) ELDO simulation

Fig. 7.7 Switched circuit simulations

Setting V_{off} to 0 and choosing a small value for V_{on} lead to a model close to (1.46) for the chosen parameters. Simulations have been done using different sets of parameters. It is noteworthy that the behavior of ELDO depends on these values. For example, using a backward scheme Euler with the time step fixed to 0.1 µs and $V_{\mathrm{on}} = 10^{-4}$ V, $V_{\mathrm{off}} = 0$ V, $R_{\mathrm{off}} = 1000\ \Omega$, $R_{\mathrm{on}} = 0.001\ \Omega$ causes trouble during the ELDO simulation: '`Newton no-convergence`' messages appear. Figure 7.7(b)

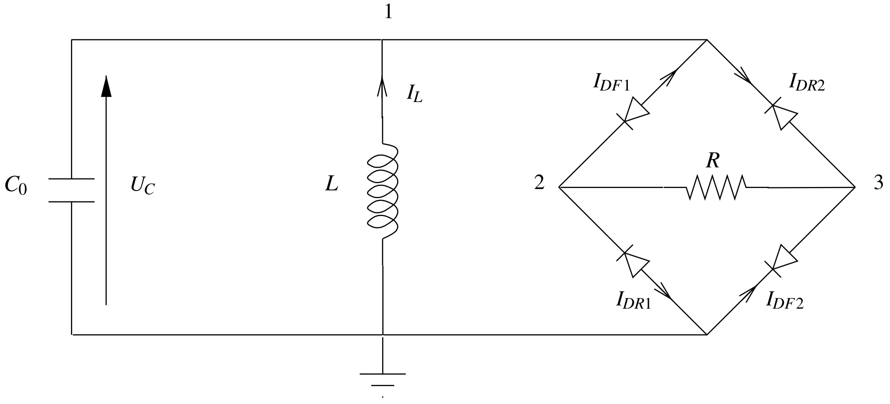


Fig. 7.8 A 4-diode bridge wave rectifier

shows the ELDO simulation. The values are very close to the SICONOS simulation, except for the steps corresponding to the ‘`no-convergence`’ messages. In this case, the resulting current value is absurd.

This academic example demonstrates that standard analog tools (SPICE-like simulators) can fail to simulate a switched circuit.

7.2 A First Diode-Bridge Wave Rectifier

The chosen example is a four-diode bridge wave rectifier as shown in Fig. 7.8. In this sample, an LC oscillator initialized with a given voltage across the capacitor and a null current through the inductor provides the energy to a load resistance through a full-wave rectifier consisting of a 4-ideal-diode bridge. Both waves of the oscillating voltage across the LC are provided to the resistor with current flowing always in the same direction. The energy is dissipated in the resistor resulting in a damped oscillation. This section presents the modeling and the simulation of this circuit using the SICONOS platform and the automatic circuit equation formulation presented in Chap. 6.

7.2.1 Dynamical Equations

The automatic circuit equation formulation leads to the system (7.7), (7.9) and (7.8). The vector of unknown variables is $(U_C, I_L, V_1, V_2, V_3, I_{DF1}, I_{DF2}, I_{DR1}, I_{DR2})^T \in \mathbb{R}^9$. The potentials V_1, V_2, V_3 are the potentials at the points indicated on the figure. We obtain:

$$\begin{cases} -C_0\dot{U}_C(t)+I_L(t)-I_{DR2}(t)+I_{DF1}(t)=0,\\ L\dot{I}_L(t)+V_1(t)=0,\\ \frac{V_3(t)-V_2(t)}{R}-I_{DR1}(t)-I_{DF1}(t)=0,\\ \frac{V_2(t)-V_3(t)}{R}+I_{DR2}(t)+I_{DF2}(t)=0,\\ U_C(t)-V_1(t)=0, \end{cases} \tag{7.7}$$

$$\begin{cases} 0\leqslant\lambda_1(t)\perp V_2(t)-V_1(t)\geqslant 0,\\ 0\leqslant\lambda_2(t)\perp -V_3(t)\geqslant 0,\\ 0\leqslant\lambda_3(t)\perp V_2(t)\geqslant 0,\\ 0\leqslant\lambda_4(t)\perp V_1(t)-V_3(t)\geqslant 0, \end{cases} \tag{7.8}$$

with:

$$\lambda_1=-I_{DF1}\qquad \lambda_2=-I_{DF2}\qquad \lambda_3=-I_{DR1}\qquad \lambda_4=-I_{DR2}. \tag{7.9}$$

One may identify this dynamics with the canonical form in (2.54), where x may be chosen as the above vector of unknown variables. One has:

$$E=\begin{pmatrix} -C_0 & 0 & 0 & \cdots & 0\\ 0 & L & 0 & \cdots & 0\\ 0 & \cdots & & & 0\\ & & \vdots & & \\ 0 & \cdots & & & 0 \end{pmatrix}\in\mathbb{R}^{9\times 9},\qquad M=I_4,$$

$$C=\begin{pmatrix} 0 & 0 & -1 & 1 & 0 & \cdots & & 0\\ 0 & 0 & 0 & 0 & -1 & 0 & & 0\\ 0 & 0 & 0 & 1 & 0 & 0 & \cdots & 0\\ 0 & 0 & 1 & 0 & -1 & 0 & \cdots & 0 \end{pmatrix}\in\mathbb{R}^{4\times 9}.$$

Here I_4 is the 4 × 4 identity matrix.

7.2.2 Simulation Results

Figure 7.9 shows a simulation with SICONOS using the following numerical values: $L=10^{-2}$ H, $C=10^{-6}$ H, $R=10^3\ \Omega$, $V_1(0)=10$ V. The initial time is zero and the total simulation time is 5×10^{-3} s with a step of 10^{-6} s. The nonsmooth problem is written as a Mixed Linear Complementarity Problem (MLCP). It has been solved using indifferently PATH and SEMISMOOTH methods. Obviously an enumerative solver is also convenient for a problem of such a size. The comparison is made with the SPICE simulator SMASH. In this case, the system is equivalent to an ODE with Lipschitz right-hand-side. The simulation with standard SPICE simulator together with low order schemes (Backward Euler) still works.

It is possible to modify the time-stepping algorithm described in Chap. 5 so that the time-step is adapted according to the local error. A practical error estimation is based on halved time-steps. Figure 7.10 presents the result of the step control mechanism using 10^{-3} for the relative and absolute tolerance. It is noteworthy that the time step is not decreased to pass through the diodes switching.

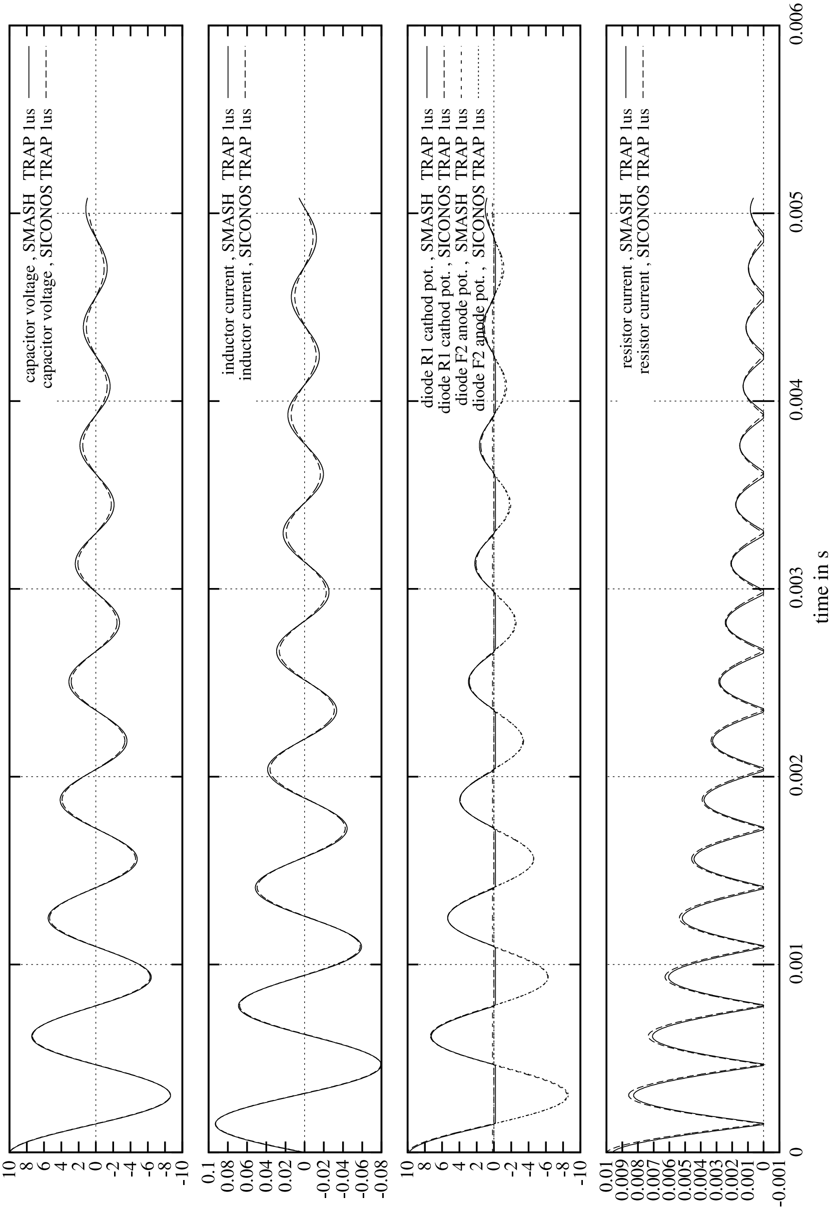


Fig. 7.9 Bridge wave rectifier simulation results

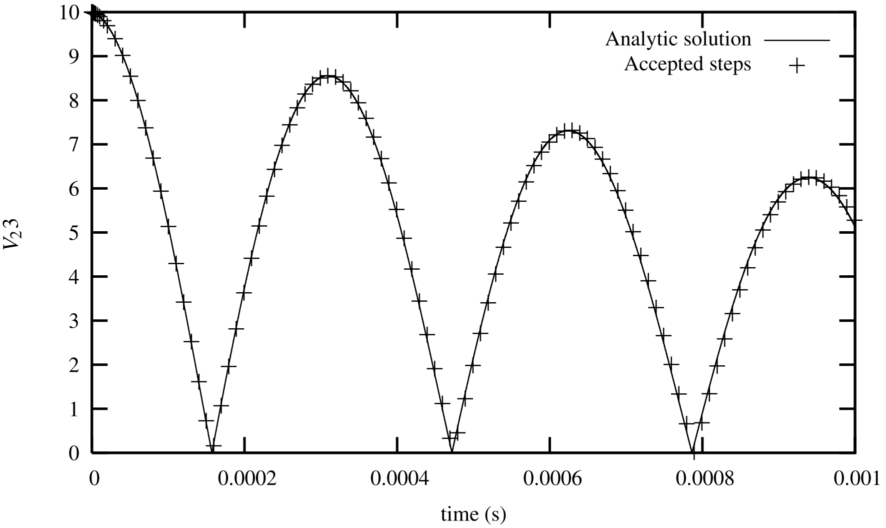
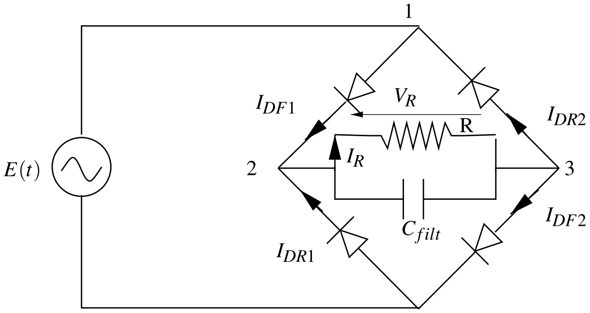


Fig. 7.10 Step size control using a 10^{-3} tolerance. 193 steps +95 rejected

Fig. 7.11 Filtered full wave rectifier



7.3 A Second Diode-Bridge Wave Rectifier

A little bit more complex example was simulated: a sinusoidal voltage supply providing energy to a resistor through a 4-diode bridge full-wave rectifier filtered with a capacitor (see Fig. 7.11). The automatic circuit equation formulation leads to the system (7.10) with the complementarity constraints in (7.9) and (7.8). The vector of unknown variables is $(U_C, V_1, V_2, V_3, I_E, I_{DF1}, I_{DF2}, I_{DR1}, I_{DR2})^T \in \mathbb{R}^9$. The potentials V_1, V_2, V_3 are the potentials at the points indicated in the figure. We obtain:

$$\begin{cases} -C\dot{U}_C(t) + \frac{V_2(t)-V_3(t)}{R} + I_{DR1}(t) + I_{DF1}(t) = 0, \\ -I_E(t) + I_{DR1}(t) - I_{DF2}(t) = 0, \\ I_{DF2}(t) - I_{DR1}(t) + I_{DR2}(t) - I_{DF1}(t) = 0, \\ V_1(t) = e(t), \\ U_C(t) = V_3(t) - V_2(t). \end{cases} \tag{7.10}$$

One may once again identify the dynamics of this circuit with the MLCS dynamics in (2.54), choosing for x the above vector of unknown variables. Figures 7.12, 7.13 and 7.14 show what happens with the SPICE algorithms when the time step is forced to a “high value” (here 10 μs): the SPICE simulator seems to converge but the results are erroneous while the nonsmooth approach provides accurate results. The SPICE package that has been used for these simulations is SMASH and the time-integration scheme is the trapezoidal rule. The results are taken from Denoyelle and Acary (2006).

Figures 7.15 and 7.16 show a comparison between SMASH results with a time step of 0.1 μs and SICONOS results with respectively time steps of 2 μs and 1 μs. The 2 μs results are already very close to SMASH ones. At 1 μs the differences are almost unnoticeable, whereas a factor of 10 is gained on the time step.

These results suggest that with a small number of stiff components in a circuit, the convergence of the Newton-Raphson algorithm is already impaired, even if several tricks were added in the SPICE software to help it. When the integration time period becomes too large, some diodes may be completely blocked at a time step and completely passing at the next time step. The SPICE algorithms are not designed to handle such a case: they need to step a sufficient number of times to cover properly all the switching period which is very short here.

On the contrary, the nonsmooth approach is able to compute a consistent solution with relatively far time steps, assuming that it exists and it is unique, which is true here.

Similar circuits like parallel resonant converters can be modeled and simulated in a similar way by SICONOS,[1] showing the wide range of applicability of the NSDS method in this field of electrical engineering.

7.4 The Ćuk Converter

In this section, we are interested in a special type of DC/DC power converter: the Ćuk converter (Middlebrook and Ćuk 1976). The circuit is described in Fig. 7.17. The converter is supplied by a constant voltage supply $E = 10$ V and loads a resistor $R = 50\ \Omega$. The switch is modeled by a linear MOS transistor described in Sect. 4.7.5 with two hyperplanes. The parameters of the nMOS model are $V_T = 5$ V and $K = 10$. The diode threshold voltage is 0.2 V. The capacitance are $C_1 = C_2 = 10$ μF and the inductances are $L_1 = L_2 = 250$ μH. There is no feedback on the regulation of the switch. The voltage V_G at the gate of nMOS model is given a periodic door

[1] See http://siconos.gforge.inria.fr/Examples/EMPowerConverter.html.

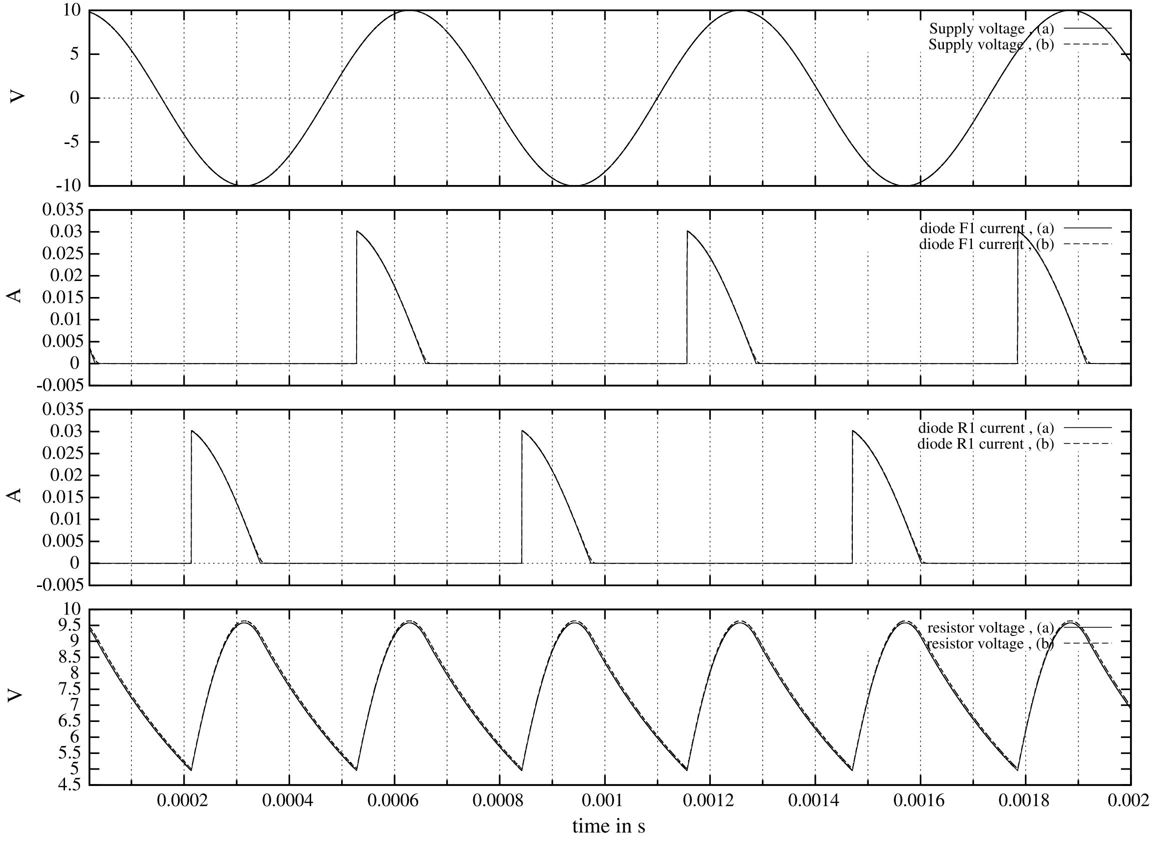


Fig. 7.12 SICONOS (**a**) and SMASH (**b**) simulations of the diode-bridge circuit, 0.1 μs time step

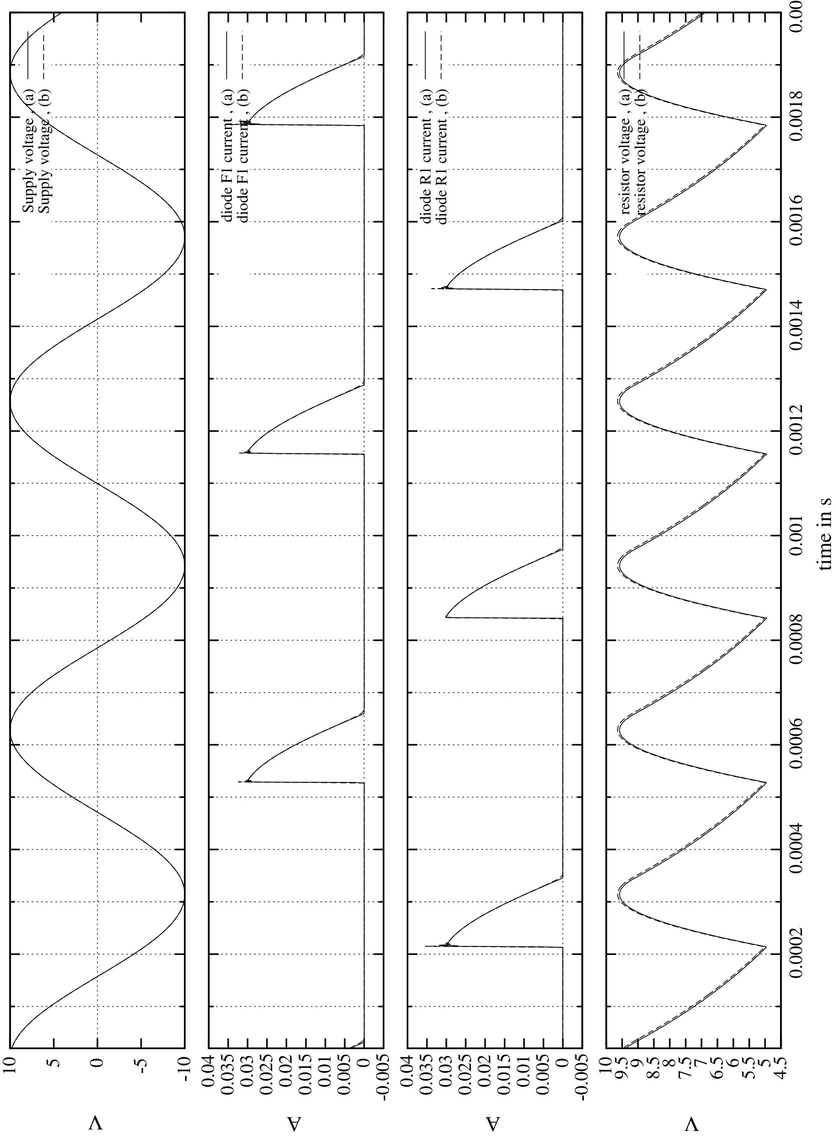


Fig. 7.13 SICONOS **(a)** and SMASH **(b)** simulations of the diode-bridge circuit, 1 μs time step

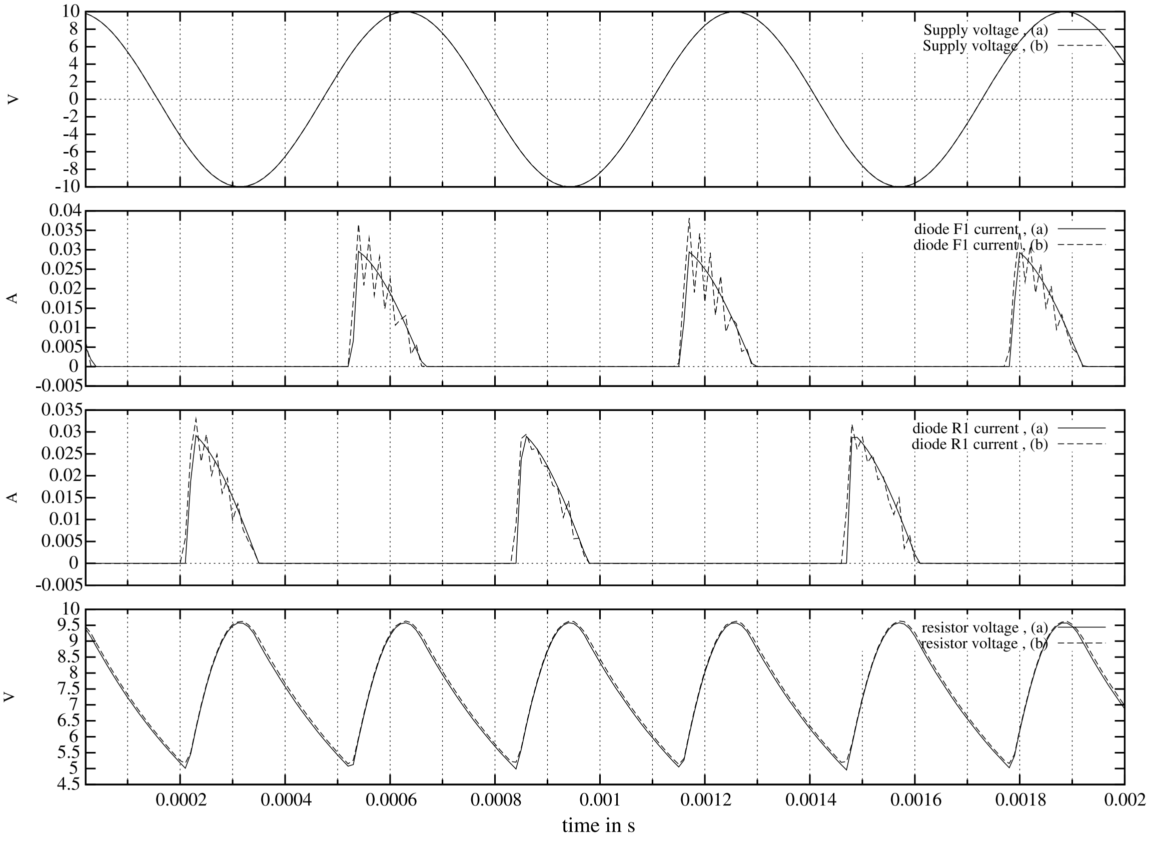


Fig. 7.14 SMASH (**b**) and SICONOS (**a**) simulations of the diode-bridge circuit, 10 μs time step

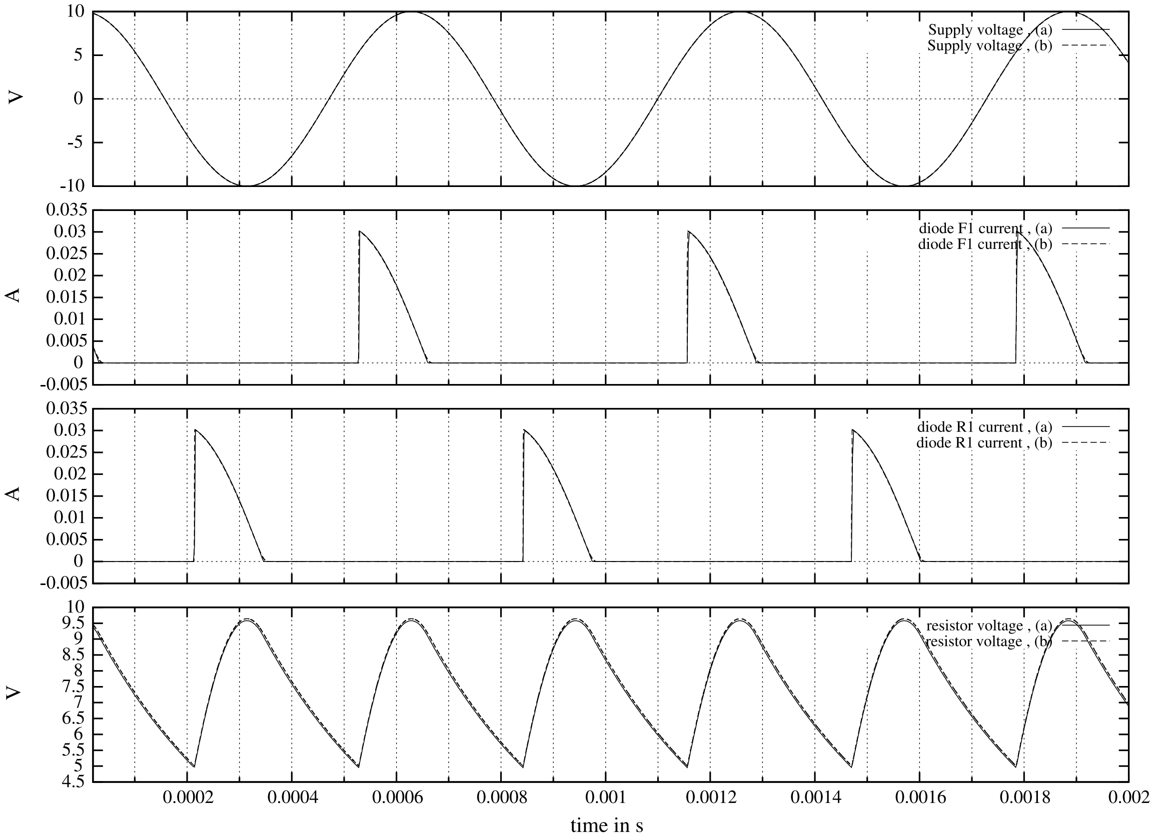


Fig. 7.15 Simulation results of the diode-bridge circuit. (**a**) SICONOS with $h = 2$ µs (**b**) SMASH with $h = 0.1$ µs

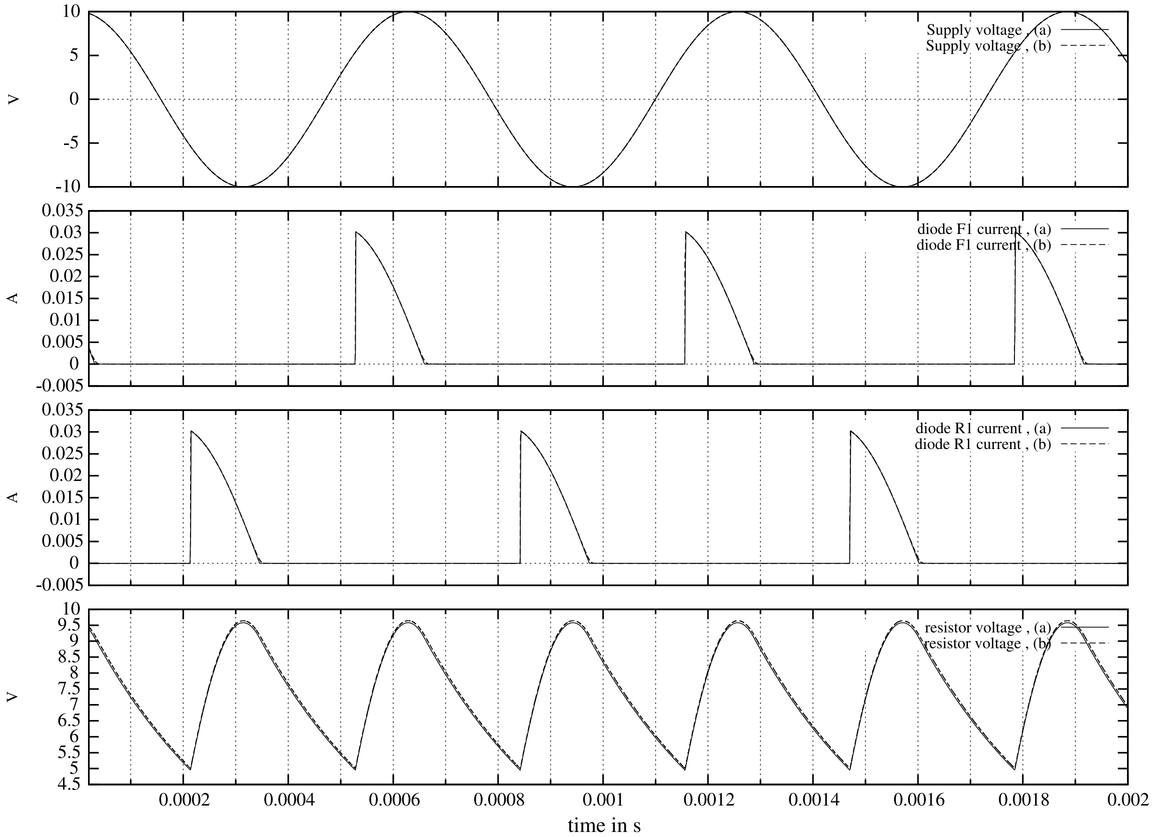


Fig. 7.16 Simulation results of the diode-bridge circuit. (**a**) SICONOS with $h = 1$ μs (**b**) SMASH with $h = 0.1$ μs

Fig. 7.17 The Ćuk converter

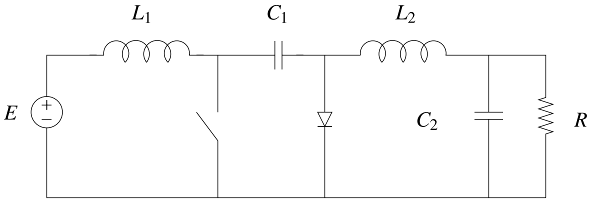
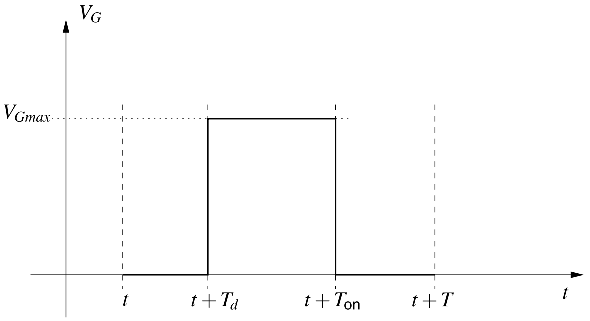


Fig. 7.18 Regulation function of the switch $V_G(t)$



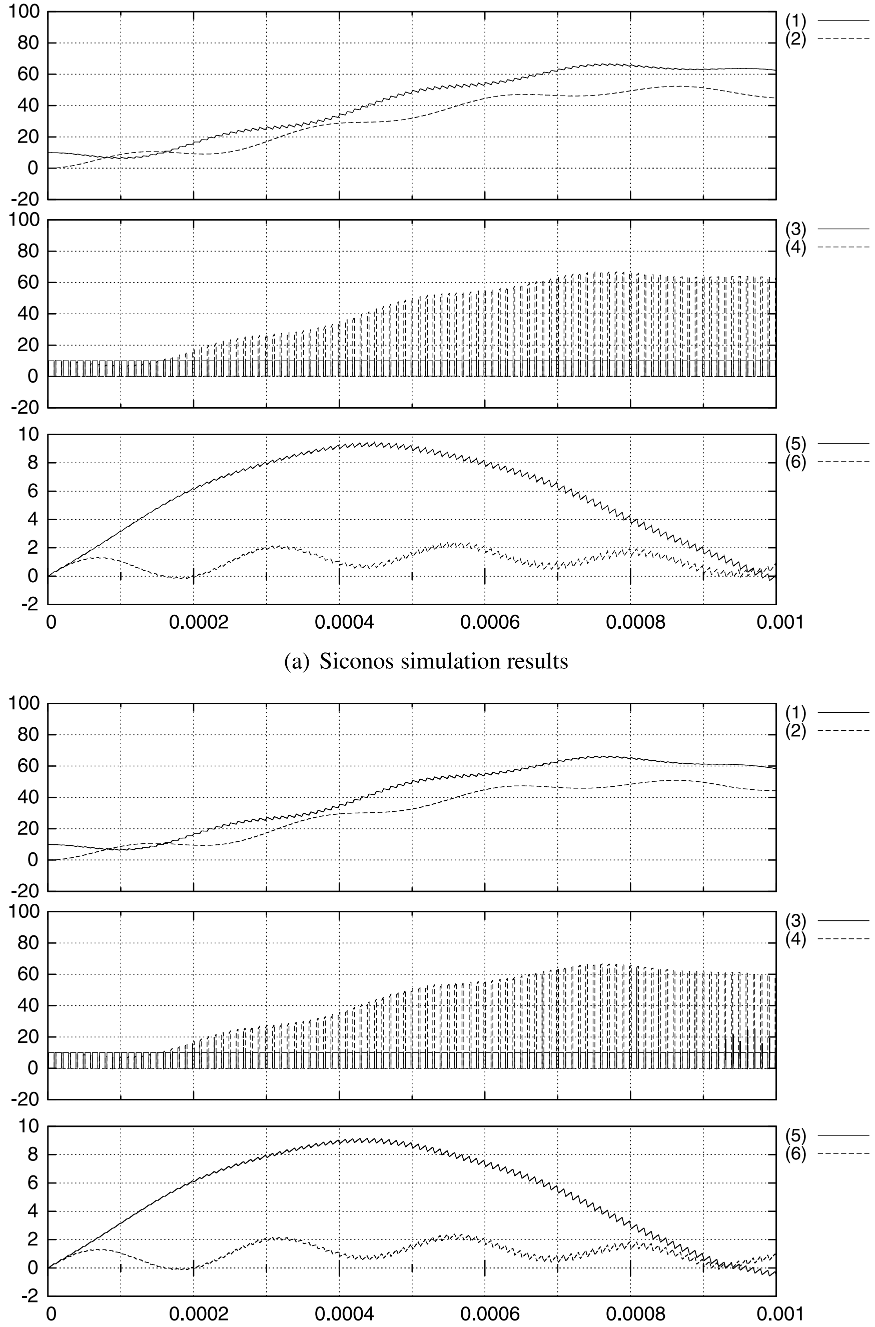
function of period T described in Fig. 7.18. In our numerical simulation, the parameters of the control of the switch are $V_{Gmax} = 10$, $T = 10$ µs, $T_d = 0.95$ µs, $T_{\rm on} = 7.5$ µs.

The results are displayed in Fig. 7.19 up to $= 0.001$ s and in Fig. 7.20 for the remaining simulation up to $t = 0.01$ s. The standard behaviour of the Ćuk converter is found with the SICONOS software as well as with the SPICE solver and the results are very similar. The main discrepancy between the nonsmooth approach and the SPICE approach is the choice of the time-step. The SICONOS simulations are performed with a time-step $h = 10^{-7}$ s and the SPICE simulations are performed with a time-step of 10^{-10} s. This last choice is mainly motivated by the numerical convergence problem of the Newton method when large time-steps are chosen. The gap between this two time-steps results in a gain of CPU time.

> The nonsmooth approach allows the use of larger time-step for the same accuracy in avoiding the numerical convergence problem of the Newton method when the electrical characteristics are stiff. This results in more robustness of the simulating process and in lower simulation times.

7.5 A Circuit Exhibiting Sliding Modes

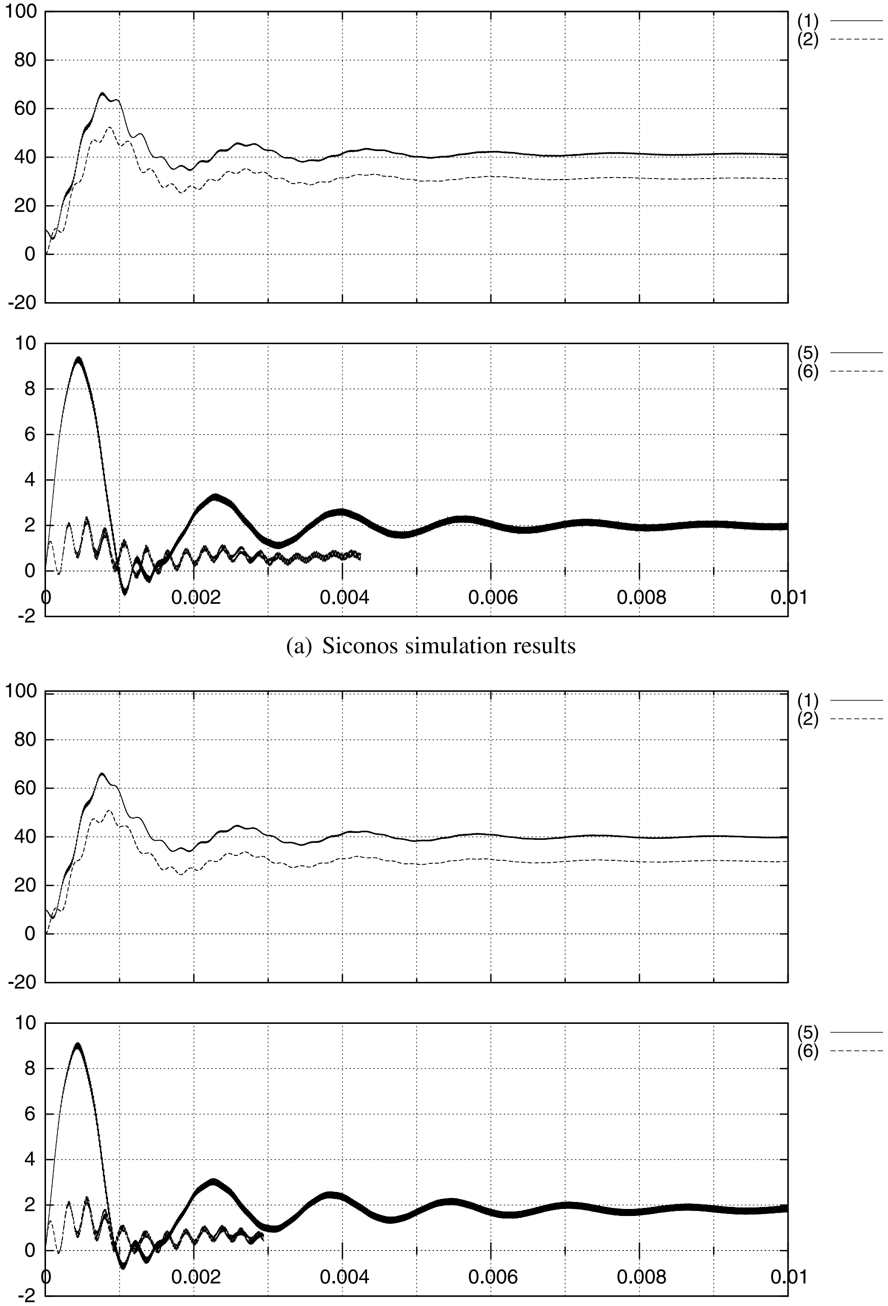
The goal of this section is to focus on the very interesting feature of the nonsmooth approach: the possibility to simulate consistently multivalued components and then



(a) Siconos simulation results

(b) SPICE simulation results

Fig. 7.19 Simulation values versus time (s) for the Ćuk Converter. (*1*) C_1 voltage, (*2*) C_1 voltage, (*3*) V_G gate voltage, (*4*) MOS switch drain voltage, (*5*) L_1 current and (*6*) L_2 current



(a) Siconos simulation results

(b) SPICE simulation results

Fig. 7.20 Simulation values versus time (s) for the Ćuk Converter. (*1*) C_1 voltage, (*2*) C_1 voltage, (*5*) L_1 current and (*6*) L_2 current

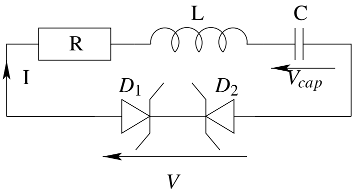


Fig. 7.21 RLC Zener-diodes circuit

coherently ideal components. The point is not to show that ideal components are better for the physical modeling accuracy, but rather that the regularization or standard hybrid approaches are not convenient for a high-level description and design.

> Our goal is to show that it is better to have a right simulation with an ideal model rather than a simulation which does not work correctly with a regularized model and pseudo-physics.

Let us consider a multivalued component, more precisely, we choose a multivalued behavior of a couple of Zener-diodes as the one which has been already introduced in Sect. 2.5.8. The goal is to outline the efficiency of the nonsmooth model by inclusions to handle such a behavior, even when a sliding mode occurs. In this context, the sliding mode has to be understood as a mode whose operating point of the component is inside the multivalued part of the graph in Fig. 2.24. The circuit is depicted in Fig. 7.21.

7.5.1 Models and Dynamical System

The choice of the vector of unknowns is $[I, V_{cap}]^T$ and it yields:

$$L\frac{dI(t)}{dt} = -RI(t) - V_{cap}(t) + V(t), \qquad C\frac{dV_{cap}(t)}{dt} = I(t). \tag{7.11}$$

7.5.1.1 Nonsmooth Model of Double Opposite Zener Diodes

For simplicity's sake, the two Zener diodes D_1 and D_2 in Fig. 7.21 are modeled as a single component. The behavior of the whole component is given by (see Fig. 2.24):

$$-I \in \mathbb{N}_{[-V_z,V_z]}(V), \tag{7.12}$$

which can be equivalently written as a complementarity problem (see the material in Sect. 2.4.6, especially (2.90)):

$$\begin{cases} V = \lambda_2 - V_z, \\ y_1 = V_z - V, \\ y_2 = I + \lambda_1. \end{cases} \quad \text{and} \quad 0 \leqslant \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \perp \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} \geqslant 0. \tag{7.13}$$

7.5.1.2 A Hybrid Single Valued Ideal Model in VERILOG/ELDO

Using a Netlist and a VERILOG description of the relation (7.14) to represent the couple of Zener diodes, a hybrid single valued model reads as:

$$V = \begin{cases} V_z & \text{if } I < 0, \\ 0 & \text{if } I = 0, \\ -V_z & \text{if } I > 0. \end{cases} \tag{7.14}$$

7.5.1.3 A Smooth Model Using Hyperbolic Tangent Function in ELDO

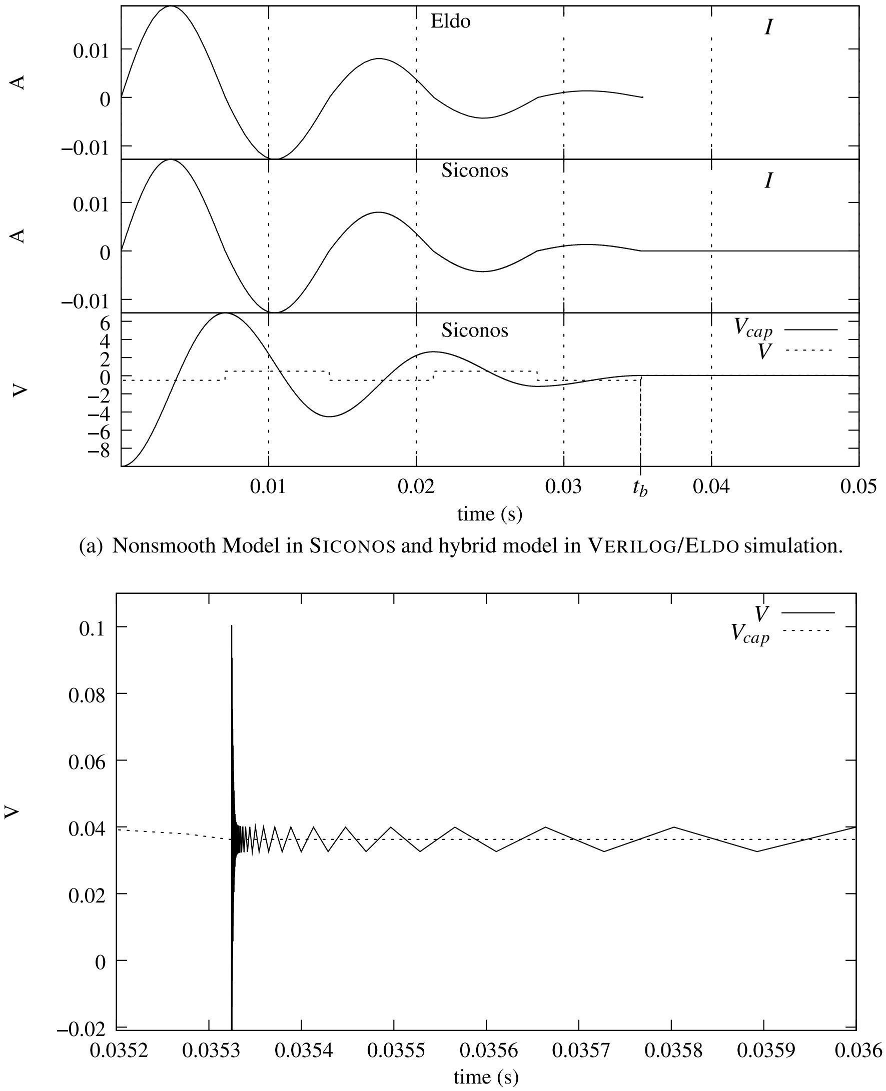
Another approach consists in using the hyperbolic tangent function to approximate the multivalued components D_1 and D_2, *i.e.* the vertical branch in Fig. 2.24 or in Fig. 1.8(b), is replaced by a stiff smooth curve. The relation between V and I is $V = -V_z \tanh(av)$. The coefficient a is chosen sufficiently large in order to neglect the influence of the regularization. We chose the value 10^5 to simulate the circuit using ELDO.

7.5.2 Simulation and Comparisons

The initial conditions are chosen as $I(0) = 0$ A, $V_{cap}(0) = -10$ V and the value of V_z is 0.5 V. The simulation using the LCP is successfully achieved with SICONOS. The result is shown in Fig. 7.22(a). This electrical circuit dissipates some energy, so V_{cap} oscillates with a decreasing amplitude up to a threshold value V_z. After the first event at $t = t_b$, the current i vanishes and the voltage through the capacitor V_{cap} is stabilized to a nonzero value, equal to v through the double Zener diodes component. Notice that this equilibrium point is located in the multivalued part of the characteristic.

Such a behaviour exactly corresponds to the dynamics (7.11–7.12), for which the segment $\{(I, V_{cap}) \mid I = 0, -V_z \leqslant V_{cap} \leqslant V_z\}$ is an attractive sliding surface, attained in a finite time. When $I = 0$, it follows from (7.12) that $V_{cap} \in [-V_z, V_z]$. Figure 7.22(a) shows also the ELDO simulation using the Netlist and the VERILOG relation (7.14). In this case the simulation is correctly done until t_b. At the first event at time t_b, the simulation cannot be continued because the equilibrium point of the circuit is not handled by the model.

The simulation using the hyperbolic function has been made using ELDO. We focus our attention on the difference due to the regularization of the multivalued



(a) Nonsmooth Model in SICONOS and hybrid model in VERILOG/ELDO simulation.

(b) Regularized model in ELDO. Zoom in the neighborhood of the switching time.

Fig. 7.22 Simulation of the circuit with a sliding mode

model. Figure 7.22(b) zooms on the moment where the current vanishes. At this instant, the circuit is equivalent to an RLC circuit where the value of R is the coefficient of the tangent to the hyperbolic curve a. Note that using a coefficient a larger than 10^5 leads to an artificial v oscillation around the value of V_{cap}. The conclusion is that we cannot expect to observe the convergence toward an ideal behavior with such a regularization.

The NSDS method with Moreau’s implicit time-stepping scheme allows one to perfectly simulate sliding mode phenomena. This is not the case for the other approaches.

More details on the simulation of sliding mode systems may be found in Acary and Brogliato (2010).

Chapter 8
Buck and Delta-Sigma Converters

This chapter is dedicated to the numerical simulation of the buck and the delta-sigma converters. Comparisons between the results obtained with the NSDS SICONOS approach and other approaches are presented.

8.1 The Buck Converter with Load Resistor

The studied buck converter with a load resistor is depicted in Fig. 8.1. The electronic components are modeled with either linear or piecewise-linear or set-valued relations, yielding a nonsmooth dynamical system of the linear-time-invariant complementarity systems class. The features of the models are given thereafter:

- Power MOSFETS pMOS/nMOS: they are described as an assembly of a piecewise-linear current source $I_{DS} = f(V_{GS}, V_{DS})$ and the intrinsic diode (DpMOS and DnMOS) with an ideal characteristic. The capacitors were not taken into account. The diodes residual voltage is 1 V. The MOSFETs transconductance KP was set to 10 A V^{-2} and their threshold voltage to respectively $V_T = -2$ V for the pMOS and $V_T = 2$ V for the nMOS. One can notice that the sum of their absolute values largely exceeds the supply voltage $V_I = 3$ V, thus providing non-overlapping conduction times. The other physical parameters are chosen as follows: $\mu = 750\ \mathrm{cm}^2\,\mathrm{V}^{-1}\,\mathrm{s}^{-1}$ for a nMOS and $\mu = 250\ \mathrm{cm}^2\,\mathrm{V}^{-1}\,\mathrm{s}^{-1}$ for a pMOS, $\varepsilon_{Ox} = \varepsilon_{r\ \mathrm{SiO}_2} \times \varepsilon_0$ with $\varepsilon_{r\ \mathrm{SiO}_2} \approx 3.9$, $t_{OX} \approx 4$ nm, $W = 130$ nm, $L = 180$ nm.

  The piecewise-linear model uses 6 segments given by the following data: $c_1 = 0.09$, $c_2 = 0.2238$, $c_3 = 0.4666$, $c_4 = 1.1605$, $c_5 = 2.8863$, $a_1 = 0$, $a_2 = 0.1$, $a_3 = 0.2487$, $a_4 = 0.6182$, $a_5 = 1.5383$. The relative error between $f(\cdot)$ and $f_{\mathsf{pwl}}(\cdot)$ (see Sect. 4.7.5) is kept below 0.1 for $0.1 \leqslant x < 3.82$. The absolute error is less than 2×10^{-3} for $0 \leqslant x < 0.1$ and 0 for negative x. In practice, the values of V_G, V_S, V_D, V_T in logic integrated circuits allow a good approximation of $f(\cdot)$ by $f_{\mathsf{pwl}}(\cdot)$.
- Compensator amplifier: It is modeled as a 1×10^5 gain and an output low-pass filter with a cutoff frequency of 30 MHz that is $R_p = 1\ \Omega$ and $C_p = 5.3$ nF.

V. Acary et al., *Nonsmooth Modeling and Simulation for Switched Circuits*,
Lecture Notes in Electrical Engineering 69,
DOI 10.1007/978-90-481-9681-4_8, © Springer Science+Business Media B.V. 2011

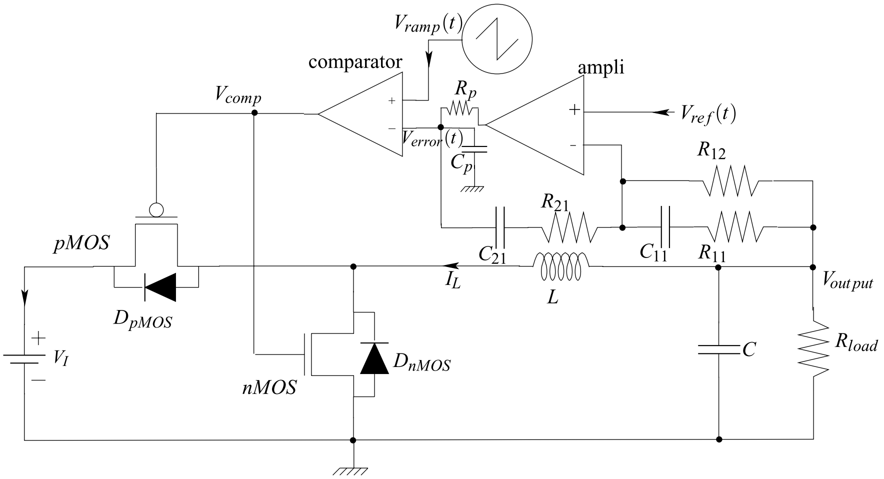


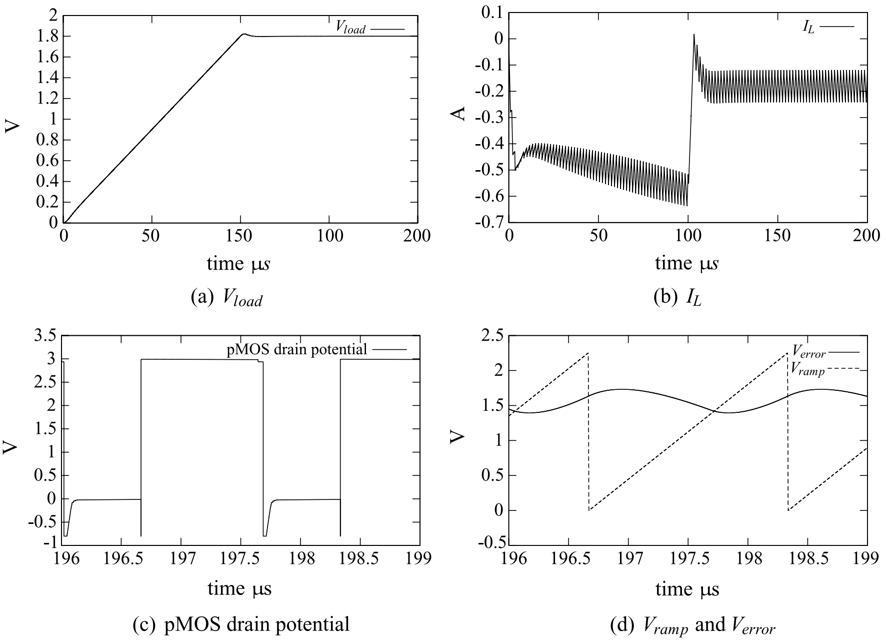
Fig. 8.1 Buck converter

- Comparator: It is modeled as a piecewise-linear function whose value is 0 if $x < -0.15$ V and 3 if $x > 0.15$ V.
- Ramp voltage: The frequency is 600 kH and the bounds are 0 and $0.75V_I = 2.25$ V. The rise time is 1.655 ns and the fall time is 10 ns.
- Standard values for other components: $V_I = 3$ V, $L = 10$ μH, $C = 22$ μF, $R_{load} = 10\ \Omega$, $R_{11} = 15.58\ \mathrm{k}\Omega$, $R_{12} = 227.8\ \mathrm{k}\Omega$, $R_{21} = 5.613\ \mathrm{M}\Omega$, $C_{11} = 20$ pF, $C_{21} = 1.9$ pF.
- Values exhibiting a sliding mode: $L = 4$ μH, $C = 10$ μF, $R_{11} = 10\ \mathrm{k}\Omega$, $R_{21} = 8\ \mathrm{M}\Omega$, $C_{11} = 10$ pF.

The reference voltage V_{ref} rises from 0 to 1.8 V in 0.1 ms at the beginning of the simulation. The output voltage V_{output} is regulated to track the reference voltage V_{ref} when V_I or V_{ref} or the load current vary. The error voltage V_{error} is a filtered value of the difference between V_{output} and V_{ref}. This voltage signal is converted into a time length thanks to a comparison with the periodic ramp signal. The comparator drives the pMOS transistor which in turn provides more or less charge to the output depending on the error level. The operation of a buck converter involves both a relatively slow dynamics when the switching elements (MOS and diodes) are keeping their conducting state, and a fast dynamics when the states change. The orders of magnitude are 50 ps for some switching details, 1 μs for a slow variation period and 100 μs at least for a settling period of the whole circuit requiring a simulation.

8.1.1 Dynamical Equations

The nonsmooth DAE has been generated using the automatic circuit equation formulation described in Chap. 6. It leads to a dynamical system with 25 variables



(a) V_{load}

(b) I_L

(c) pMOS drain potential

(d) V_{ramp} and V_{error}

Fig. 8.2 SICONOS buck converter simulation using standard parameters

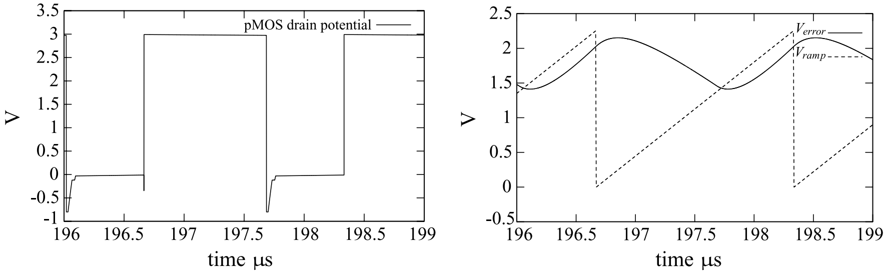
coupled to an inclusion rule. The dimension of the inclusion rule is 24. The size of the x vector is 5, composed of the capacitor voltages, and of the inductor current.

8.1.2 *Numerical Results with* SICONOS

The start-up of the converter was simulated thanks to SICONOS. As initial conditions, all state variables are zeroed. The detailed analysis of the switching events requires to use a time step as small as 50 ps. The simulations are carried with a fixed time-step, 4×10^6 steps are then computed for the 200 µs long settling of the output voltage. The OSNSP solvers which are used are PATH with a convergence tolerance of 10^{-7}, and a semi-smooth Newton method based on the Fischer-Bursmeister reformulation (DeLuca et al. 1996), implemented in SICONOS and using a convergence tolerance of 10^{-12}. The overall result is shown in Fig. 8.2.

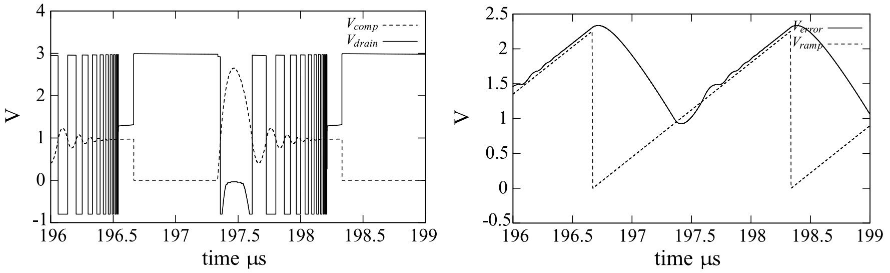
Simulation Time The CPU time required to achieve the simulation is 60 s on a Pentium 4 processor clocked at 3 GHz. It includes 19 s in the MLCP solvers, 40 s in matrices products. The time to export the resulting data is not included.

- Figure 8.2(a) is the output potential, following the ramp V_{ref}.
- Figure 8.2(b) is the current through the inductor. Until 0.0001 s, I_L is loading the capacitor C. After 0.0001 s, I_L has to keep the capacitor charge constant.



(a) pMOS drain potential (b) V_{ramp} and V_{error}

Fig. 8.3 SICONOS buck converter simulation using $L = 7$ µH, $C = 15$ µF, $R_{11} = 12$ kΩ, $R_{21} = 6$ MΩ, $C_{11} = 15$ pF



(a) V_{comp} and V_{drain} (b) V_{ramp} and V_{error}

Fig. 8.4 SICONOS buck converter simulation using sliding mode parameters

- Figures 8.2(c) and 8.3(a) zoom on the pMOS drain potential with standard parameters.
- Figures 8.2(d) and 8.3(b) zoom on the V_{error} and V_{ramp} voltages.
- Figure 8.4(a) using sliding mode parameters,[1] shows the stabilization of the comparator output to an unsaturated value. It also shows the stabilization of the current through the pMOS allowing the V_{error} signal to follow the V_{ramp} signal.
- Figure 8.4(b) using sliding mode parameters, shows the V_{error} and V_{ramp} voltages.

8.1.2.1 Focus on the pMOS Component During the Sliding Mode

The Fig. 8.5 focuses on the pMOS component during the interval from 196 µs to 199 µs. The goal is to show the stabilization of the pMOS current during the sliding mode. From Fig. 8.5 two behaviors can be distinguished. The first one happens during the interval [197.350 µs, 198.055 µs], when the transistor *pMOS* oscillates between two states:

[1] See Sect. 2.4.4 for a definition of sliding surfaces in switching systems.

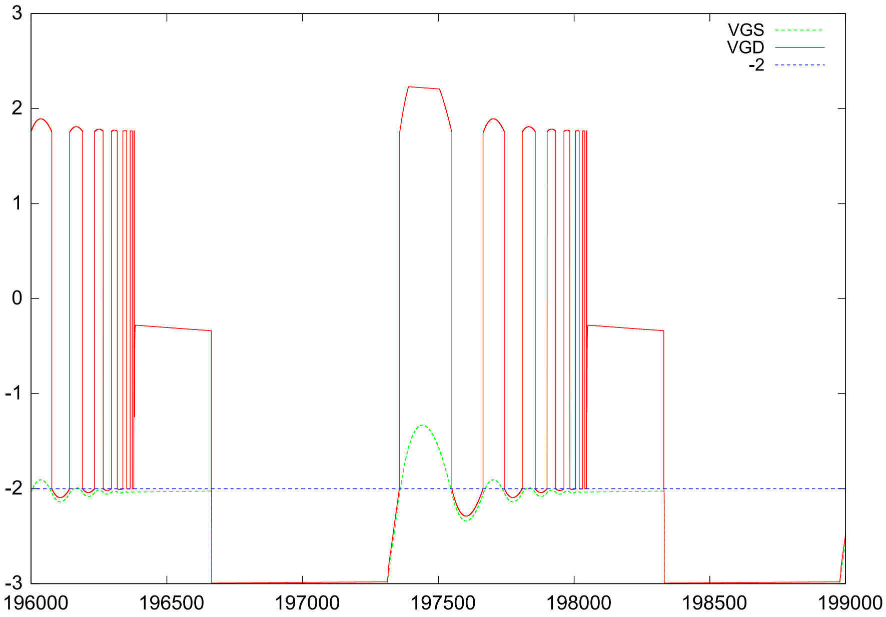


Fig. 8.5 Focus on the pMOS component

- The comparator output tension is high. V_{GD} and V_{GS} are below $VT0$, so the current I_{SD} is positive, consequently I_L grown. It can be said that the transistor is ON.
- The comparator output tension is low. The transistor is locked, the current I_{SD} is null and the current through the diode D_{nMOS} is positive. It can be said that the transistor is OFF, the system is braking.

The second behavior is observable during the interval [198.055 μs, 198.32 μs]. It is the stabilization of V_{GS} below $VT0$, implying a current through the transistor P that is equal to the current dissipated through the loading resistor. Using a multivalued comparator leads to the stabilization of the comparator output to an intermediate value.

8.1.2.2 Robustness of the NSDS Method

The simulation has been tested with many parameter values, see Figs. 8.2, 8.3 and 8.4. The robustness of the nonsmooth modeling and solving algorithms enables one to perform with the same CPU time the simulation of such cases. This is not the case with the SPICE algorithms, see Sect. 8.1.3.2.

8.1.2.3 Simulation Using a Nonlinear MOS Model

Similar results are obtained with the nonlinear MOS model presented in Sect. 4.7.5. They are not detailed here because no important discrepancy can be noticed.

8.1.3 Comparisons and Discussions

In this section the results obtained with different software packages relying on various modeling approaches, are compared.

8.1.3.1 Simulation with SPICE

The simulation of the buck converter was done with several versions of SPICE (the open-source package NGSPICE and ELDO) and two kinds of MOS models:

The MOS Level 3 Model This model takes more physical effects into account than the piecewise-linear model used in SICONOS simulations, in particular the voltage-dependent capacitors. It is an important issue since these varying capacitances cause some convergence problems when node 2 switches between V_I and the ground. Adding a small capacitor of a few picoFarad between this node and the ground helps to solve the problem but may yield artifacts (spikes) on the current of the V_I alim and the MOS transistors.

An nMOS Simplified Model (Sah Model) with fixed capacitors and a quadratic static characteristic:

$$I_{DS} = \max(0, V_{GS} - Vt_N)^2 - \max(0, V_{GD} - Vt_N)^2.$$

This model is very close to the piecewise-linear model used in SICONOS simulations. The implementation in Netlists was done thanks to voltage-dependent current sources that are very likely not compiled by the various SPICE simulators tested. Thus the measured CPU time is increased with respect to a compiled version. An estimation of the CPU time with a compiled simplified model may be given by multiplying the MOS level 3 CPU time by the ratio of the Newton-Raphson iterations required respectively during the simulations with each model. An additional correction should be done to reflect that the computation of the Jacobian matrix entries linked to a compiled simplified model would require less time than with a MOS level 3 model. Even if the SPICE simulation includes other operations, the Jacobian matrix loading time is indeed known to be generally predominant.

- Power MOSFETS intrinsic diodes are modeled by the classical Shockley equation with an emission coefficient $N = 1$:

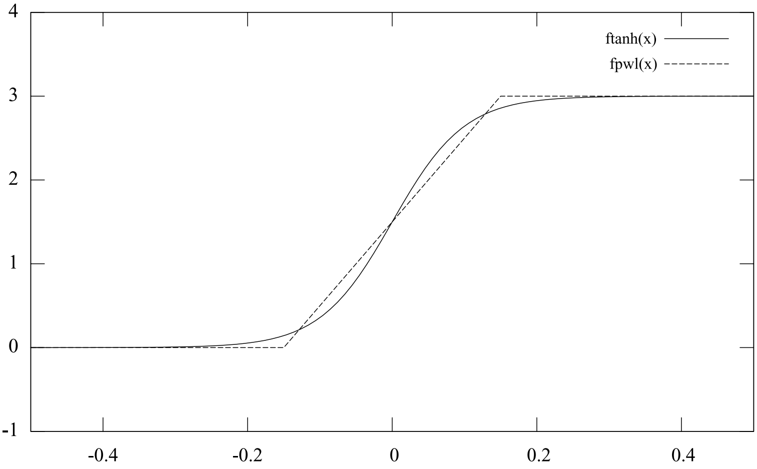


Fig. 8.6 Comparison of piecewise-linear and Spice (tanh based) comparator models

$$I = I_S(e^{\frac{q.V}{N.k.T}} - 1) \quad \text{when } V > -5N\frac{k.T}{q},$$

$$I = -I_S \quad \text{when } V < -5N\frac{k.T}{q},$$

with V and I the voltage and the current through the diode, I_S the saturation current, a default value 10^{-14} A, electron q charge 1.6×10^{-19} C, k Boltzmann constant 1.38×10^{-23} J K^{-1}, T temperature in K and N emission coefficient.

- The comparator is modeled as a nonlinear voltage controlled-voltage-source defined as $V_{out} = 1.5(\tanh(10V_{in}) + 1)$. Thus the 3-segment characteristic used as the nonsmooth model of Fig. 4.10(b) is regularized to help the convergence of Spice (see a comparison of the piecewise-linear comparator as used in Siconos simulations with the Spice one in Fig. 8.6).

The power supply V_I is raised from 0 in 50 ns at the beginning to help the convergence.[2] The Spice tolerance values used are 1 nA for currents, 1 µV for voltages and 0.00075 for relative differences. The maximum number of Newton-Raphson iterations is set to 100 (the default values are 10 for Ngspice and 13 for Eldo).

Usually, Spice simulators integrate with a time step adjusted according to different strategies based on an estimation of the local truncation error (LTE) or the number of Newton-Raphson iterations required by previous steps. Since Siconos simulations were carried with a fixed time step of 50 ps, simulators were forced to use this value as a maximum. Even when Spice simulators use a fixed time step, they may compute LTE to assess a solution found by the Newton-Raphson algo-

[2] This is not required with the Siconos algorithms that find a consistent initial solution from scratch.

Table 8.1 Numerical comparisons on the buck converter example

SIMULATOR	MODEL	Number of Newton iterations	CPU time (s)
Standard compensator values			
Ngspice	simple	8024814	632
Ngspice	level 3	failed	
Eldo	simple	4547579	388
Eldo	level 3	4554452	356
Siconos	LCP	–	60
Sliding mode compensator values			
Ngspice	simple	8070324	638
Ngspice	level 3	8669053	385
Eldo	simple	5861226	438
Eldo	level 3	5888994	367
Siconos	LCP	–	60

rithm. This computation of LTE was disabled because it could impair the performance of Spice with respect to Siconos.[3]

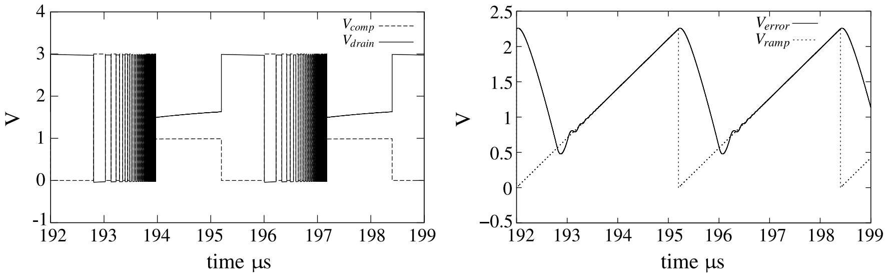
8.1.3.2 Simulation Comparisons

The Table 8.1 displays the results with the standard and the sliding mode values of compensator components. An estimation of the CPU time with a compiled simplified model is added.

These results have to be compared to the 60 s CPU time achieved with the NSDS method with Siconos. Depending on the model and the Spice simulator, the CPU time is from 5.9 to 10.6 larger than with Siconos. Moreover, it was necessary to add a parasitic capacitor on the connection between the pMOS and nMOS transistors to allow the convergence of the Ngspice simulator with the MOS level 3 model.

All the Siconos simulations presented in this chapter have been obtained in one-shot from the dynamical equations automatically generated from the Netlist, without any further parameter tuning.

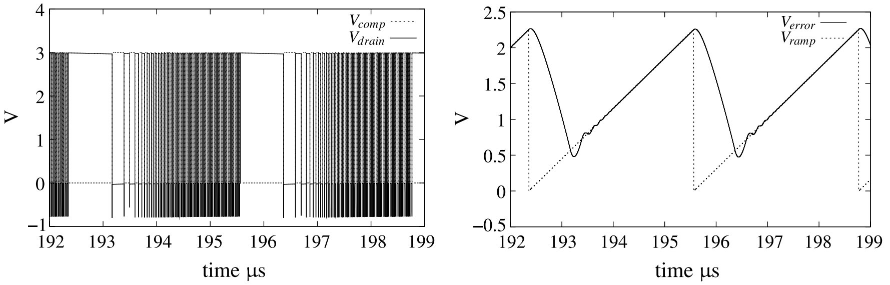
[3] For Ngspice, it implied a slight modification of the source code since no standard option is provided to do it.



(a) V_{comp} and V_{drain}

(b) V_{ramp} and V_{error}

Fig. 8.7 SICONOS buck converter simulation using sliding mode parameters and multivalued comparator



(a) V_{comp} and V_{drain}

(b) V_{ramp} and V_{error}

Fig. 8.8 ELDO buck converter simulation using sliding mode parameters and $V_{out} = 1.5$ $(\tanh(10000V_{in}) + 1)$ for the comparator

8.1.3.3 Sliding Mode Using a Multivalued Comparator

This section focuses on the simulation with sliding parameters and using a multivalued model for the comparator, *i.e.* a model whose graph possesses a vertical branch at zero, like the relay multifunction. The rise time of the ramp voltage has been increased to 3.2 ns. The model used in SICONOS consists in setting the ε gap to 0 in the model depicted in Fig. 4.10(b). Figure 8.7 shows the SICONOS simulation using a fully implicit time-stepping method. It could be noted that the comparator output is stabilized to an unsaturated value, corresponding to intermediate value in the multivalued part of the characteristic. The simulation using ELDO has been done using the model $V_{out} = 1.5(\tanh(10000V_{in}) + 1)$ for the comparator. The MOS level 3 leads to "Newton no-convergence" messages, so the MOS Sah model has been used to run the simulation displayed in Fig. 8.8. It is noteworthy that this simulation with ELDO does not handle the stabilization of the comparator output on the sliding surface. Indeed despite this is not visible in Fig. 8.8(b), the trajectory keeps oscillating around the attractive surface, as witnessed by the values observed in Fig. 8.8(a).

8.1.3.4 Simulation with PLECS

PLECS is a SIMULINK/MATLAB toolbox dedicated to the simulation of power electronics circuits.[4] The electronic circuits are modeled as hybrid systems: at each instant, the circuit is described according to one of a set of topologies specified by the ON or OFF state of ideal switches (diodes, transistors …). A topology is valid when the computed value of some variables (for instance a diode current or voltage) is kept within some bounds. When a topology is no more valid, a new topology has to be found as well as the possible jump of variables linked to this topology switching. Even if this approach targets the same kind of systems as the NSDS method, its models and algorithms differ mainly in:

Hybrid systems approach	**NSDS approach**
Each topology is described by a separate set of equations.	A single set of equations and constraints describes the whole system.
Checking the topology switching conditions is critical and may be computationally expensive.	There is no topology switching: constraints are met at each time step.
Determining the new topology and the new state value after a topology switch is not obvious. It may be quick if it can benefit from prior knowledge about the circuit's operation but may also involve heavy computations to check all possible transitions if no rule is available.	At each time step, the new values are computed to meet equations and constraints thanks to proper time integration schemes and one step problem solving based on optimization algorithms

The switch models (diodes and transistors) available in the PLECS toolbox are ideal: the transistors are supposed to be controlled by a boolean signal forcing a conducting or blocking state. The power nMOS and pMOS are controlled by opposite signals issued from the feedback loop, thus there is no flyback conduction by the nMOS diode (see the description of the PLECS circuit and the SIMULINK model in Figs. 8.9 and 8.10).

The CPU time required to achieve the simulation of 200 µs varies between 2 min 15 s and 6 min 50 s on a Pentium 4 clocked at 3 GHz, depending on the values of the resistors, capacitors and inductor. This should be compared to the 24 s of the SICONOS simulation, **obtained independently from these components values**. This demonstrates the robustness and efficiency of the time-stepping scheme and the one step solving algorithms of SICONOS. Figures 8.11 and 8.12 show the results when the standard parameters are used: $L = 10$ µH, $C = 22$ µF, $R_{11} = 15.58$ kΩ, $R_{12} = 227.8$ kΩ, $R_{21} = 5.613$ MΩ, $C_{11} = 20$ pF, $C_{21} = 1.9$ pF.

Remark 8.1 On both Figs. 8.7(a) and 8.4(b) it is seen that the sliding surface is attained in finite time after an accumulation of switches. This is a classical phenomenon in nonsmooth dynamical systems, see Filippov's example in Acary and Brogliato (2010).

[4] See http://www.plexim.com.

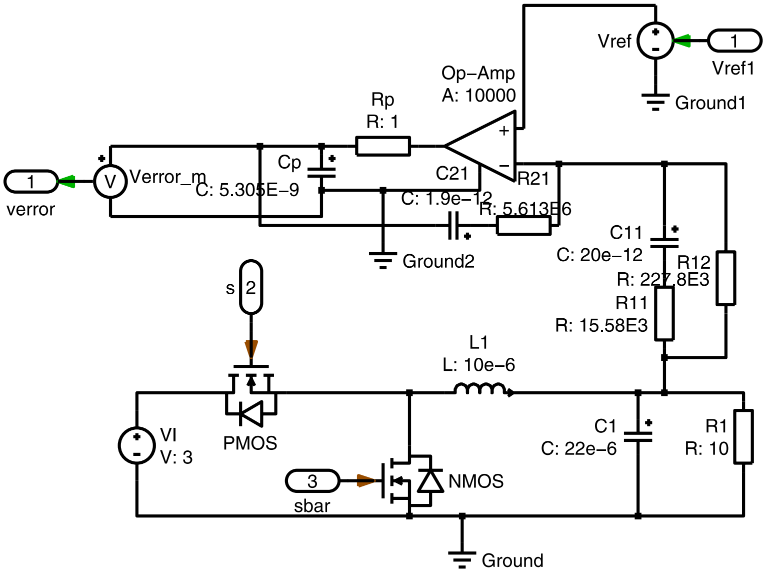


Fig. 8.9 PLECS circuit part of the buck converter with a load resistor

8.1.3.5 Global Error Evaluation

It consists in computing the experimental order of the simulation. Since the analytic solution is unknown, the reference trajectory is a simulation using a very small time step. In order to compute the order of SICONOS and ELDO, the following simulations have been performed:

Time step (ps)	10	20	40	80	160	320	640	1200
ELDO Global error (nA)	0.15	0.443	0.966	2.1	4.5	8.2	20	34
SICONOS Global error (nA)	0.246	0.513	1.026	2.2	4.7	9.2	21	40
ELDO Log2 (global error)	−32.6	−31	−30	−28.8	−27.7	−26.8	−25.6	−24.8
SICONOS Log2 (global error)	−31.9	−30.8	−29.8	−28.7	−27.7	−26.7	−25.5	−24.5

A 5 ps trajectory is the reference used to compute the global error. Figure 8.13 shows the log2 (Global error) as a function of log2 (time step): the experimental order of the global error is 1. So the experimental order of the method used at each step is 2.

8.1.3.6 Global Error Evaluation as a Function of the Number of Hyperplanes

In this section, we focus on the incidence of the number of hyperplanes used to approximate the square function of the transistor. It consists in comparing the simulation output using the model of transistor (4.52) to the simulation output using the model of transistor (4.45). To achieve this goal, the simulation has been done

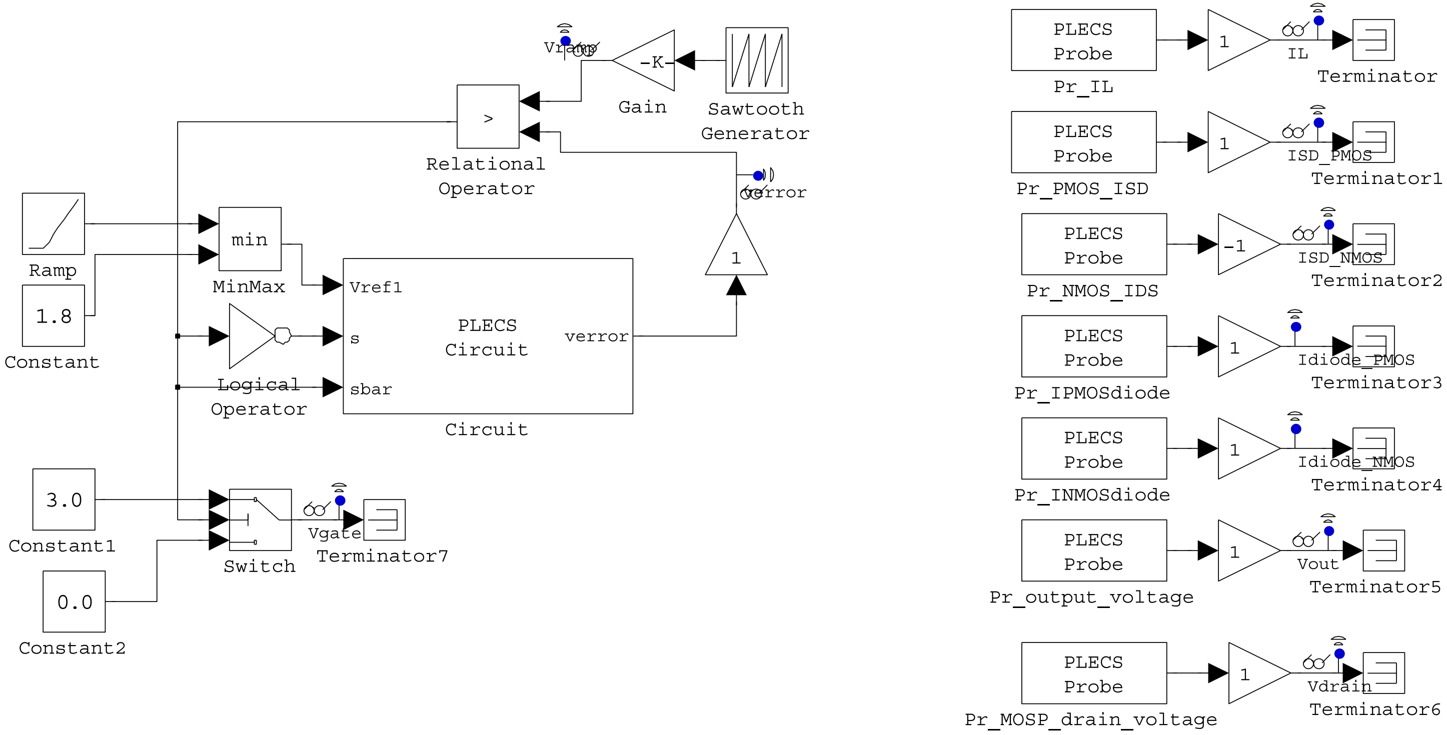
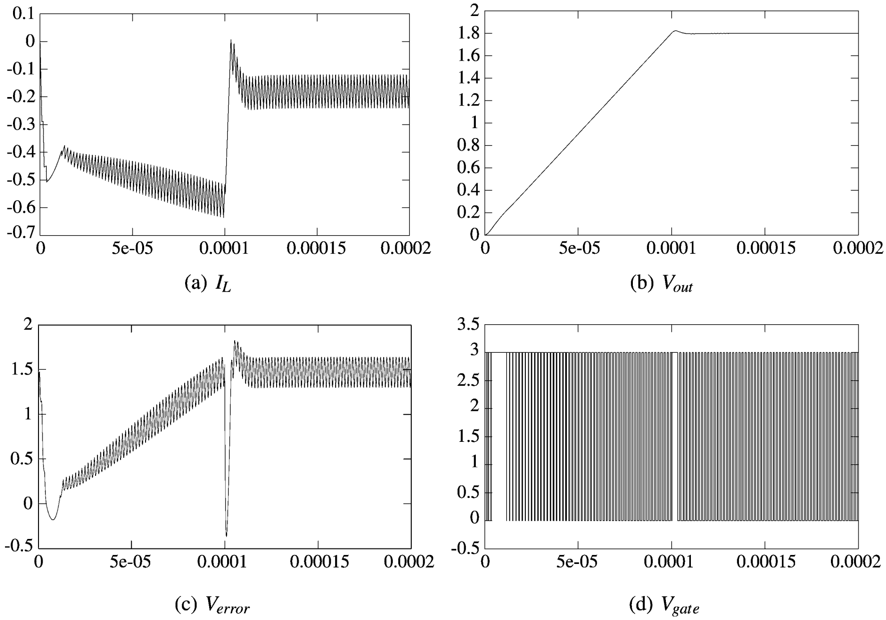


Fig. 8.10 SIMULINK model of the buck converter with a load resistor



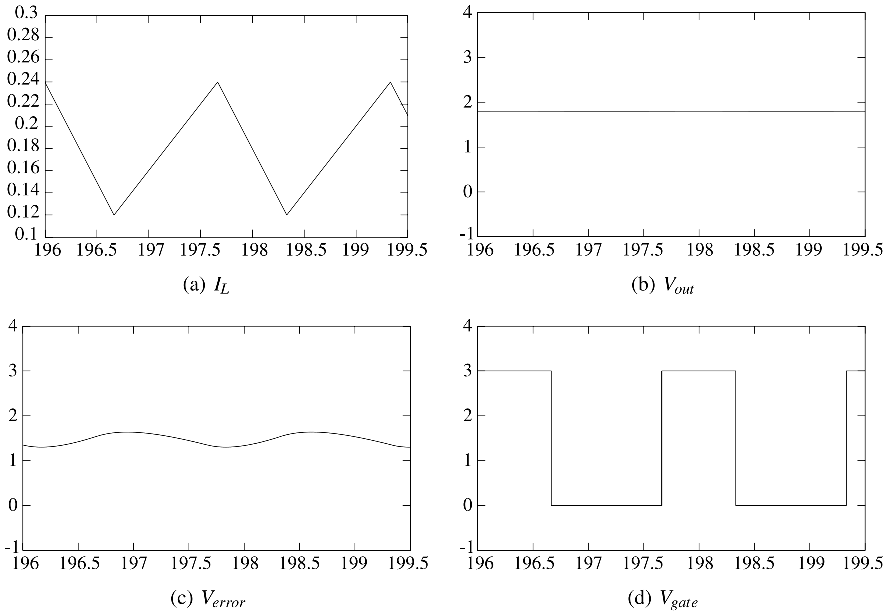
(a) I_L

(b) V_{out}

(c) V_{error}

(d) V_{gate}

Fig. 8.11 PLECS buck converter simulation using standard parameters. Time in seconds



(a) I_L

(b) V_{out}

(c) V_{error}

(d) V_{gate}

Fig. 8.12 PLECS buck converter simulation using standard parameters, zoom view of steady state. Time in μs

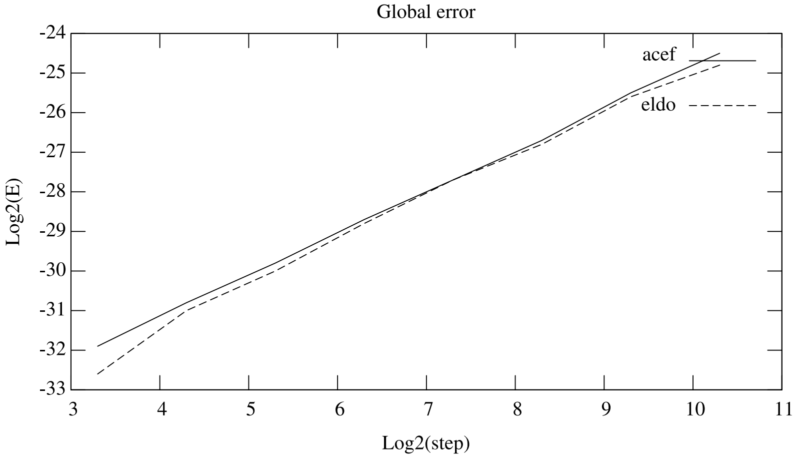


Fig. 8.13 Global error

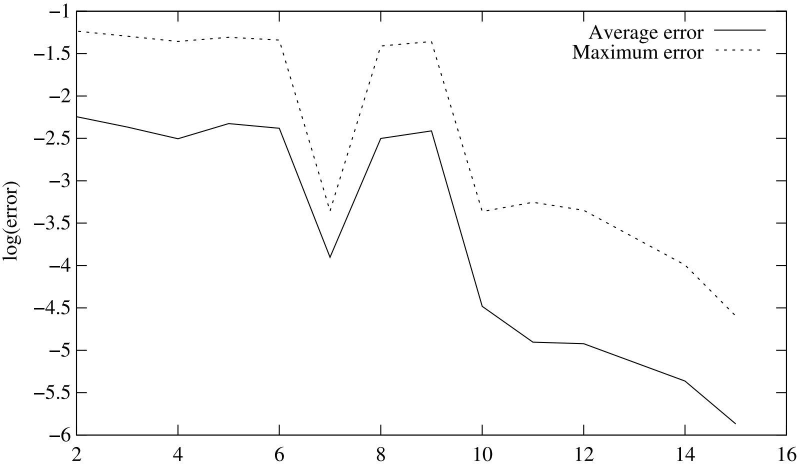


Fig. 8.14 Global error evaluation as a function of the number of hyperplanes

using a transistor model composed of 2 until 15 hyperplanes. For each step of the simulation, the state vector error has been computed. Figure 8.14 shows the average error and the maximum error. It is noteworthy that the obtained curve is not monotone decreasing. It comes from the fact that more the number of hyperplanes is important, more the maximal error to the square model is small, but for any value, the approximation is not necessarily better. Nevertheless, Fig. 8.14 shows that the errors become small using a large number of hyperplanes.

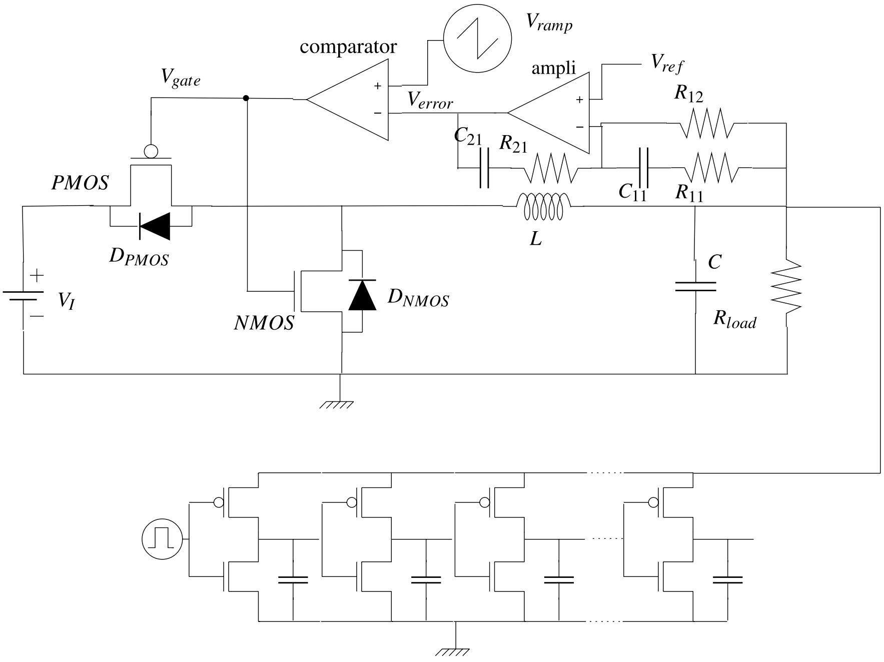


Fig. 8.15 Buck converter supplying a resistor load and an inverter chain

8.2 The Buck Converter Loaded by a Resistor and an Inverter Chain

An inverter chain is supplied by the converter in parallel with the resistor. The input of this chain starts to oscillate between 0 and 1.8 V after 150 μs, *i.e.* 50 μs after the reference voltage reaches 1.8 V. The simulated circuit is shown in Fig. 8.15.

8.2.1 Simulation as a Nonsmooth Dynamical System with SICONOS

The inverter MOSFETs transistors are modeled with a piecewise-linear characteristic

$$I_{DS} = f(V_{GS}, V_{DS}).$$

Their parameters are:

- Transconductance *KP*: 4.3×10^{-5} A V^{-2} for the pMOS, 12.9×10^{-5} A V^{-2} for the nMOS.
- Threshold voltage: −0.6 V for the pMOS and 0.6 V for the nMOS.
- Output capacitive load: 20 fF.

Two chain lengths were tested. In the first case, the chain includes only 2 inverters to enable a CPU time comparison with PLECS whose evaluation version is limited to 6 switches. To emulate a large current drawn from the output of the converter, the 2-inverter chain is supposed to be 8000 times large, *i.e.* 8000 chains switching simultaneously as a synchronous logic circuit. The transconductance and capacitance are therefore multiplied by 8000. The switching frequency of the chain input is 200 MHz. The CPU time required to achieve the simulation of 200 μs with a 1 ns time step is 52 seconds on a Pentium 4 clocked at 3 GHz. Results are displayed in Figs. 8.16 and 8.17. Figure 8.18 shows the start-up of the inverters switching simulated with a 0.5 ns time step to enhance the waveform accuracy.

A 100 inverters long chain was also simulated. The reference voltage rises from 0 to 1.8 V in 50 μs, and the chain input starts to switch 30 μs later at 25 MHz. A 3 Ω load resistor is supplied in parallel with the inverter chain. To emulate a large current drawn from the output of the converter, the 100-inverter chain is supposed to be 2400 times large, i.e. 2400 chains switching simultaneously as a synchronous logic circuit. The transconductance and capacitance are therefore multiplied by 2400. The CPU time required to achieve the simulation of 100 μs with a 0.25 ns time step is 3 hours on a Pentium 4 clocked at 3 GHz. The results are displayed in Figs. 8.19, 8.20 and 8.21.

In Fig. 8.20, one can notice a decrease of 0.02 V of the converter output voltage between times 80 μs and 100 μs caused by the start of the inverters switchings that suddenly increases the current drawn. The effect of this supply voltage variation on the inverters delay is in turn displayed in Fig. 8.21 showing the inverters voltages at times 83 μs and 99 μs. This effect is computed thanks to the modeling of the MOS transistor characteristic. Simulators based on an ideal switch model in series with a resistor cannot show it.

8.2.2 Simulation with PLECS

Only 2 inverters could be used due to the limitations of the evaluation version. The inverter MOS transistors are ideal switches with a R_{ON} value of 0.54 Ω to approximately match the transconductance of the 8000 parallel MOS modeled in SICONOS. The load capacitor is set to 8000×20 fF $= 0.16$ nF (see the description of the PLECS circuit and the SIMULINK model in Figs. 8.22 and 8.23).

The CPU time required to achieve the simulation of 200 μs is 4 hours 8 min on a Pentium 4 clocked at 3 GHz, *i.e.* 300 times the SICONOS simulation time.

8.3 The Delta-Sigma Converter

This section is dedicated to the numerical simulation of a double-bit second order Delta-Sigma converter as depicted in the Fig. 8.24. The inputs are the nodes spg_1

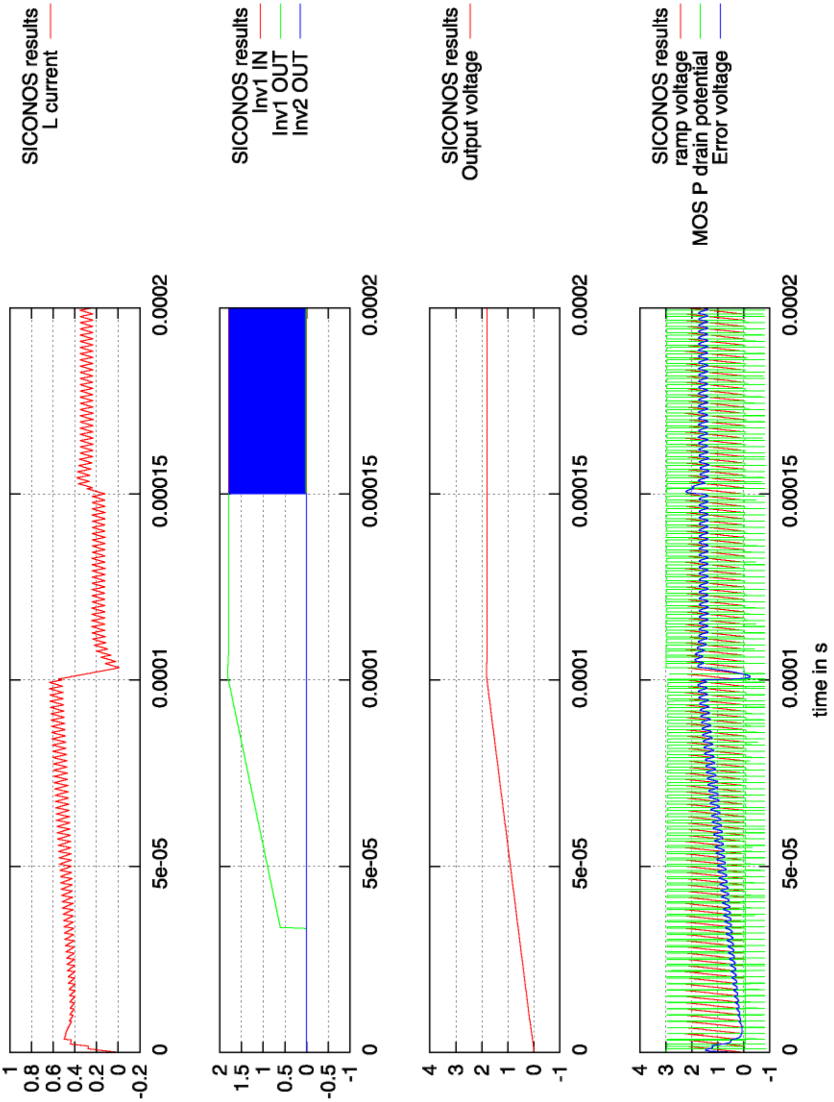


Fig. 8.16 SICONOS simulation results, buck converter supplying a load resistor and an inverter chain (first 200 μs)

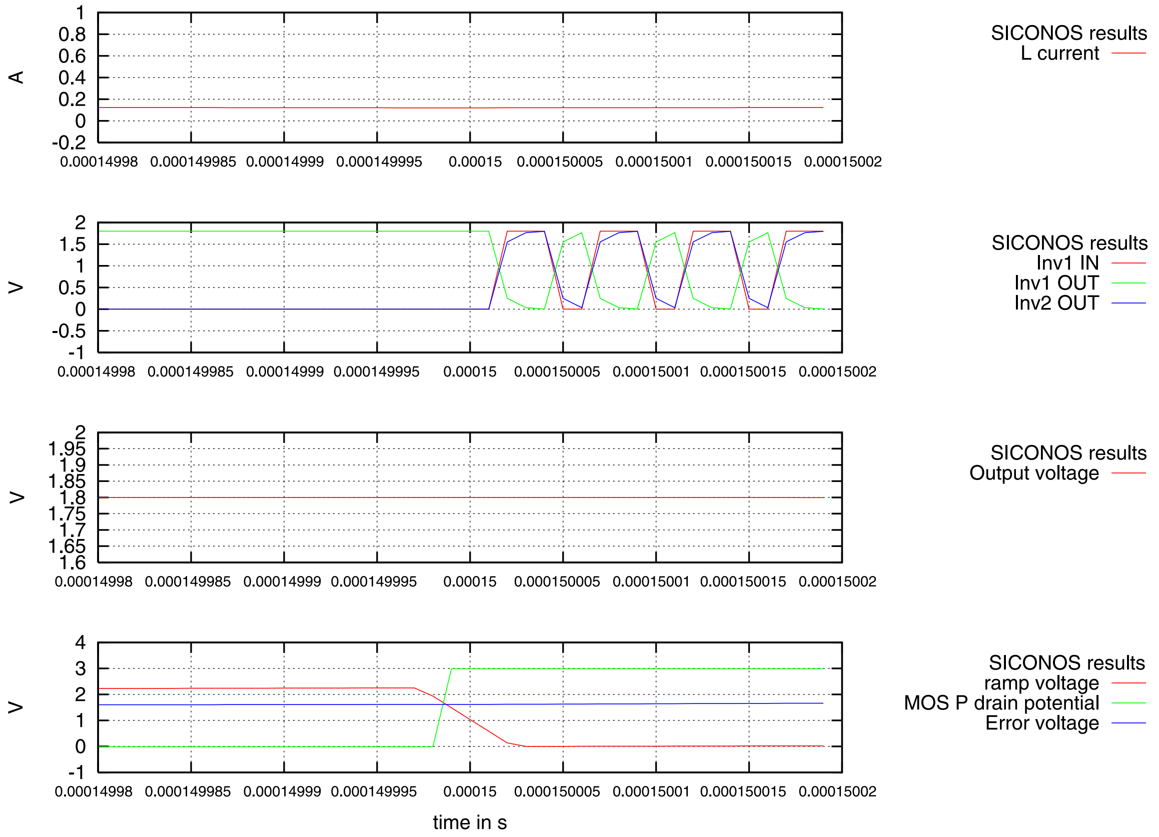


Fig. 8.17 SICONOS simulation results, buck converter supplying a load resistor and an inverter chain; Zoom on the start-up of the inverter chain (time step = 1 ns)

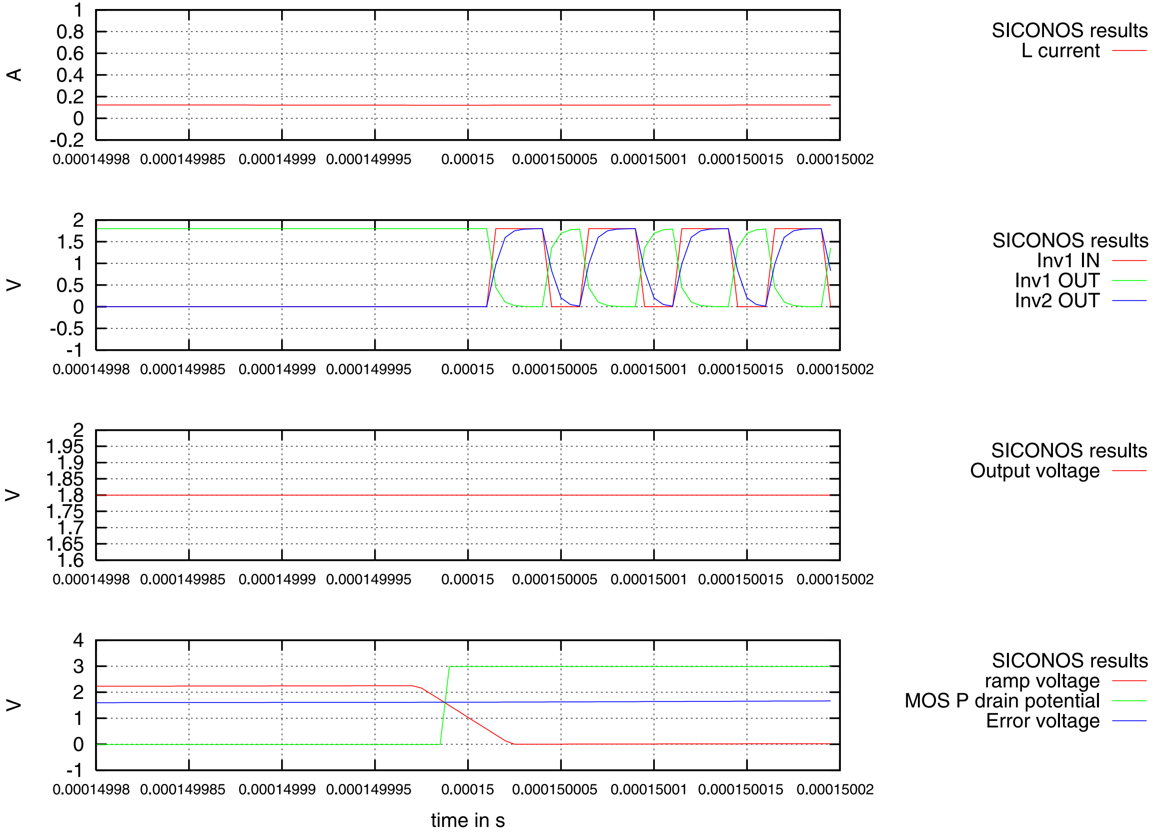


Fig. 8.18 SICONOS simulation results, buck converter supplying a load resistor and an inverter chain; Zoom on the start-up of the inverter chain (time step = 0.5 ns)

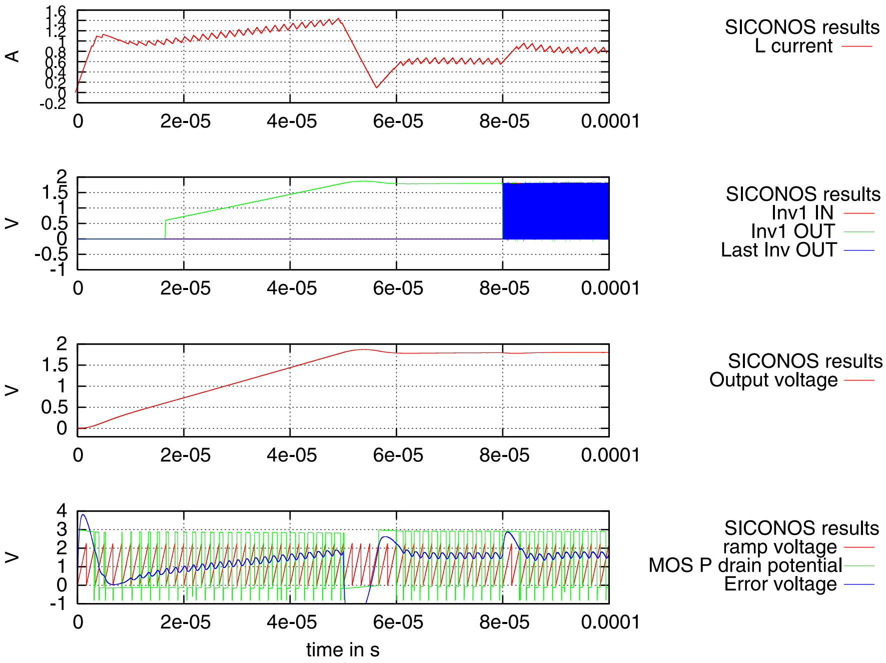


Fig. 8.19 SICONOS simulation results, buck converter supplying a load resistor and a 100 inverters chain (first 100 µs)

and spg_2. The circuit is composed of two switched capacitor integrators. The reference levels from the quantizer are $\{-2.5, 2.5\}$, it oscillates between theses two values in such a manner that its local average equals the average input value. The components are modeled as piecewise-linear components. The features of the non-smooth models are given thereafter:

- The switches are modeled using the model of transistor in Sect. 4.7.5, their threshold voltage is $V_T = 1$ V, the transconductance KP was set to 2.45×10^{-4} A V^{-2} and the supply voltage is 3 V.
- The comparator is modeled as a piecewise-linear function depicted in Fig. 4.10(b) using the parameters value $v_{min} = -2.5$ V, $v_{max} = 2.5$ V and $\varepsilon = 0.01$ V.

The value of the first and second integrator are given thereafter:

- The values of the capacitors are: $C_{int10} = C_{int11} = 4$ pF, $C_{10} = C_{11} = C_{12} = C_{13} = C_{14} = 2$ pF, $C_{int20} = C_{int21} = 1$ pF and $C_{20} = C_{21} = C_{22} = C_{23} = C_{24} = 0.5$ pF.
- The OPA are represented using a comparator modeled with the piecewise-linear function described in Fig. 4.10(b) using the parameters $V_{min} = -5$ V, $V_{max} = 5$ V and $\varepsilon = 1 \times 10^{-3}$ V. The output of this comparator is the output node plus the output of an OPA. The negative output of the OPA is a voltage-controlled source.

The values used for the quantizer are given thereafter:

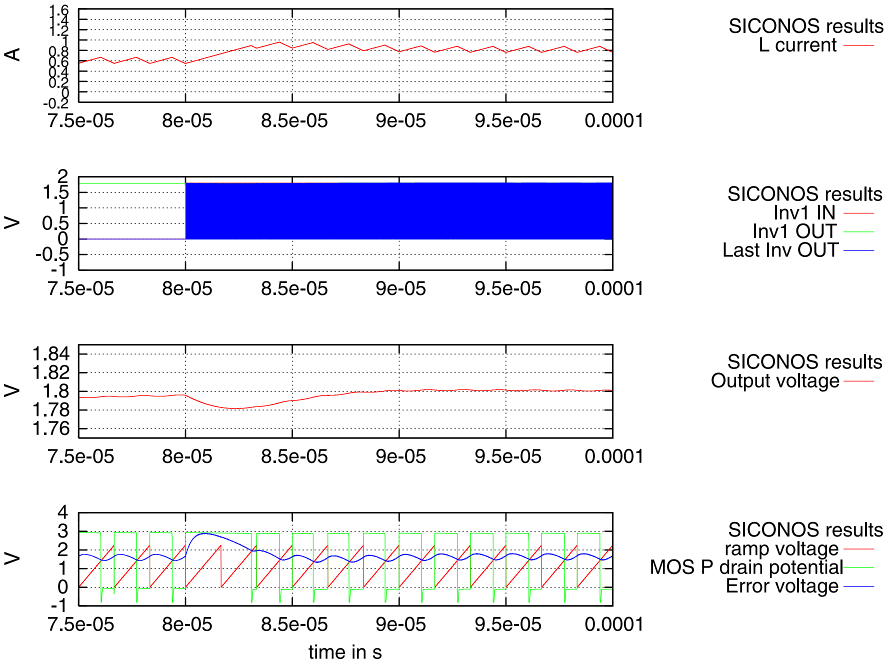


Fig. 8.20 SICONOS simulation results, buck converter supplying a load resistor and a 100 inverters chain; Zoom on the start-up of the inverter chain

- The inverters are represented using a comparator modeled with the piecewise-linear function described in Fig. 4.10(b) using the parameters $V_{min} = 2.5$ V, $V_{max} = -2.5$ V and $\varepsilon = 1 \times 10^{-4}$ V. Their capacitor value is 0.01 pF.

The unnamed capacitors are all set to 0.025 pF.

8.3.1 Dynamical Equations

The dynamics is obtained from the automatic equations generation algorithm. The size of the state vector is 39 and the number of algebraic equations is 65 in the system (5.14). Using a switch model composed of two hyperplanes leads to a nonsmooth law of dimension 156. The simulation consists in doing 4×10^6 fixed steps with a size of 0.1 ns. It needs 10 minutes on a CPU Pentium 4 processor clocked at 3 GHz. It must be underline that our implementation is not based on the sparse matrices. The nonsmooth problem has been solved using either the PATH library or a semi-smooth Newton method based on the Fischer-Burmeister reformulation (DeLuca et al. 1996) that is our own implementation in SICONOS using a convergence tolerance of 10^{-12}.

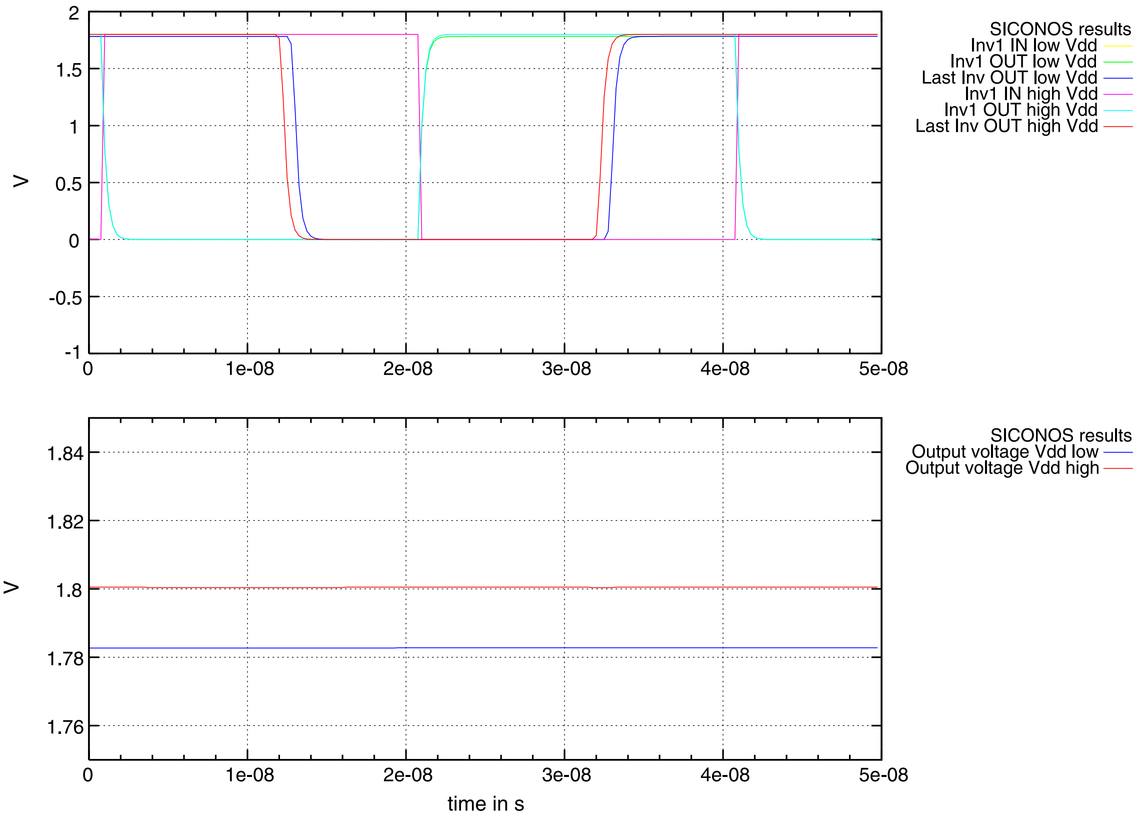


Fig. 8.21 SICONOS simulation results, buck converter supplying a load resistor and a 100 inverters chain; Comparison of inverters delays at times 83 µs (low Vdd) and 99 µs (high Vdd)

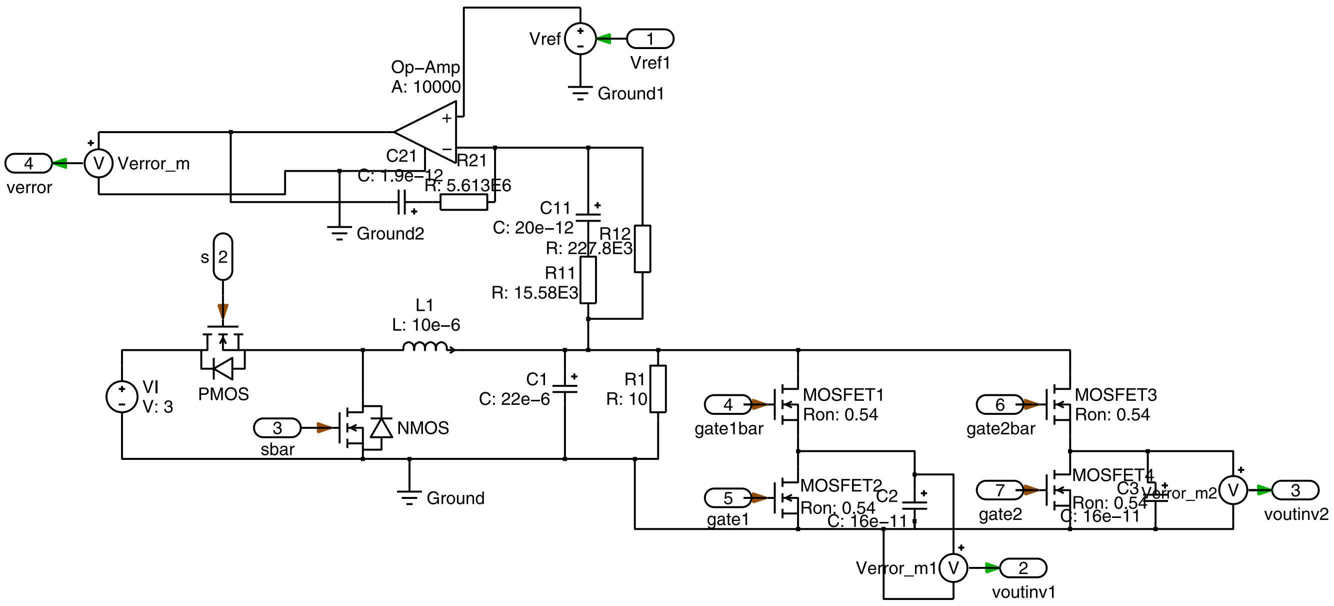


Fig. 8.22 PLECS circuit part of the buck converter supplying a resistor and inverters

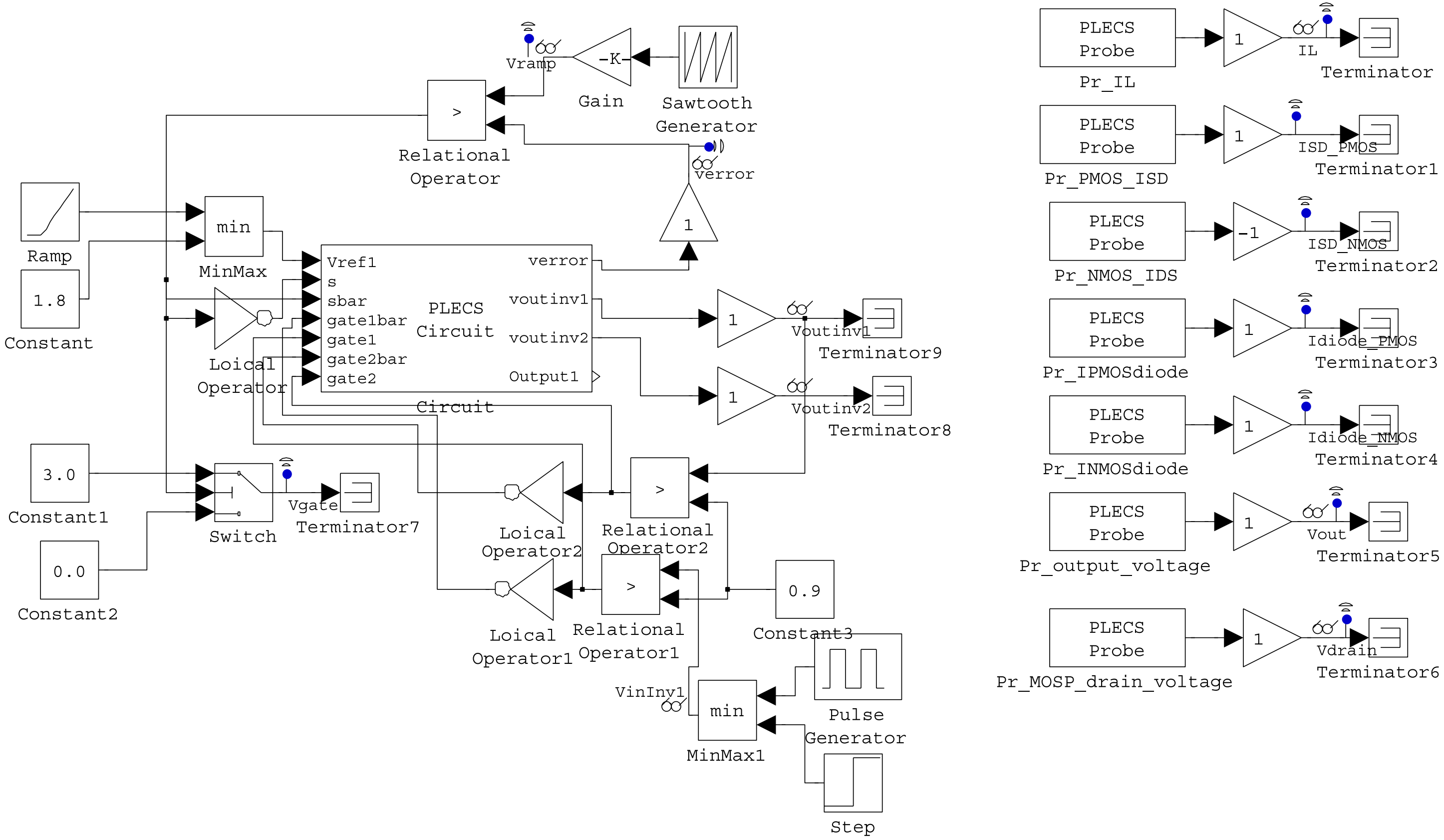


Fig. 8.23 SIMULINK model of the buck converter supplying a resistor and inverters

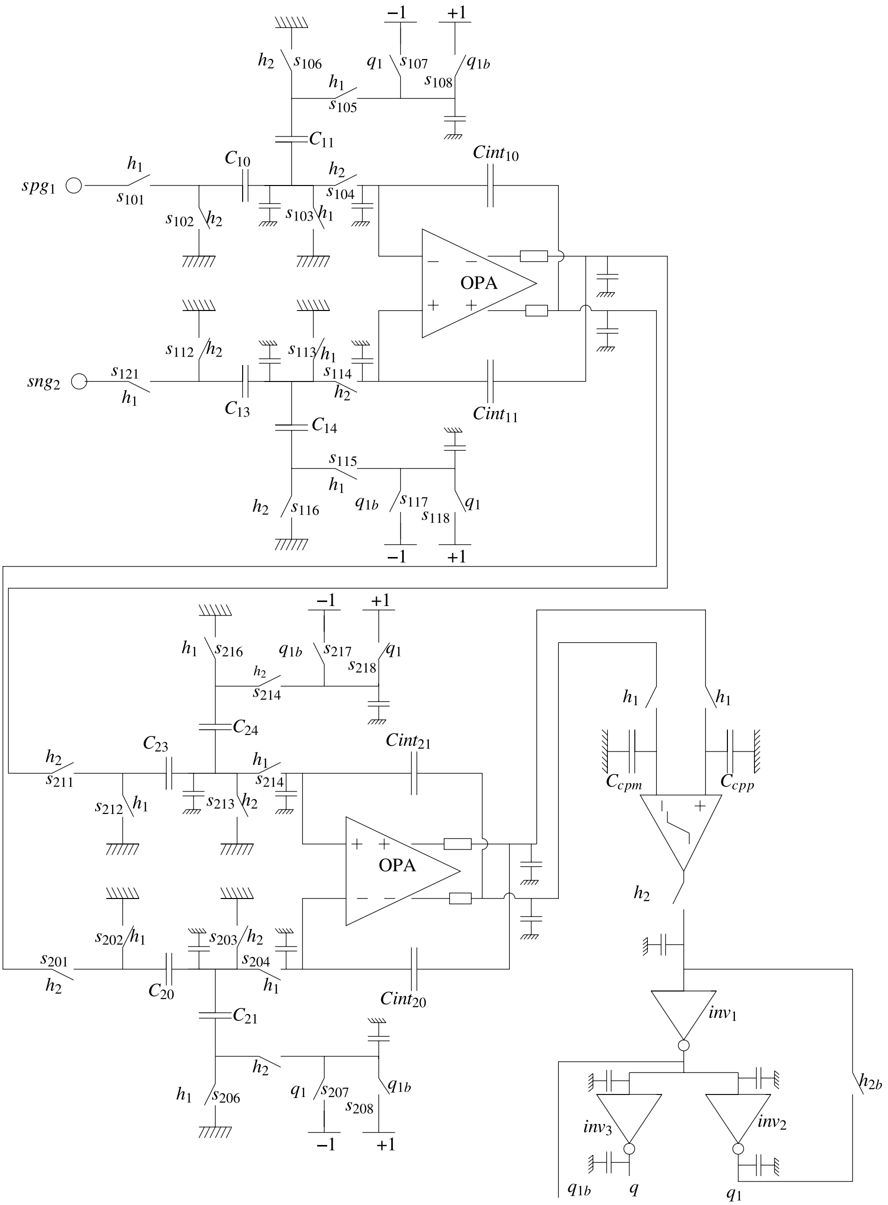


Fig. 8.24 Delta-Sigma converter

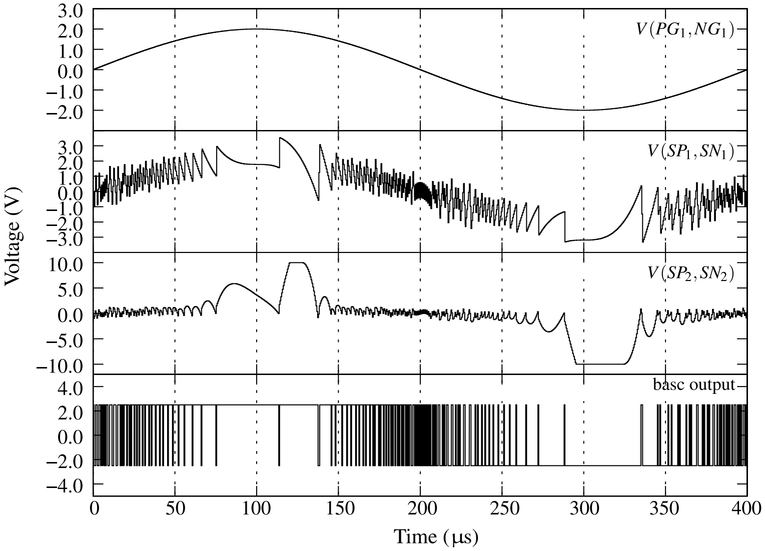


Fig. 8.25 ELDO simulation

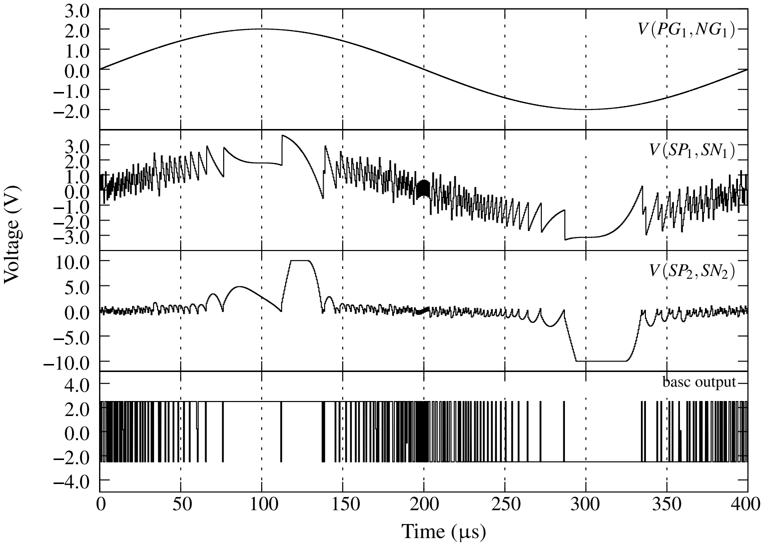


Fig. 8.26 SICONOS simulation

8.3.2 Numerical Results with SICONOS

The results obtained with SICONOS are depicted in Fig. 8.26. The same simulation is made with ELDO in Fig. 8.25, showing that both the NSDS and the SPICE approaches yield very close results.

8.3.3 *Comparisons and Discussions*

It is noteworthy that the simulations using a quadratic model for the transistor leads to messages of "no-convergence" of the ELDO Newton-Raphson algorithm. Moreover such simulations provide a wrong value of the OPA output. This is not the case with SICONOS.

Remark 8.2 It is crucial to keep in mind that the NSDS SICONOS package is an open-source academic tool, that does not possess the degree of optimization of the commercial software packages like ELDO or SMASH. The reported computation times for SICONOS have therefore to be considered as rough upper bounds that may significantly be improved.

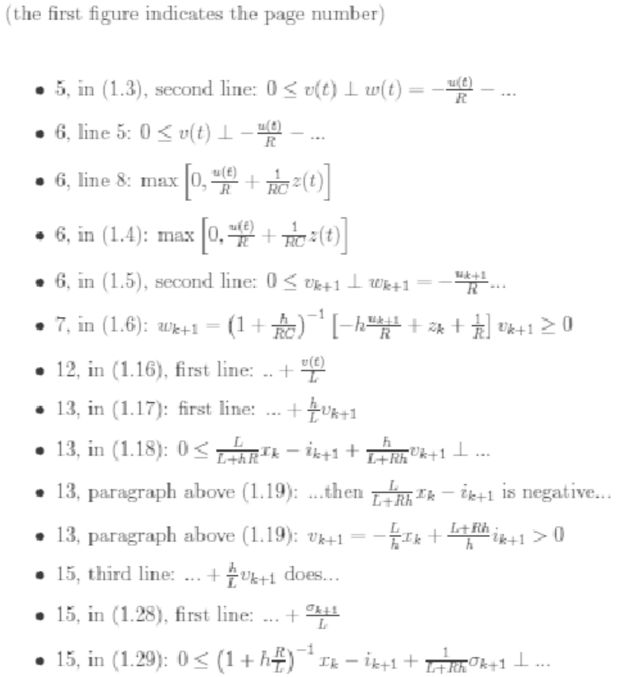
8.4 Conclusions

Chapters 7 and 8 show through numerous examples that the NSDS approach may supersede the SPICE and hybrid approaches in many instances of switched circuits. Despite SICONOS is a software package developed in an academic context (and which consequently can not benefit from the global code optimization of commercially available packages), the results which are shown in Chaps. 7 and 8 clearly demonstrate its power and its efficiency.

Erratum to: Chapters 1 and 2 of *Nonsmooth Modeling and Simulation for Switched Circuits*

Vincent Acary, Olivier Bonnefon and Bernard Brogliato

Erratum to:
Chapter 1 in : Vincent Acary et al., *Nonsmooth Modeling and Simulation for Switched Circuits,*
DOI 10.1007/978-90-481-9681-4_1

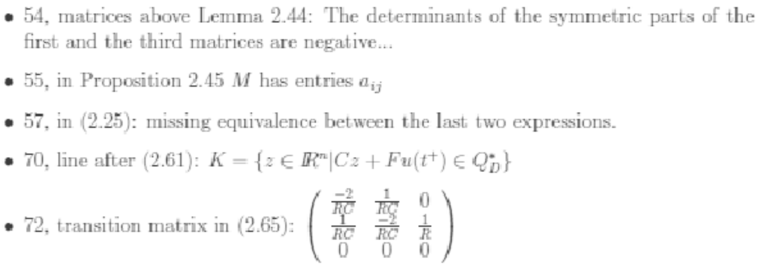


(the first figure indicates the page number)

- 5, in (1.3), second line: $0 \leq v(t) \perp w(t) = -\frac{u(t)}{R} - ...$
- 6, line 5: $0 \leq v(t) \perp -\frac{u(t)}{R} - ...$
- 6, line 8: $\max\left[0, \frac{u(t)}{R} + \frac{1}{RC}z(t)\right]$
- 6, in (1.4): $\max\left[0, \frac{u(t)}{R} + \frac{1}{RC}z(t)\right]$
- 6, in (1.5), second line: $0 \leq v_{k+1} \perp w_{k+1} = -\frac{u_{k+1}}{R}...$
- 7, in (1.6): $w_{k+1} = \left(1 + \frac{h}{RC}\right)^{-1}\left[-h\frac{u_{k+1}}{R} + z_k + \frac{1}{R}\right]v_{k+1} \geq 0$
- 12, in (1.16), first line: $.. + \frac{v(t)}{L}$
- 13, in (1.17): first line: $... + \frac{h}{L}v_{k+1}$
- 13, in (1.18): $0 \leq \frac{L}{L+hR}x_k - i_{k+1} + \frac{h}{L+Rh}v_{k+1} \perp ...$
- 13, paragraph above (1.19): ...then $\frac{L}{L+Rh}x_k - i_{k+1}$ is negative...
- 13, paragraph above (1.19): $v_{k+1} = -\frac{L}{h}x_k + \frac{L+Rh}{h}i_{k+1} > 0$
- 15, third line: $... + \frac{h}{L}v_{k+1}$ does...
- 15, in (1.28), first line: $... + \frac{\sigma_{k+1}}{L}$
- 15, in (1.29): $0 \leq \left(1 + h\frac{R}{L}\right)^{-1}x_k - i_{k+1} + \frac{1}{L+Rh}\sigma_{k+1} \perp ...$

V. Acary et al., *Nonsmooth Modeling and Simulation for Switched Circuits,*
Lecture Notes in Electrical Engineering 69,
DOI 10.1007/978-90-481-9681-4_9, © Springer Science + Business Media B.V. 2011

Erratum to:
Chapter 2 in : Vincent Acary et al., *Nonsmooth Modeling and Simulation for Switched Circuits,*
DOI 10.1007/978-90-481-9681-4_2



- 54, matrices above Lemma 2.44: The determinants of the symmetric parts of the first and the third matrices are negative...
- 55, in Proposition 2.45 M has entries a_{ij}
- 57, in (2.25): missing equivalence between the last two expressions.
- 70, line after (2.61): $K = \{z \in \mathbb{R}^n | Cz + Fu(t^+) \in Q_D^*\}$
- 72, transition matrix in (2.65): $\begin{pmatrix} \frac{-2}{RC} & \frac{1}{RC} & 0 \\ \frac{1}{RC} & \frac{-2}{RC} & \frac{1}{R} \\ 0 & 0 & 0 \end{pmatrix}$

____

The online versions of the original chapter can be found at
DOI 10.1007/978-90-481-9681-4_1
DOI 10.1007/978-90-481-9681-4_2

____

V. Acary et al., *Nonsmooth Modeling and Simulation for Switched Circuits,*
Lecture Notes in Electrical Engineering 69,
DOI 10.1007/978-90-481-9681-4_9, © Springer Science + Business Media B.V. 2011

Appendix A
Some Facts in Real Analysis

This chapter provides an introduction to the mathematical tools which are needed to rigorously define what is a measure differential inclusion as the one in Sect. 2.4.1. This is not mandatory reading for those who prefer to focus on the numerical and modeling aspects only.

A.1 Absolutely Continuous Functions and Sets

Definition A.1 (Absolutely continuous function) Let $f : I \subset \mathbb{R} \to \mathbb{R}$ be a function. It is said *absolutely continuous* if for all $\epsilon > 0$, there exists $\delta > 0$ such that for all finite sequences of disjoint intervals (a_k, b_k) of I such that $\sum_k |b_k - a_k| < \delta$, one has $\sum_k |f(b_k) - f(a_k)| < \epsilon$.

An absolutely continuous function $f(\cdot)$ on an interval $I \subset \mathbb{R}$ is such that $f \in L^1(I)$, and there exists $g(\cdot)$ with $g \in L^1(I)$ with $\int_I f(s)\dot{\varphi}(s)ds = -\int_I g(s)\varphi(s)ds$ for all test functions $\varphi(\cdot)$ which are continuously differentiable and with compact support in I. This means that $g(\cdot)$ is the derivative of $f(\cdot)$ in the sense of distributions, so that $\dot{f}(\cdot)$ is equal to $g(\cdot)$ almost everywhere in I. One denotes $\dot{f} = g$ and one has

$$f(s) - f(t) = \int_t^s g(\tau)d\tau \tag{A.1}$$

which is known as the Lebesgue-Vitali theorem. The distance of a point x to a set C is defined by $d_C(x) = \inf\{\|x - y\|,\ y \in C\}$.

Definition A.2 (Absolutely continuous multifunction) A set-valued mapping $C : I \to \mathbb{R}$ varies in an absolutely continuous way if there exists an absolutely continuous function $v : I \to \mathbb{R}$ such that, for any $x \in \mathbb{R}$ and $s, t \in I$, one has:

$$|d_{C(t)}(x) - d_{C(s)}(x)| \leqslant |v(t) - v(s)|$$

V. Acary et al., *Nonsmooth Modeling and Simulation for Switched Circuits*,

DOI 10.1007/978-90-481-9681-4, © Springer Science+Business Media B.V. 2011

A.2 Lipschitz Continuous Functions and Sets

Definition A.3 (Lipschitz continuous function) Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be a function. It is said *Lipschitz continuous* if there exists a bounded constant $k > 0$ such that for all $x, y \in \mathbb{R}^n$ one has $\|f(x) - f(y)\| \leqslant k\|x - y\|$.

A function that is Lipschitz continuous is also absolutely continuous. Lipschitz continuity of $f(\cdot)$ is equivalent to absolute continuity of $f(\cdot)$ plus boundedness of the derivative $\dot{f} = g$.

Before introducing this notion for set-valued maps, let us introduce the Hausdorff's distance.

Definition A.4 (Hausdorff distance) Let A and B be two non empty sets of $\mathbb{R}^n$. We define the distance between a point x and a set A as

$$\rho(x, A) = \inf_{a \in A} \|x - a\|$$

and

$$d_H(A, B) = \max\left\{\sup_{x \in A} \rho(x, B), \sup_{x \in B} \rho(x, A)\right\} \tag{A.2}$$

which is the Hausdorff's distance between A and B.

Definition A.5 (Lipschitz continuous multifunction) Let $C : \mathbb{R}^n \to \mathbb{R}^m$ be a set-valued map. It is Lipschitz continuous in the Hausdorff distance if there exists a constant $k > 0$ such that $d_H(C(t), C(s)) \leqslant k|t - s|$ for all $t, s \in I$.

There exists a geometric definition of Lipschitz continuity for set-valued maps, in terms of inclusions of sets (see *e.g.* Definition 2.2 in Acary and Brogliato 2008). This is equivalent to the above definition when the Hausdorff distance is replaced by the so-called Pompeiu-Hausdorff distance, see *e.g.* Rockafellar and Wets (1997, Chap. 9).

A.3 Functions of Bounded Variations in Time

Let I be an interval, and define a subdivision S_n of I as $x_0 < x_1 < \cdots < x_n$. The variation of a function $f : \mathbb{R} \to \mathbb{R}^n$ on I with respect to the subdivision S_n is defined as

$$\mathrm{var}_{I,S_n}(f) = \sum_{i=0}^{n} \|f(x_{i+1}) - f(x_i)\|.$$

The function $f(\cdot)$ is said to have a bounded variations on I if

$$\sup_{S_n} \mathrm{var}_{I,S_n}(f) \leqslant C$$

for some bounded constant C. Then $\mathrm{var}_I(f)$ is called the total variation of $f(\cdot)$ on I. A function that has a bounded variations on any compact subinterval of I is said to be of *local bounded variations* (LBV). If it is right continuous and LBV it will be denoted RCLBV. If it is right continuous and BV it is denoted as RCBV.

BV functions have the following fundamental properties:

- Let E_f be the set of points x where $f(\cdot)$ has discontinuities. Then E_f is countable.
- If $f(\cdot)$ is BV, then it is Riemann integrable.
- BV functions have left and right limits at all points (of their domain of definition).[1]
- The derivative of a BV function can be decomposed into three parts: a Lebesgue integrable part, a purely atomic measure, and a measure that is singular with respect to the Lebesgue measure and is non atomic (see below).
- Functions of special bounded variations possess a derivative that is the sum of a Lebesgue integrable function, and a purely atomic measure. The third part vanishes for SBV functions.

In most engineering applications, it may be reasonably assumed that the derivative of a BV function is just the sum of an integrable function, and a purely atomic measure of the form $\sum_i \delta_i$ for some set of i.

- We denote by $\mathrm{LBV}(I;\mathbb{R}^n)$ the space of functions of locally bounded variations, i.e. of bounded variations on every compact subinterval of I.
- We denote by $\mathrm{RCLBV}(I;\mathbb{R}^n)$ the space of right-continuous functions of locally bounded variations. It is known that if $x \in \mathrm{RCLBV}(I;\mathbb{R}^n)$ and $[a,b]$ denotes a compact subinterval of I, then x can be represented in the form (see e.g. Shilov and Gurevich 1966):
$$x(t) = \mathscr{J}_x(t) + [x](t) + \zeta_x(t), \quad \forall t \in [a,b] \tag{A.3}$$
where $\mathscr{J}_x$ is a jump function, $[x]$ is an absolutely continuous function and ζ_x is a singular function. Here $\mathscr{J}_x$ is a jump function in the sense that $\mathscr{J}_x$ is right-continuous and given any $\varepsilon > 0$, there exist finitely many points of discontinuity $t_1,\dots,t_N$ of $\mathscr{J}_x$ such that $\sum_{i=1}^N \|\mathscr{J}_x(t_i) - \mathscr{J}_x(t_i^-)\| + \varepsilon > \mathrm{var}(\mathscr{J}_x,[a,b])$, $[x]$ is an absolutely continuous function in the sense that for every $\varepsilon > 0$, there exists $\delta > 0$ such that $\sum_{i=1}^N \|[x](\beta_i) - [x](\alpha_i)\| < \varepsilon$, for any collection of disjoint subintervals $]\alpha_i,\beta_i] \subset [a,b] (1 \leqslant i \leqslant N)$ such that $\sum_{i=1}^N(\beta_i - \alpha_i) < \delta$, and ζ_x is a singular function in the sense that ζ_x is a continuous and of bounded variations function on $[a,b]$ such that $\dot{\zeta}_x = 0$ almost everywhere on $[a,b]$.
- By $u \in \mathrm{RCSLBV}(I;\mathbb{R}^n)$ it is meant that x is a right-continuous function of special locally bounded variations, i.e. x is of bounded variations and can be written as the sum of a jump function and an absolutely continuous function on every compact subinterval of I. So, if $x \in \mathrm{RCSLBV}(I;\mathbb{R}^n)$ then
$$x = [x] + \mathscr{J}_x \tag{A.4}$$

[1] Throughout the book the right (left) limits at t are denoted either as $f^+(t)$ $(f^-(t))$ or as $f(t^+)$ $(f(t^-))$.

where $[x]$ is a locally absolutely continuous function called the absolutely continuous component of x and $\mathscr{J}_x$ is uniquely defined up to a constant by

$$\mathscr{J}_x(t) = \sum_{t \geqslant t_n} x(t_n^+) - x(t_n^-) = \sum_{t \geqslant t_n} x(t_n) - x(t_n^-) \tag{A.5}$$

where $t_1, t_2, \ldots, t_n, \ldots$ denote the countably many points of discontinuity of x in I.

A.4 Multifunctions of Bounded Variation in Time

A moving set $t \mapsto K(t)$ is said of right continuous bounded variation in time on $[0, T]$, if there exists a right continuous non-decreasing function $r : [0, T] \to \mathbb{R}$ such that

$$d_H(K(t), K(s)) \leqslant r(t) - r(s), \quad \text{for all } 0 \leqslant s \leqslant t \leqslant T.$$

Let $r(0) = 0$. For any partition $0 = t_0 < t_1 < \cdots < t_N = T$ of $[0, T]$, this yields

$$\sum_{i=0}^{N-1} d_H(K(t_{i+1}), K(t_i)) \leqslant \sum_{i=0}^{N-1} [r(t_{i+1}) - r(t_i)] = r(T).$$

Therefore the first inequality can be interpreted as requiring that $t \mapsto K(t)$ is of bounded variations. We conclude that the above definition of the variation of a function, can be extended to set-valued functions, where the Euclidean distance is replaced by the Hausdorff's distance.

A.5 Differential Measures

Details on differential measures may be found in Schwartz (1993), Monteiro Marques (1993), and Moreau (1988).

Definition A.6 Let $x : I \to \mathbb{R}^n$ be a BV function, $I \neq \emptyset$, $I \subseteq \mathbb{R}$. Let $\varphi(\cdot)$ be a continuous real function on I, with compact support. Let $\mathscr{P}$ denote the set of finite partitions of I, each partition P_N with nodes $t_0 < t_1 < \cdots < t_N$. Let $\theta_k \in [t_{k-1}, t_k]$ for all intervals of the partition P_N. The Riemann-Stieltjes sums $S(\varphi, P_N, \theta; x) = \sum_{k=1}^{N} \varphi(\theta_k)(x(t_k) - x(t_{k-1}))$ converge as $N \to +\infty$ to a limit independent of the θ_k. This limit is denoted as

$$\int \varphi dx \tag{A.6}$$

where dx is the differential measure associated to $x(\cdot)$. The map $x \mapsto dx$ is linear.

If $x(\cdot)$ is constant, $dx = 0$. If $dx = 0$ and $x(\cdot)$ is right-continuous in the interior of I, then $x(\cdot)$ is constant. If $x(\cdot)$ is a step function, then dx is the sum of a finite collection of Dirac measures with atoms at the discontinuity points of $x(\cdot)$. For $a \leqslant b$, $a, b \in I$:

$$\begin{aligned} dx([a,b]) &= x(b^+) - x(a^-), \\ dx([a,b)) &= x(b^-) - x(a^-), \\ dx((a,b]) &= x(b^+) - x(a^+), \\ dx((a,b)) &= x(b^-) - x(a^+). \end{aligned}$$

In particular, we have

$$dx(\{a\}) = x(a^+) - x(a^-)$$

Obviously when measuring a singleton only the part of the differential measure that corresponds to the jump function $\mathscr{J}_x(t)$ may play a role because the other two parts corresponding to $[x](t)$ and $\zeta_x(t)$ are non atomic: therefore necessarily $d[x](\{a\}) = [\dot{x}(t)]dt(\{a\}) = 0$ and $d\zeta_x(\{a\}) = 0$.

For any RCLBV mapping $x : I \to \mathbb{R}^n$ on a subinterval I of $\mathbb{R}$ one has:

$$x(t) = x(s) + \int_{(s,t]} dx \quad \text{for all } s, t \in I \text{ with } s \leqslant t.$$

which is the BV counterpart of (A.1).

The next results are useful when dealing with quadratic functionals of BV functions (like quadratic Lyapunov functions in stability analysis; Brogliato 2004). For $x \in \mathrm{LBV}(I;\mathbb{R}^n)$, x^+ and x^- denote the functions defined by

$$x^+(t) = x(t^+) = \lim_{s \to t, s > t} x(s), \quad \forall t \in I, t < \sup\{I\}$$

and

$$x^-(t) = x(t^-) = \lim_{s \to t, s < t} x(s), \quad \forall t \in I, t > \inf\{I\}$$

(where $\sup\{I\}$ (resp. $\inf\{I\}$) denotes the supremum (resp. infimum) of the set I). If $x, y \in LBV(I;\mathbb{R}^n)$ then $x^T y \in LBV(I;\mathbb{R})$ and

$$d(x^T y) = (y^-)^T dx + (x^+)^T dy = (y^+)^T dx + (x^-)^T dy. \tag{A.7}$$

Let us also recall that

$$2(x^-)^T dx \leqslant d(x^T x) = (x^+ + x^-)^T dx \leqslant 2(x^+)^T dx. \tag{A.8}$$

A.6 Measure Differential Inclusion (MDI)

Roughly speaking, a measure differential inclusion is a differential inclusion whose solution may jump, so that its derivative contains Dirac measures. Consequently the left-hand-side of the inclusion is a measure, say dv, and the set-valued right-hand-side $F(w(t))$ contains measures. One has to give a rigorous meaning to such inclusions. Recall that λ is here the Lebesgue measure ($d\lambda = dt$), μ is the measure associated with the variation function $\mathrm{var}_C(\cdot)$, that is the variation of the set valued map $C(\cdot)$ over $[0,T]$. The following definition is adapted from Edmond and Thibault (2006, Definition 2.1).

Definition A.7 Let $C(\cdot)$ and $f(\cdot,\cdot)$ satisfy the conditions of Theorem 2.56. A function $x:[0,T]\to\mathbb{R}^n$ is a solution of the differential inclusion (2.36) provided that:

- $x(\cdot)$ is of bounded variation, right continuous, and it satisfies $x(0)=x_0$ and $x(t)\in C(t)$ for all $t\in[0,T]$.
- there exists a positive Radon measure ν that is absolutely continuously equivalent to the measure $\mu+\lambda$, with respect to which the differential measure dx is absolutely continuous with density $\frac{dx}{d\nu}\in L^1_\nu([0,T],\mathbb{R}^n)$ and

$$\begin{aligned}\frac{dx}{d\nu}(t)+f(t,x(t))\frac{d\lambda}{d\nu}&\in -N_{C(t)}(x(t)),\\ \nu\text{-almost everywhere } t&\in[0,T].\end{aligned}\tag{A.9}$$

One sees that the inclusion is now written in (A.9) with the densities of the measures dx and $d\lambda$, which are functions of time. This is equivalent to the writing in (2.36) that is an inclusion of measures. Suppose that $\nu=\delta_t+dt$ for some $t\in[0,T]$. Then if t is an atom of dx (a discontinuity time of the function $x(\cdot)$) one obtains that $\frac{dx}{d\nu}(t)=(x(t^+)-x(t^-))$. Outside atoms of dx one obtains simply $\frac{dx}{d\nu}(t)=\frac{dx}{dt}(t)=\dot{x}(t)$. The fact that one may choose other "basis" measures ν enlarges the scope of the formalism.

Notice in passing that (A.9) allows us to give a meaning to the inclusion of a measure into a set (here a closed convex cone). Let the measure ν be non negative and let dx be absolutely continuous with respect to ν. Writing $dx=x_\nu d\nu\in K(t)$ for some closed convex cone $K(t)$ means that the density of dx with respect to ν, $\frac{dx}{d\nu}(\cdot)$, belongs to $K(t)$. So in case t is an atom of dx one simply gets $(x(t^+)-x(t^-))\in K(t)$.

References

V. Acary, B. Brogliato, *Numerical Methods for Nonsmooth Dynamical Systems: Applications in Mechanics and Electronics*. Lecture Notes in Applied and Computational Mechanics, vol. 35 (Springer, Berlin, 2008)

V. Acary, B. Brogliato, Implicit Euler numerical scheme and chattering-free implementation of sliding mode systems. Systems and Control Letters **59**, 284–293 (2010)

V. Acary, B. Brogliato, D. Goeleven, Higher order Moreau's sweeping process: mathematical formulation and numerical simulation. Mathematical Programming, Series A **113**(1), 133–217 (2008)

V. Acary, O. Bonnefon, B. Brogliato, Improved circuit simulator. Patent number 09/02605, May 2009.

V. Acary, O. Bonnefon, B. Brogliato, Time-stepping numerical simulation of switched circuits with the nonsmooth dynamical systems approach. IEEE Transactions on Computer-Aided Design for Integrated Circuits and Systems **29**(7), 1042–1055 (2010)

K. Addi, B. Brogliato, D. Goeleven, A qualitative mathematical analysis of a class of linear variational inequalities via semi-complementarity problems: applications in electronics. Mathematical Programming A (2010). doi:10.1007/s10107-009-0268-7

K. Addi, S. Adly, B. Brogliato, D. Goeleven, A method using the approach of Moreau and Panagiotopoulos for the mathematical formulation of non-regular circuits in electronics. Nonlinear Analysis: Hybrid Systems **1**(1), 30–43 (2007)

U. Ascher, L. Petzold, *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations* (SIAM, Philadelphia, 1998)

S. Bächle, F. Ebert, Element-based topological index reduction for differential-algebraic equations in circuit simulation. Technical Report Preprint 05-246 (Matheon), Inst. f. Mathematik, TU Berlin, 2005a

S. Bächle, F. Ebert, Graph theoretical algorithms for index reduction in circuit simulation. Technical Report Preprint 05-245 (Matheon), Inst. f. Mathematik, TU Berlin, 2005b

J. Bastien, M. Schatzman, Numerical precision for differential inclusions with uniqueness. ESAIM M2AN: Mathematical Modelling and Numerical Analysis **36**(3), 427–460 (2002)

C. Batlle, E. Fossas, A. Miralles, Generalized discontinuous conduction modes in the complementarity formalism. IEEE Transactions on Circuits and Systems II, Express Briefs **52**(8), 447–451 (2005)

D. Bedrosian, J. Vlach, Analysis of switched networks. International Journal of Circuit Theory and Applications **20**, 309–325 (1992)

S. Billups, S. Dirkse, M. Ferris, A comparison of large scale mixed complementarity problem solvers. Computational Optimization and Applications **7**, 3–25 (1997)

D. Biolek, J. Dobes, Computer simulation of continuous-time and switched circuits: limitations of SPICE-family programs and pending issues, in *Radioelektronika, 17th Int. Conference*, Brno, Czech Republic, 24–25 April 2007

V. Acary et al., *Nonsmooth Modeling and Simulation for Switched Circuits,*
Lecture Notes in Electrical Engineering 69,
DOI 10.1007/978-90-481-9681-4, © Springer Science+Business Media B.V. 2011

W. Bliss, S. Smith, K. Loh, A switched-capacitor realization of discrete-time block filters, in *35th IEEE Midwest Symposium on Circuits and Systems*, vol. 1, 9–12 August 1992, pp. 429–432

F. Branin Jr., Computer methods of network analysis. Proceedings of the IEEE **55**(11), 1787–1801 (1967). ISBN: 0018-9219

K. Brenan, S. Campbell, L. Petzold, *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations* (North-Holland, Amsterdam, 1989)

B. Brogliato, *Nonsmooth Mechanics: Models, Dynamics and Control*, 2nd edn. (Springer, London, 1999)

B. Brogliato, Some perspectives on the analysis and control of complementarity systems. IEEE Transactions on Automatic Control **48**(6), 918–935 (2003)

B. Brogliato, Absolute stability and the Lagrange-Dirichlet theorem with monotone multivalued mappings. Systems and Control Letters **51**, 343–353 (2004)

B. Brogliato, D. Goeleven, The Krasovskii-LaSalle invariance principle for a class of unilateral dynamical systems. Mathematics of Control, Signals and Systems **17**(1), 57–76 (2005)

B. Brogliato, D. Goeleven, Well-posedness, stability and invariance results for a class of multivalued Lur'e dynamical systems. Nonlinear Analysis: Theory, Methods and Applications (2010, in press)

B. Brogliato, L. Thibault, Existence and uniqueness of solutions for non-autonomous complementarity dynamical systems. Journal of Convex Analysis **17**(3–4) (2010). Special issue in the honour of H. Attouch 60th birthday

B. Brogliato, A. Daniilidis, C. Lemaréchal, V. Acary, On the equivalence between complementarity systems, projected systems and differential inclusions. Systems and Control Letters **55**, 45–51 (2006)

B. Brogliato, R. Lozano, B. Maschke, O. Egeland, *Dissipative Systems Analysis and Control Theory and Applications*, 2nd edn. Communications and Control Engineering (Springer, London, 2007)

J. Butcher, *The Numerical Analysis of Ordinary Differential Equations—Runge-Kutta and General Linear Methods* (Wiley, New York, 1987)

K. Camlibel, Complementarity methods in the analysis of piecewise linear dynamical systems, PhD thesis, Katholieke Universiteit Brabant, 2001. ISBN 90 5668 073X

M.K. Camlibel, W. Heemels, J. Schumacher, Consistency of a time-stepping method for a class of piecewise-linear networks. IEEE Transactions on Circuits and Systems I **49**, 349–357 (2002a)

M. Camlibel, W. Heemels, J. Schumacher, On linear passive complementarity systems. European Journal of Control **8**(3), 220–237 (2002b)

M. Cao, M. Ferris, A pivotal method for affine variational inequalities. Mathematics of Operations Research **21**(1), 44–64 (1996)

A. Carbone, F. Palma, Discontinuity correction in piecewise-linear models of oscillators for phase noise characterization. International Journal of Circuit Theory and Applications **35**(1), 93–104 (2006)

L. Chua, A. Dang, Canonical piecewise-linear analysis: Part II—tracing driving-point and transfer characteristics. IEEE Transactions on Circuits and Systems **CAS-32**(5), 417–444 (1985)

L. Chua, R. Ying, Canonical piecewise-linear analysis. IEEE Transactions on Circuits and Systems **CAS-30**(3), 125–140 (1983)

L. Chua, C. Desoer, E. Kuh, *Linear and Non Linear Circuits* (McGraw-Hill, New York, 1991)

H. Chung, A. Ioinovici, Fast computer aided simulation of switching power regulators based on progressive analysis of the switches' state. IEEE Transactions on Power Electronics **9**(2), 206–212 (1994)

F. Clarke, Generalized gradients and its applications. Transactions of AMS **205**, 247–262 (1975)

B. Cornet, Existence of slow solutions for a class of differential inclusions. Journal of Mathematical Analysis and Applications **96**, 130–147 (1983)

J. Cortés, Discontinuous dynamical systems. a tutorial on solutions, nonsmooth analysis, and stability. IEEE Control Systems Magazine, 36–73 (2008)

R.W. Cottle, J. Pang, R.E. Stone, *The Linear Complementarity Problem* (Academic Press, Boston, 1992)

L. De Kelper, A. Dessaint, K. Al-Haddad, H. Nakra, A comprehensive approach to fixed-step simulation of switched circuits. IEEE Transactions on Power Electronics **17**(2), 216–224 (2002)

K. Deimling, *Multivalued Differential Equations* (de Gruyter, Berlin, 1992)

K. Dekker, J. Verwer, *Stability of Runge-Kutta Methods for Stiff Nonlinear Differential Equations*. CWI Monographs (North-Holland, Amsterdam, 1984)

T. DeLuca, F. Facchinei, C. Kanzow, A semismooth equation approach to the solution of nonlinear complementarity problems. Mathematical Programming **75**, 407–439 (1996)

P. Denoyelle, V. Acary, The non-smooth approach applied to simulating integrated circuits and power electronics. Evolution of electronic circuit simulators towards fast-SPICE performance. INRIA Research Report 0321, 2006. http://hal.inria.fr/docs/00/08/09/20/PDF/RT-0321.pdf

A. Dontchev, F. Lempio, Difference methods for differential inclusions: a survey. SIAM Reviews **34**(2), 263–294 (1992)

T. Dontchev, E. Farkhi, Stability and Euler approximation of one-sided-Lipschitz differential inclusions. SIAM Journal of Control and Optimization **36**(2), 780–796 (1998)

R. Dzonou, M. Monteiro Marques, A sweeping process approach to inelastic contact problems with general inertia operators. European Journal of Mechanics A, Solids **26**(3), 474–490 (2007)

R. Dzonou, M. Monteiro Marques, L. Paoli, Algorithme de type sweeping process pour un problème de vibro-impact avec un opérateur d'inertie non trivial (sweeping process algorithm for a vibro-impact problem with a nontrivial inertia operator). Comptes Rendus Mécanique **335**(1), 55–60 (2007)

J. Edmond, L. Thibault, Relaxation of an optimal control problem involving a perturbed sweeping process. Mathematical Programming B **104**, 347–373 (2005)

J. Edmond, L. Thibault, BV solutions of nonconvex sweeping process differential inclusion with perturbation. Journal of Differential Equations **226**, 135–179 (2006)

O. Enge, P. Maisser, Modelling electromechanical systems with electrical switching components using the linear complementarity problem. Multibody System Dynamics **13**, 421–445 (2005)

D. Estèvez Schwarz, C. Tischendorf, Structural analysis for electric circuits and consequences for MNA. International Journal of Circuit Theory and Applications **28**, 131–162 (2000)

F. Facchinei, J.S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer Series in Operations Research, vols. I & II (Springer, New York, 2003)

A. Filippov, Differential equations with discontinuous right-hand-side. AMS Transactions **42**, 199–231 (1964)

A.F. Filippov, *Differential Equations with Discontinuous Right Hand Sides* (Kluwer, Dordrecht, 1988)

R. Frasca, M. Camlibel, I. Goknar, L. Ianelli, F. Vasca, State jump rules in linear passive networks with ideal switches. GRACE Report no 460, University of Benevento, 2007

R. Frasca, M. Camlibel, I. Goknar, L. Ianelli, F. Vasca, State discontinuity analysis of linear switched systems via energy function optimization, in *IEEE Int. Symposium on Circuits and Systems, ISCAS2008*, 18–21 May 2008, pp. 540–543

T. Fujisawa, E. Kuh, T. Ohtsuki, A sparse matrix method for analysis of piecewise linear resistive circuits. IEEE Transactions on Circuit Theory **19**(6), 571–584 (1972)

M. Fukushima, Equivalent differentiable optimization problems and descent methods for asymmetric variational inequality problems. Mathematical Programming **53**, 99–110 (1992)

C. Gear, Simultaneous numerical solution of differential-algebraic equations. IEEE Transactions on Circuit Theory **18**(1), 89–95 (1971). ISSN: 0018-9324

C. Glocker, *Set-Valued Force Laws: Dynamics of Non-Smooth Systems*. Lecture Notes in Applied Mechanics, vol. 1 (Springer, Berlin, 2001)

C. Glocker, Models of non-smooth switches in electrical systems. International Journal of Circuit Theory and Applications **33**, 205–234 (2005)

D. Goeleven, Existence and uniqueness for a linear mixed variational inequality arising in electrical circuits with transistors. Journal of Optimization Theory and Applications **138**(3), 397–406 (2008)

D. Goeleven, B. Brogliato, Stability and instability matrices for linear evolution variational inequalities. IEEE Transactions on Automatic Control **49**(4), 521–534 (2004)

G. Golub, C. Van Loan, *Matrix Computations*, 3rd edn. (The Johns Hopkins University Press, Baltimore, 1996)

E. Griepentrog, R. März, *Differential-Algebraic Equations and Their Numerical Treatment* (Teubner, Leipzig, 1986)

M. Günther, U. Feldmann, The DAE-index in electric circuit simulation. Technical Report TUM-M9319, Facultät für Mathematiik. Technische Universität München, 1993

M. Günther, U. Feldmann, E. ter Maten, Modelling and discretization of circuit problems, in *Handbook of Numerical Analysis, Special Volume on Numerical Methods in Electromagnetics*, vol. XIII, ed. by W. Schilders, E. ter Maten (Elsevier, Amsterdam, 2005), pp. 523–659

G. Hachtel, R. Brayton, F. Gustavson, The sparse tableau approach to network analysis and design. IEEE Transactions on Circuit Theory **18**(1), 101–113 (1971). ISSN: 0018-9324

E. Hairer, G. Wanner, *Solving Ordinary Differential Equations II. Stiff and Differential-Algebraic Problems* (Springer, Berlin, 1996)

W. Heemels, B. Brogliato, The complementarity class of hybrid dynamical systems. European Journal of Control **9**, 311–349 (2003)

W. Heemels, M. Camlibel, J. Schumacher, A time-stepping method for relay systems, in *Proceedings of the 39th IEEE Conference on Decision and Control*, Sydney, Australia, December 2000, pp. 4461–4466

W. Heemels, M. Camlibel, J. Schumacher, On event-driven simulation of electrical circuits with ideal diodes. APII Journal Européen des Systèmes Automatisés, Numéro Spécial ADPM **1**, 1–22 (2001)

W. Heemels, M. Camlibel, A. van der Schaft, J. Schumacher, Modelling, well-posedness and stability of switched electrical networks, in *HSCC 2003*, ed. by O. Maler, A. Pnueli. Lecture Notes in Computer Science, vol. 2623 (Springer, Berlin, 2003), pp. 249–266

C. Henry, An existence theorem for a class of differential equations with multivalued right-hand side. Journal of Mathematical Analysis and Applications **41**, 179–186 (1973)

J. Hiriart-Urruty, C. Lemaréchal, *Fundamentals of Convex Analysis* (Springer, Berlin, 2001)

J.-B. Hiriart-Urruty, A. Seeger, A variational approach to copositive matrices. SIAM Review **52**(4) (2010, in press)

C. Ho, A. Ruehli, P. Brennan, The modified nodal approach to network analysis. IEEE Transactions on Circuits and Systems **22**(6), 504–509 (1975). ISSN: 0098-4094

M. Huang, S. Liu, A fully differential comparator-based switched-capacitor $\delta\sigma$ modulator. IEEE Transactions on Circuits and Systems II, Express Briefs **56**(5), 369–373 (2009)

M. Jean, The non smooth contact dynamics method. Computer Methods in Applied Mechanics and Engineering **177**, 235–257 (1999). Special issue on computational modeling of contact and friction, J.A.C. Martins and A. Klarbring, editors

M. Johansson, *Piecewise Linear Control Systems*. Lecture Notes in Control and Information, vol. 284 (Springer, London, 2003)

S. Kang, L. Chua, A global representation of multidimensional piecewise-linear functions with linear partitions. IEEE Transactions on Circuits and Systems **CAS-25**(11), 938–940 (1978)

A. Kastner-Maresch, Implicit Runge-Kutta methods for differential inclusions. Numerical Functional Analysis and Optimization **11**(9–10), 937–958 (1990–1991)

T. Kato, Accretive operators and nonlinear evolution equations in Banach spaces, in *Proceedings of Symposia in Pure Mathematics*. Nonlinear Functional Analysis, vol. 18 (Chicago, 1968), pp. 138–161. Part 1

T. Kevenaar, D. Leenaerts, A comparison of piecewise-linear model description. IEEE Transactions on Circuits and Systems I, Fundamental Theory and Applications **39**(12), 996–1004 (1992)

M. Kunze, M. Monteiro Marquès, An introduction to Moreau's sweeping process, in *Impact in Mechanical Systems: Analysis and Modelling*, ed. by B. Brogliato. Lecture Notes in Physics, vol. 551 (Springer, Berlin, 2000), pp. 1–60

D. Leenaerts, On linear dynamic complementarity systems. IEEE Transactions on Circuits and Systems I, Fundamental Theory and Applications **46**(8), 1022–1026 (1999)

D. Leenaerts, W. Van Bokhoven, *Piecewise Linear Modeling and Analysis* (Kluwer Academic, Norwell, 1998). ISBN: 0792381904

R. Leine, N. van de Wouw, *Stability and Convergence of Mechanical Systems with Unilateral Constraints*. Lecture Notes in Applied and Computational Mechanics, vol. 36 (Springer, Berlin, 2008)

C. Lin, Q.G. Wang, On uniqueness of solutions to relay feedback systems. Automatica **38**, 177–180 (2002)

C. Liu, J. Hsieh, C. Chang, J. Bocek, Y. Hsiao, A fast-decoupled method for time-domain simulation of power converters. IEEE Transactions on Power Electronics **8**(1), 37–45 (1993)

Y. Lootsma, A. van der Schaft, M. Camlibel, Uniqueness of solutions of linear relay systems. Automatica **35**(3), 467–478 (1999)

T. Lukl, J. Vrana, J. Misurec, Scisip—program for switched circuit analysis in matlab, in *IEEE International Behavioral Modeling and Simulation Workshop*, San Jose, California, 2006, pp. 61–66

M. Mabrouk, A unified variational for the dynamics of perfect unilateral constraints. European Journal of Mechanics A, Solids **17**, 819–842 (1998)

P. Maffezzoni, L. Codecasa, D. D'Amore, Event-driven time-domain simulation of closed-loop switched circuits. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems **25**(11), 2413–2426 (2006)

D. Maksimovic, A. Stankovic, V. Thottuvelil, G. Verghese, Modeling and simulation of power electronic converters. Proceedings of the IEEE **89**(6), 898–912 (2001)

R. März, Numerical methods for differential algebraic equations. Acta Numerica **1**, 141–198 (1992)

R. März, C. Tischendorf, Recent results in solving index 2 differential algebraic equations in circuit simulation. SIAM Journal on Scientific and Statistical Computing **18**(1), 139–159 (1997)

K. Mayaram, D. Lee, D. Moinian, J. Roychowdhury, Computer-aided circuit analysis tools for RFIC simulation: algorithms, features, and limitations. IEEE Transactions on Circuits and Systems II, Analog and Digital Signal Processing **47**(4), 274–286 (2000)

R. Middlebrook, S. Ćuk, A general unified approach to modelling switching—converter power stages, in *IEEE Power Electronics Specialists Conference*, Cleveland, OH, 8–10 June 1976

M. Moeller, C. Glocker, Non-smooth modelling of electrical systems using the flux approach. Nonlinear Dynamics **50**, 273–295 (2007)

M. Monteiro Marques, Chocs inélastiques standards: un résultat d'existence, in *Séminaire d'Analyse Convexe, Exposé no 4*, vol. 15, USTL, Montpellier, France, 1985

M. Monteiro Marques, *Differential Inclusions in Nonsmooth Mechanical Problems. Shocks and Dry Friction*. Progress in Nonlinear Differential Equations and Their Applications, vol. 9 (Birkhäuser, Basel, 1993)

B. Mordukhovich, Generalized differential calculus for nonsmooth and set-valued analysis. Journal of Mathematical Analysis and Applications **183**, 250–288 (1994)

J. Moreau, La notion de surpotentiel et les liaisons unilatérales en élastoplastique. Comptes Rendus de l'Académie des Sciences **267a**, 954–957 (1968)

J. Moreau, Rafle par un convexe variable (première partie), exposé no 15, in *Séminaire d'Analyse Convexe*, University of Montpellier, 1971, p. 43

J. Moreau, Rafle par un convexe variable (deuxième partie) exposé no 3, in *Séminaire d'Analyse Convexe*, University of Montpellier, 1972, p. 36

J. Moreau, Problème d'évolution associé à un convexe mobile d'un espace hilbertien. Comptes Rendus de L'Académie des Sciences Paris, Série A–B **276**, 791–794 (1973)

J. Moreau, Evolution problem associated with a moving convex set in a Hilbert space. Journal of Differential Equations **26**, 347–374 (1977)

J. Moreau, Bounded variation in time, in *Topics in Nonsmooth Mechanics*, ed. by J. Moreau, P. Panagiotopoulos, G. Strangus (Birkhäuser, Basel, 1988), pp. 1–74

K. Murty, *Linear Complementarity, Linear and Nonlinear Programming* (Heldermann, Berlin, 1988)

A. Opal, Sampled data simulation of linear and nonlinear circuits. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems **15**(3), 295–307 (1996)

M. Parodi, M. Storace, P. Julian, Synthesis of multiport resistors with piecewise-linear characteristics: a mixed-signal architecture. International Journal of Circuit Theory and Applications **33**(4), 307–319 (2005)

E. Pratt, A. Léger, M. Jean, Critical oscillations of mass-spring systems due to nonsmooth friction. Archives of Applied Mechanics **78**(2), 89–104 (2008)

G. Reissig, U. Feldmann, Computing the generic index of the circuit equations of linear active networks, in *IEEE Int. Symp. on Circuits and Systems (ISCAS)*, 1996, pp. 190–193

L. Repetto, M. Parodi, M. Storace, A procedure for the computation of accurate pwl-approximations of non-linear dynamical systems. International Journal of Circuit Theory and Applications **34**(2), 237–248 (2006)

S. Robinson, Generalized equations and their solutions. I. Basic theory. Mathematical Programming Study **10**, 128–141 (1979)

R. Rockafellar, *Convex Analysis* (Princeton University Press, Princeton, 1970)

R. Rockafellar, R. Wets, *Variational Analysis*, vol. 317 (Springer, New York, 1997)

H. Samelson, R. Thrall, O. Wesler, A partition theorem for Euclidean n-space. Proc. Am. Math. Soc. (1958)

R. Sargent, An efficient implementation of the Lemke algorithm and its extension to deal with upper an lower bounds. Mathematical Programming Study **7**, 36–54 (1978)

L. Schwartz, *Analyse III, Calcul Intégral* (Hermann, Paris, 1993)

S. Seshu, M. Reed, *Linear Graphs and Electrical Networks* (Addison-Wesley, Reading, 1961)

J. Shen, J. Pang, Semicopositive linear complementarity systems. International Journal of Robust and Nonlinear Control **17**, 1367–1386 (2007)

G. Shilov, B. Gurevich, *Integral Measure and Derivative. A Unified Approach* (Prentice-Hall, Englewood Cliffs, 1966). Hermann, Paris, 1993

G. Smirnov, *Introduction to the Theory of Differential Inclusions*. Graduate Studies in Mathematics, vol. 41 (American Mathematical Society, Providence, 2002)

S. Stevens, P. Lin, Analysis of piecewise-linear resistive networks using complementarity pivot theory. IEEE Transactions on Circuits and Systems **CAS-28**(5), 429–441 (1981)

D. Stewart, A high accuracy method for solving ODEs with discontinuous right-hand-side. Numerische Mathematik **58**, 299–328 (1990)

D. Stewart, A numerical method for friction problems with multiple contacts. Journal of Australian Mathematical Society, Series B **37**, 288–308 (1996)

C. Studer, *Numerics of Unilateral Contacts and Friction. Modeling and Numerical Time Integration in Non-Smooth Dynamics*. Lecture Notes in Applied and Computational Mechanics, vol. 47 (Springer, Berlin, 2009)

C. Tischendorf, Topological index calculation of DAE in circuit simulation. Surveys on Mathematics for Industry **8**(3–4), 187–189 (1999)

V. Utkin, J. Guldner, J. Shi, *Sliding Mode Control in Electro-Mechanical Systems*. Automation and Control Engineering (CRC Press, Boca Raton, 2009)

J. Valsa, J. Vlach, Swann—a programme for analysis of switched analogue non-linear networks. International Journal of Circuit Theory and Applications **23**, 369–379 (1995)

W. van Bokhoven, Piecewise linear modelling and analysis. PhD thesis, Technical University of Eindhoven, TU/e, 1981. Available at alexandria.tue.nl/extra3/proefschrift/PRF3B/8105755.pdf

W. van Bokhoven, J. Jess, Some new aspects of P and P_0 matrices and their application to networks with ideal diodes, in *IEEE Int. Symp. Circuits and Systems*, 1978, pp. 806–810

A. van der Schaft, J. Schumacher, Complementarity modeling of hybrid systems. IEEE Transactions on Automatic Control **43**(4), 483–490 (1998)

W. van Eijndhoven, A piecewise linear simulator for large scale integrated circuits. PhD thesis, Technical University of Eindhoven, TU/e, 1984

M.T. van Stiphout, Plato—a piecewise linear analysis for mixed-level circuit simulation, PhD thesis, Technical University of Eindhoven, TU/e, 1990

L. Vandenberghe, B. De Moor, J. Vandewalle, The generalized linear complementarity problem applied to the complete analysis of resistive piecewise-linear circuits. IEEE Transactions on Circuits and Systems **36**(11), 1382–1391 (1989)

F. Vasca, L. Iannelli, M. Camlibel, R. Frasca, A new perspective for modelling power electronics converters: complementarity framework. IEEE Transactions on Power Electronics **24**(2), 456–468 (2009)

J. Vlach, A. Opal, Modern CAD methods for analysis of switched networks. IEEE Transactions on Circuits and Systems I, Fundamental Theory and Applications **44**(8), 759–762 (1997)

J. Vlach, K. Singhai, M. Vlach, Computer oriented formulation of equations and analysis of switched-capacitor networks. IEEE Transactions on Circuits and Systems **31**(9), 753–765 (1984)

J. Vlach, J. Wojciechowski, A. Opal, Analysis of nonlinear networks with inconsistent initial conditions. IEEE Transactions on Circuits and Systems I, Fundamental Theory and Applications **42**(4), 195–200 (1995)

Y. Wang, S. Joeres, R. Wunderlich, S. Heinen, Modeling approaches for functional verification of RF-socs: limits and future requirements. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems **28**(5), 769–773 (2009)

C. Wen, S. Wang, H. Zhang, M. Khan, A novel compact piecewise-linear representation. International Journal of Circuit Theory and Applications **33**(1), 87–97 (2005)

K. Yamamura, A. Machida, An efficient algorithm for finding all dc solutions of piecewise-linear circuits. International Journal of Circuit Theory and Applications **36**(8), 989–1000 (2008)

X. Yu, O. Kaynak, Sliding-mode control with soft computing: a survey. IEEE Transactions on Industrial Electronics **56**(9), 3275–3285 (2009)

F. Yuan, A. Opal, Computer methods for switched circuits. IEEE Transactions on Circuits and Systems I, Fundamental Theory and Applications **50**(8), 1013–1024 (2003)

D. Zhu, P. Marcotte, Modified descents methods for solving the monotone variational inequality problem. Operations Research Letters **14**, 111–120 (1993)

D. Zhu, P. Marcotte, An extended descent framework for monotone variational inequalities. Journal of Optimization Theory and Applications **80**, 349–366 (1994)

L. Zhu, J. Vlach, Analysis and steady state of nonlinear networks with ideal switches. IEEE Transactions on Circuits and Systems I, Fundamental Theory and Applications **42**(4), 212–214 (1995)

Index

V. Acary et al., *Nonsmooth Modeling and Simulation for Switched Circuits,*
Lecture Notes in Electrical Engineering 69,
DOI 10.1007/978-90-481-9681-4, © Springer Science+Business Media B.V. 2011