

## 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.

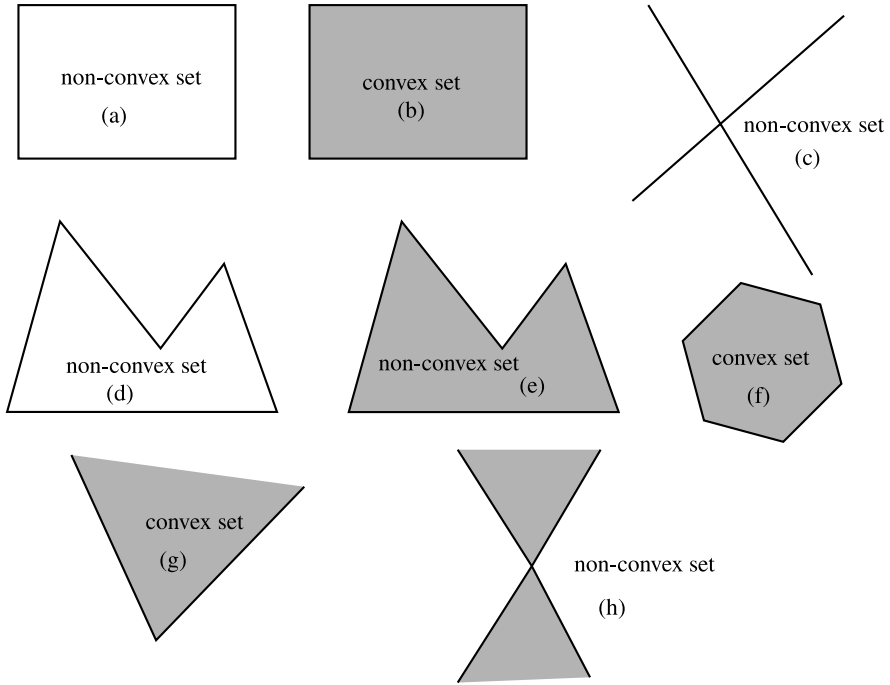
### 2.1 Basics from Convex and Nonsmooth Analysis

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  $\lambda \in (0, 1)$ .



**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  $Ax \geq 0$  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

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

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  $C^\circ$  is its orthogonal subspace. The polar cone obtained from  $C$  depends on the scalar product that is used in the definition: changing the scalar

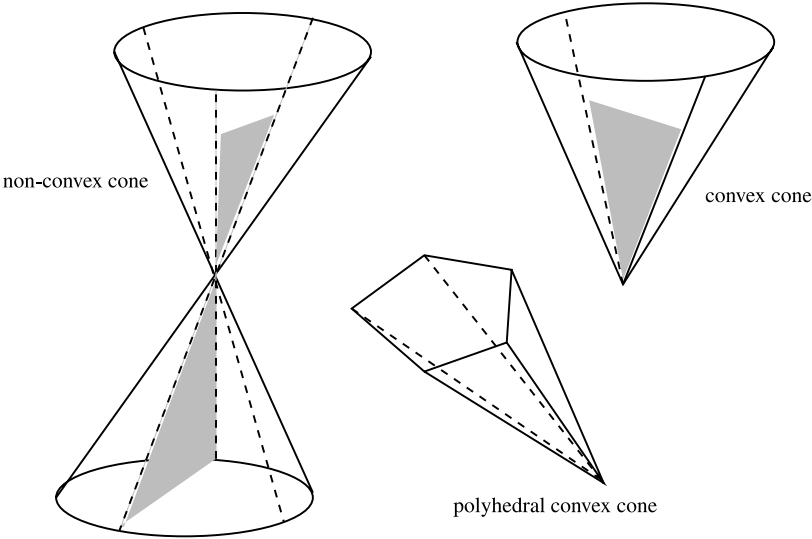
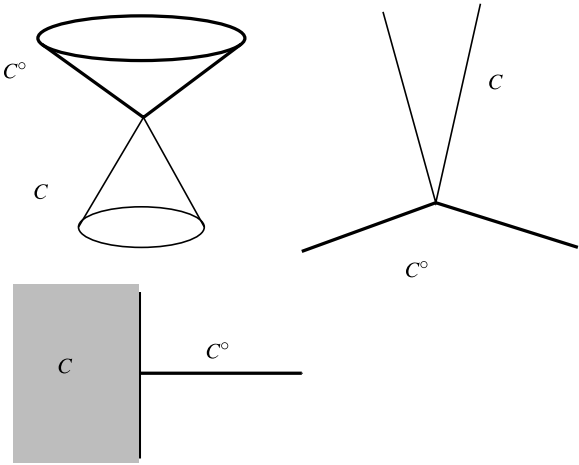


Fig. 2.2 Convex and non-convex cones

Fig. 2.3 Convex cones and their polar cones



product changes  $C^\circ$ . When  $C$  is a non empty closed convex cone, then  $C^\circ$  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  $C^* = \{s \in \mathbb{R}^n \mid \langle s, x \rangle \geq 0 \text{ for all } x \in C\}$ . 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  $C^* = \{s \in \mathbb{R}^n \mid s^T y \geq 0 \text{ for all } y \in C\}$ . This is indeed a cone as can be checked.

An interesting result is the next one:

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

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  $f : C \rightarrow \mathbb{R}$  is said convex on  $C$  when, for all pairs  $(x, y) \in C \times C$  and all  $\lambda \in (0, 1)$ , it holds that:

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

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  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  with a linear mapping  $A : \mathbb{R}^m \rightarrow \mathbb{R}^n$ , denoted as  $(f \circ A)(\cdot) = f(A(\cdot))$ , is again convex. The domain of a function  $f(\cdot)$  is defined as  $\text{dom}(f) = \{x \in \mathbb{R}^n \mid f(x) < +\infty\}$ , 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  $f : U \rightarrow \mathbb{R}$  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  $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$  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  $f : U \rightarrow \mathbb{R}$  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.*  $\langle \nabla^2 f(x)y, y \rangle \geq 0$  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  $f : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto |x|$ , 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  $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$  be a proper convex function. The *conjugate* of  $f(\cdot)$  is the function  $f^*(\cdot)$  defined by:

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

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.

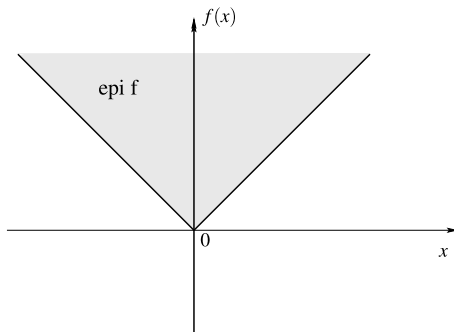
*Example 2.11* Let us compute the conjugate function  $g(y) = f^*(y)$  of the absolute value function  $f(x) = |x|$ . We get:

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

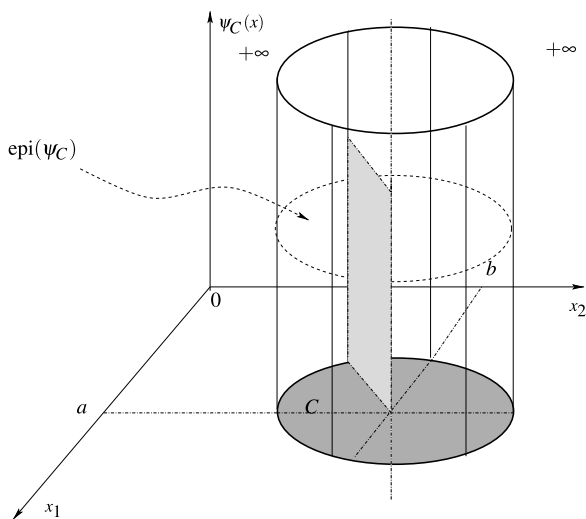
If  $x > 0$  then  $g(y) = \sup_{x \in \mathbb{R}} x(y - 1)$ . So if  $y > 1$  one obtains  $g(y) = +\infty$ , and if  $y \leq 1$  one obtains  $g(y) = 0$ . If  $x < 0$  then  $g(y) = \sup_{x \in \mathbb{R}} x(y + 1)$ . So if  $y \geq -1$  one obtains  $g(y) = 0$ , and if  $y < -1$  one obtains  $g(y) = +\infty$ . 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  $f : \mathbb{R}^n \rightarrow \mathbb{R}, x \mapsto \|x\|$  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  $f : \mathbb{R} \rightarrow \mathbb{R}, x \mapsto ax$  if  $x \leq 0, x \mapsto bx$  if  $x \geq 0$ . Then  $f^*(y) = \psi_{[a, b]}(y)$ .

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

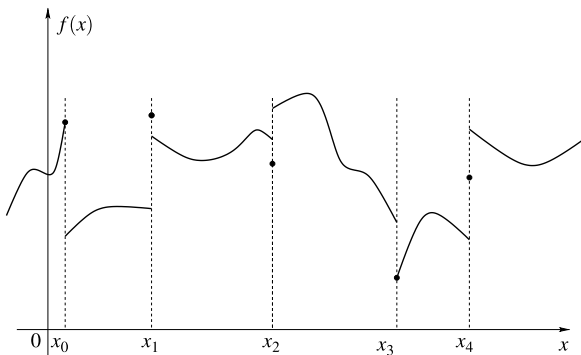


Let us now introduce a notion that is useful to characterize the convexity of a function, and which also permits to link convex functions and convex sets.

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

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

Notice that  $\eta$  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  $\text{epi}(|x|) = \{(x, \eta) \in \mathbb{R} \times \mathbb{R} \mid \eta \geq |x|\} \subset \mathbb{R}^2$ . Consider now the set  $C = \{x \in \mathbb{R}^2 \mid (x_1 - a)^2 + (x_2 - b)^2 \leq r^2\}$  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:  $\text{epi}(\psi_C) = \{(x, \eta) \in C \times \mathbb{R} \mid \eta \geq 0\}$ . 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  $\liminf$  means the lower limit. Given a subset  $S \subseteq \mathbb{R}^n$ ,  $l = \liminf_{y \rightarrow x} f(y)$  for  $x \in \text{cl}S$  means that:<sup>1</sup> for all  $\epsilon > 0$ , there exists a neighborhood  $N(x)$  such that  $f(y) \geq l - \epsilon$  for all  $y \in N(x)$ , and in any neighborhood  $N(x)$ , there is  $y \in N(x)$  such that  $f(y) \leq l + \epsilon$ .

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

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.

<sup>1</sup> $\text{cl}S$  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  $\text{epi}(f)$  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.  $\text{dom}(f) = \mathbb{R}^n$ ) 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:

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

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

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

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

$$T_C(x) = \text{cl} \left( \bigcup_{y \in C} \bigcup_{\lambda > 0} \lambda(y - x) \right),$$

where  $\text{cl}(\cdot)$  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.

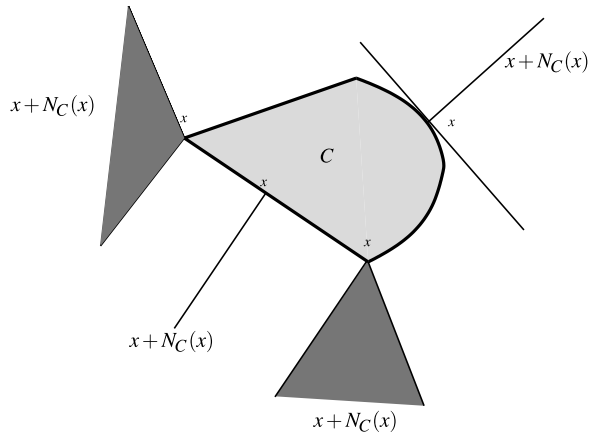
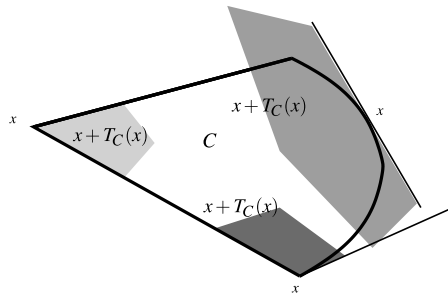
**Proposition 2.19** Let  $C \subset \mathbb{R}^n$  be a closed non empty convex set and let  $x \in C$ . Then the tangent and normal cones are closed convex cones, and  $N_C(x) = (T_C(x))^\circ$  and  $T_C(x) = (N_C(x))^\circ$ .

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

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

which is also a variational definition of the tangent cone. One finds that when  $x \in \text{Int}(C)$ , 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  $C = \{x \in \mathbb{R}^n \mid Ex + F \leq 0, E \in \mathbb{R}^{m \times n}, F \in \mathbb{R}^m\}$ . In other words  $C$  is defined with  $m$  inequalities  $E_i x + F_i \leq 0$  where the  $m$  vectors  $E_i \in \mathbb{R}^{1 \times n}$  are the rows of the matrix  $E$  and the  $F_i$ s 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:

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

and

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

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 usually 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:

$$f(y) - f(x) \geq \gamma^T(y - x) \quad (2.6)$$

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  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  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  $x \in \text{Int}(\text{dom}(f))$ , then  $\partial f(x) \neq \emptyset$ . In particular, if  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  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  $\gamma$  such that  $|y| \geq \gamma y$  for all reals  $y$ . If  $y > 0$  one finds  $\gamma \leq 1$ . If  $y < 0$  then one finds  $\gamma \geq -1$ . One concludes that  $-1 \leq \gamma \leq 1$ . 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  $\text{Int}(C)$  contains an  $n$ -dimensional ball of radius  $r > 0$ . Then the subgradients of the indicator function of  $C$  at  $x$  are the vectors  $\gamma$  satisfying  $\psi_C(y) - \psi_C(x) \geq \gamma^T(y - x)$  for all  $y \in \mathbb{R}^n$ . Let  $x \in \text{Int}(C)$ . We get  $\psi_C(y) \geq \gamma^T(y - x)$ . Let  $y \in \text{Int}(C)$ , so that  $0 \geq \gamma^T(y - x)$  for all  $y \in \text{Int}(C)$ . In view of the assumptions on  $x$  and  $C$  there exists a ball of positive radius centered at  $x$ , contained in  $\text{Int}(C)$ . 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  $x \in \text{Int}(C)$ . Let now  $x \notin C$ , so that  $\psi_C(y) \geq +\infty + \gamma^T(y - x)$  for all  $y \in \mathbb{R}^n$ . Take for instance  $y \in C$  so that we get  $\gamma^T(x - y) \geq +\infty$ . This is impossible and we conclude that  $\partial\psi_C(x) = \emptyset$  when  $x \notin C$ . Let now  $x \in \text{Bd}(C)$ , the boundary of the set  $C$ . We get  $\psi_C(y) \geq \gamma^T(y - x)$  for all  $y \in \mathbb{R}^n$ . Take  $y \in C$ , then the subgradients have to satisfy  $\gamma^T(y - x) \leq 0$  for all  $y \in C$ . Precisely, such vectors  $\gamma$  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  $\partial\psi_C(\cdot) = N_C(\cdot)$ .

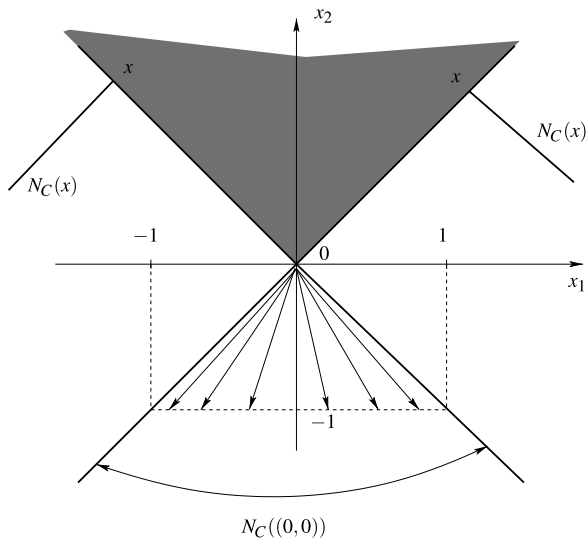
**Example 2.27** (Normal cone to a finitely represented set) If  $C$  is finitely represented, i.e.  $C = \{x \in \mathbb{R}^n \mid g(x) \leq 0\}$ , with  $g(\cdot)$  lower semi-continuous, proper, and convex such that  $0 \notin \partial g(x)$ , then:

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

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  $\{\lambda \eta \mid \lambda > 0 \text{ and } \eta \in \partial g(x)\}$ . One can say that on  $\text{Bd}(C)$  the normal cone is generated by the subgradients of the function  $g(\cdot)$ . Consider for instance the set of  $\mathbb{R}^2$  defined as  $C = \{(x_1, x_2) \mid x_2 \geq |x_1|\}$ . Thus  $g(x) = |x_1| - x_2$ , and there is a corner at  $x_1 = x_2 = 0$ . One has  $\partial g(0, 0) = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$ . 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  $f : \mathbb{R} \rightarrow \mathbb{R}$  with  $f(x) = -1$  if  $x < 0$ ,  $f(x) = 1$  if  $x > 0$ , and  $f(0)$  can be given any bounded value. The distribution is then defined as  $\langle T, \varphi \rangle = \int_{\text{dom}(\varphi)} f(t) \varphi(t) dt$ . 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  $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$  be a convex lower semi-continuous function, and  $A : \mathbb{R}^m \rightarrow \mathbb{R}^n$  be an affine mapping.<sup>2</sup> 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  $f \circ A : \mathbb{R}^m \rightarrow \mathbb{R} \cup \{+\infty\}$  is given by*

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

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

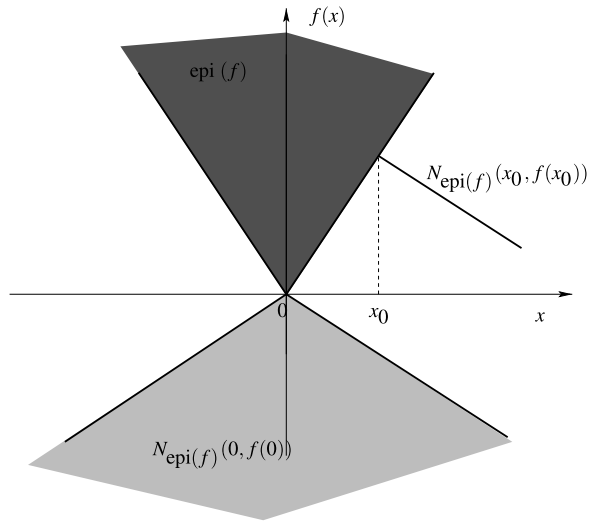
**Theorem 2.30** (Moreau-Rockafellar: subdifferentiate of a sum) *Let  $f_i : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ ,  $1 \leq i \leq 2$  be proper convex functions, and let  $f(\cdot) = \sum_{i=1}^2 f_i(\cdot)$ . Assume that the convex sets  $\text{dom}(f_i)$ ,  $1 \leq i \leq 2$ , have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

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

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

<sup>2</sup>I.e.  $Ax = A_0x + b$  with  $A_0$  linear.

**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:

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

The normal cone to the epigraph is therefore generated by the vectors  $(\gamma, -1)$  where  $\gamma$  is a subgradient. Normal cones to the epigraph of the function  $f(x) = ax$  if  $x \leq 0$  and  $f(x) = bx$  if  $x \geq 0$  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:

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

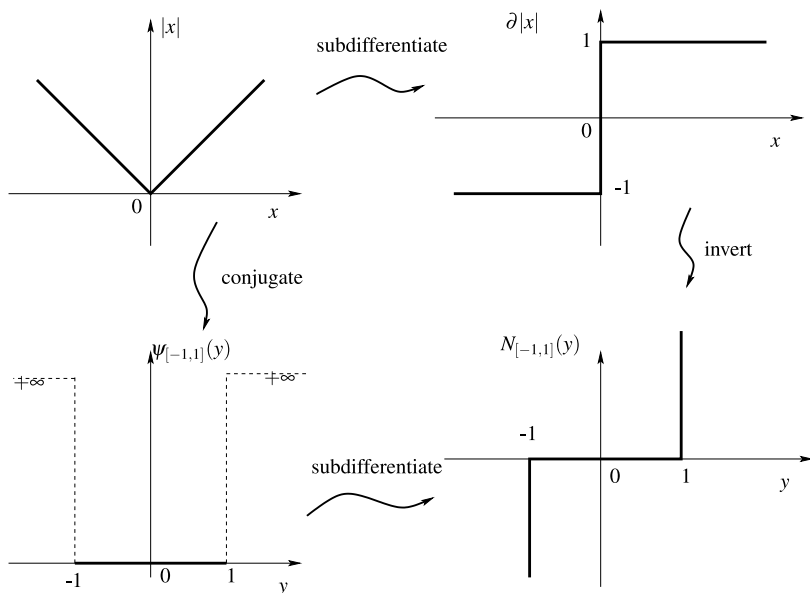
This is illustrated in Fig. 2.11 for the absolute value function (see Examples 2.11, 2.25 and 2.26). In particular one has  $N_{[-1,1]}(1) = \mathbb{R}_+$  and  $N_{[-1,1]}(-1) = \mathbb{R}_-$ . Inversion of graphs occurs when passing from  $(i(t), v(t))$  to  $(v(t), i(t))$  characteristics of electronic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

$$0 \in \partial f(x) + \partial \psi_C(x) = \partial f(x) + N_C(x). \quad (2.10)$$

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 fulfilled 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*  $F^{-1} : Y \rightarrow X$  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:

- $F : \mathbb{R} \rightarrow \mathbb{R}$ ,  $x \mapsto [-1, 1]$ , which assigns to each  $x$  an interval, see Fig. 2.12(a).
- $F : \mathbb{R} \rightarrow \mathbb{R}$ ,  $x \mapsto [-|x|, |x|]$ , 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

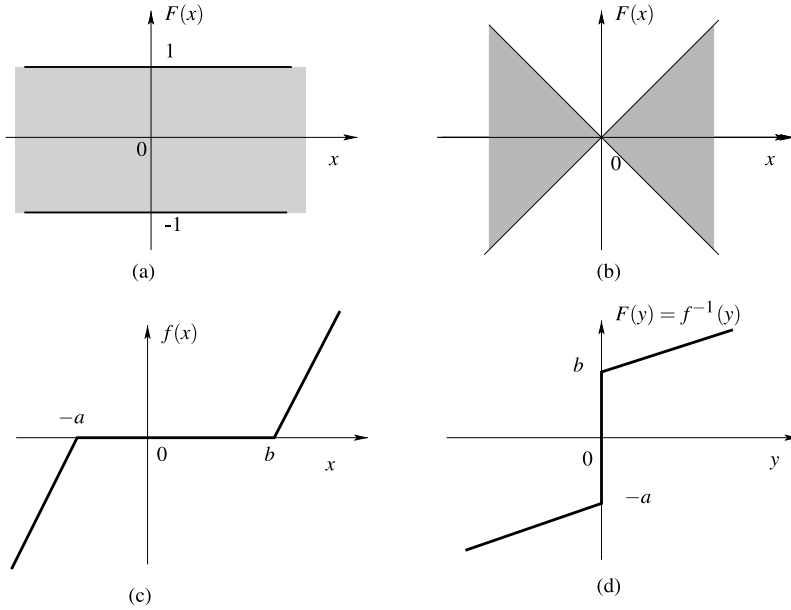
**Definition 2.33** (Maximal monotone mapping) A multivalued mapping  $F : S \subseteq \mathbb{R}^n \rightarrow \mathbb{R}^n$  is said to be *monotone* on  $S$  if for every pairs  $(x_1, y_1)$  and  $(x_2, y_2)$  in its graph one has:

$$\langle x_1 - x_2, y_1 - y_2 \rangle \geq 0. \quad (2.11)$$

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:

$$\langle x_1 - x_2, y_1 - y_2 \rangle \geq c \|x_1 - x_2\|^\xi. \quad (2.12)$$

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.



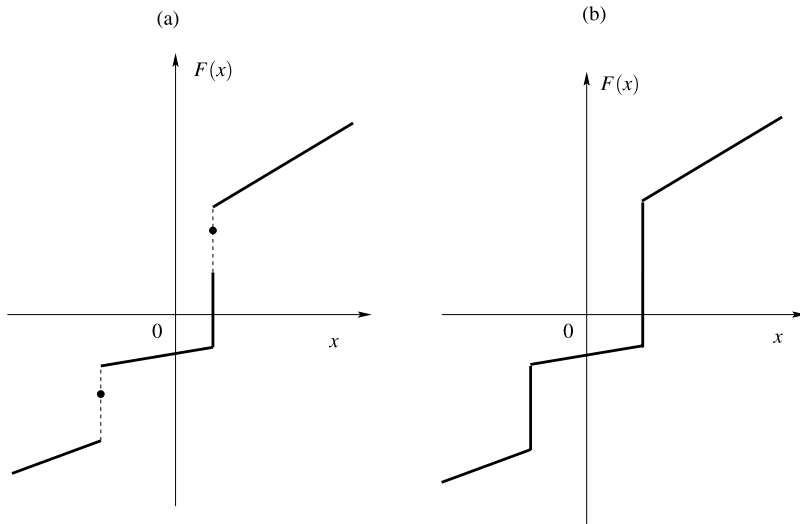
**Fig. 2.12** Multivalued functions

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  $(x, y) \in (\mathbb{R}^n \times \mathbb{R}^n) \setminus \text{gr}(F)$  there exists  $(x', y') \in \text{gr}(F)$  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  $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$  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  $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$  is monotone then  $\lambda F(\cdot)$  is monotone for any  $\lambda > 0$ .
- If  $F_1 : \mathbb{R}^n \rightarrow \mathbb{R}^n$  and  $F_2 : \mathbb{R}^n \rightarrow \mathbb{R}^n$  are monotone, then  $(F_1 + F_2)(\cdot)$  is monotone.
- $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$  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  $\mathbb{R} \rightarrow \mathbb{R}$  is differentiable, then its gradient is non decreasing.

**Theorem 2.34** *Let  $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$  be convex and proper. Then the multivalued mapping  $\partial f : \mathbb{R}^n \rightarrow \mathbb{R}^n$  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:

$$0 \in F(x) + N_C(x), \quad (2.13)$$

where  $C \subseteq \mathbb{R}^n$  and  $F : C \rightarrow \mathbb{R}^n$  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  $\xi$ -monotone on  $C$  for some  $\xi > 1$ , the generalized equation in (2.13) has a unique solution.*

**Example 2.36** Let  $F : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ ,  $x \mapsto \begin{pmatrix} \cos x \\ \sin x \end{pmatrix}$ , and  $C = \{(x_1, x_2) \in \mathbb{R}^2 \mid x_1 \geq 0\}$ . We know from (2.5) that  $N_C(0) = \{\alpha \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \alpha \geq 0\} = \mathbb{R}_- \mathbf{e}_1$  where  $\mathbf{e}_1 = (1 \ 0)^T$ . We deduce that all  $x$  with  $x_1 = 2k\pi$ ,  $k \geq 0$ , 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*

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

where  $\operatorname{proj}_M$  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  $f_i : \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $1 \leq i \leq m$ , such that  $\mathbb{R}^n \ni C = \{x \in \mathbb{R}^n \mid f_i(x) \leq 0, 1 \leq i \leq m\}$ . 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:

$$\begin{cases} z \geq 0, \\ w = Mz + q \geq 0, \\ w^T z = 0, \end{cases} \quad (2.15)$$

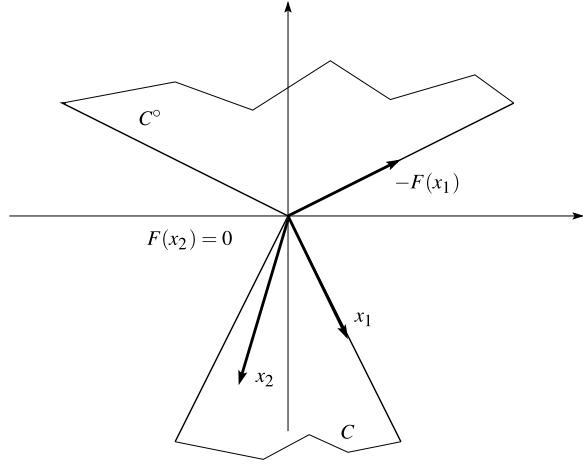
where  $z$  is the unknown of the LCP.

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

$$0 \leq w \perp z \geq 0. \quad (2.16)$$

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  $i \in \{1, \dots, n\}$ .

**Fig. 2.14** Cone complementarity problem



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

$$0 \leq x \perp F(x) \geq 0 \quad (2.17)$$

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  $F : C \rightarrow \mathbb{R}^n$  a mapping. A *Cone Complementarity Problem* (CCP) is a problem of the form:

$$C \ni x \perp F(x) \in C^* \quad (2.18)$$

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  $\text{MLCP}(A, B, C, D, a, b)$  consists in finding two vectors  $u \in \mathbb{R}^n$  and  $v \in \mathbb{R}^m$  such that

$$\begin{cases} Au + Cv + a = 0, \\ 0 \leq v \perp Du + Bv + b \geq 0. \end{cases} \quad (2.19)$$

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

$$\begin{cases} w = Mz + q, \\ w_i = 0, & \forall i \in \mathcal{E}, \\ 0 \leq z_i \perp w_i \geq 0, & \forall i \in \mathcal{J}, \end{cases} \quad (2.20)$$

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

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  $z = \begin{pmatrix} u \\ v \end{pmatrix}$ ,  $M = \begin{pmatrix} A & C \\ D & B \end{pmatrix}$ ,  $q = \begin{pmatrix} a \\ b \end{pmatrix}$ .

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

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

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

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

Note that the problem (2.21) implies that

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

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

$$0 \leq \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \perp \begin{pmatrix} 0 \\ -1 \end{pmatrix} + \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \geq 0$$

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

$$0 \leq x \perp -x + q \geq 0$$

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  $-1 \geq 0$ . 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  $0 \leq x \perp Mx + q \geq 0$  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.<sup>3</sup> 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  $x^T Mx \geq 0$  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  $\begin{pmatrix} 1 & -1 \\ 1 & 0 \end{pmatrix}$  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  $\mathbb{R}_+^4$ .

<sup>3</sup>If  $A$  is an  $m \times n$  matrix,  $I$  is a subset of  $\{1, \dots, m\}$  with  $k$  elements and  $J$  is a subset of  $\{1, \dots, n\}$  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  $[A]_{I,I}$  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  $G^T M G$  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  $a_{11} \geq 0$ ,  $a_{22} \geq 0$ ,  $a_{12} + \sqrt{a_{11}a_{22}} \geq 0$ . Let  $M = M^T \in \mathbb{R}^{3 \times 3}$ . Then  $M$  is copositive on  $\mathbb{R}_+^3$  if and only if  $a_{11} \geq 0$ ,  $a_{22} \geq 0$ ,  $a_{33} \geq 0$ ,  $b_{12} \triangleq a_{12} + \sqrt{a_{11}a_{22}} \geq 0$ ,  $b_{13} \triangleq a_{13} + \sqrt{a_{11}a_{33}} \geq 0$ ,  $b_{23} \triangleq a_{23} + \sqrt{a_{22}a_{33}} \geq 0$ , and*

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

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  $x^* \geq 0$  and  $Mx^* + q \geq 0$ . 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  $0 \leq x \perp Mx \geq 0$  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  $0 \leq x \perp Mx \geq 0$  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  $0 \leq x \perp y \geq 0$  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  $x \geq 0$ . 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  $0 \leq x \perp y \geq 0$ . Conversely let  $0 \leq x \perp y \geq 0$ . If  $y > 0$  then  $x = 0$ . If  $y = 0$  then  $x \geq 0$  so that  $-xz \leq xy$  for any  $z \geq 0$ . If  $y > 0$  then  $x = 0$  so that  $xz = 0 \leq xy = 0$  for any  $z \geq 0$ . In any case the scalar  $s \triangleq -x$  satisfies  $s(z - y) \leq 0$  for all  $z \geq 0$ , which precisely means that  $s \in N_C(y)$ , see Definition 2.18. We have shown that for  $x \in \mathbb{R}$  and  $y \in \mathbb{R}$

$$0 \leq x \perp y \geq 0 \quad \Leftrightarrow \quad -x \in N_C(y). \quad (2.23)$$

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:*

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

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:

$$\begin{aligned} 0 \leq x \perp Mx + q &\geq 0 \\ &\Leftrightarrow \\ -Mx - q &\in N_{\mathbb{R}_+^n}(x) \\ &\Leftrightarrow \\ -x - M^{-1}q &\in M^{-1}N_{\mathbb{R}_+^n}(x), \\ x &= \text{proj}_M(\mathbb{R}_+^n; -M^{-1}q) \end{aligned} \quad (2.25)$$

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:

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

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  $\mathbb{R}^n \rightarrow \mathbb{R}$ . Then for each  $y \in \mathbb{R}^n$  there exists a unique  $x \triangleq P_\phi(y) \in \mathbb{R}^n$  such that

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

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

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

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:

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

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:

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

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

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

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:

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

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

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

where  $x : \mathbb{R} \rightarrow \mathbb{R}^n$  is a function of time  $t$ ,  $\dot{x} : \mathbb{R} \rightarrow \mathbb{R}^n$  is its time derivative,  $F : \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}^n$  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

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

where  $C : [0, T] \rightarrow \mathbb{R}^n$  is a moving set. A function  $x : [0, T] \rightarrow \mathbb{R}^n$  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:

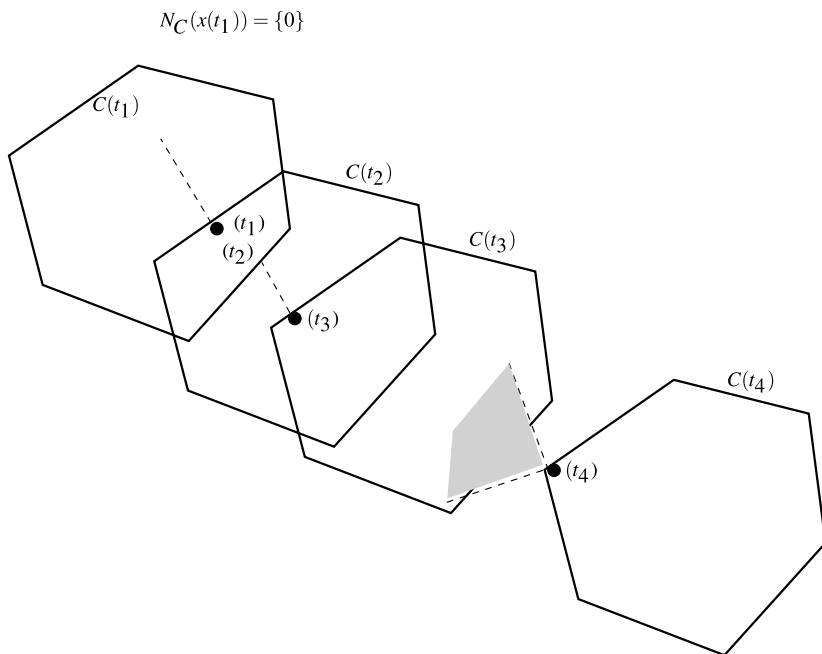
$$\begin{aligned} -\dot{x}(t) &\in N_{C(t)}(x(t) + f(t, x(t))), \\ &\text{almost everywhere on } [0, T], \quad x(0) \in C(0). \end{aligned} \quad (2.35)$$

**Remark 2.54** Why the name *sweeping process*? When  $x(t) \in \text{Int}(C(t))$ , where  $\text{Int}$  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  $0 \leq a \leq b \leq T$ .

**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:  $\|f(t, x) - f(t, y)\| \leq k_\eta(t) \|x - y\|$ ;*



**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  $\|f(t, x)\| \leq \beta(t)(1 + \|x\|)$ .

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.<sup>4</sup> 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  $[0, T]$ .<sup>5</sup> Suppose there exists some non

<sup>4</sup>It is a fact that the solutions functional set is a copy of the multifunction  $C(t)$  functional set.

<sup>5</sup>See Sect. A.4.

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

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

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.  $\lambda$  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  $f_i(q) \geq 0$ ,  $1 \leq i \leq m$ . 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:  $\Phi = \{q \in \mathbb{R}^n \mid f_i(q) \geq 0, 1 \leq i \leq m\}$ , 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  $\Phi$  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:

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

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

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

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:

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

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  $\frac{1}{1+e}$ :

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

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

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

that is:

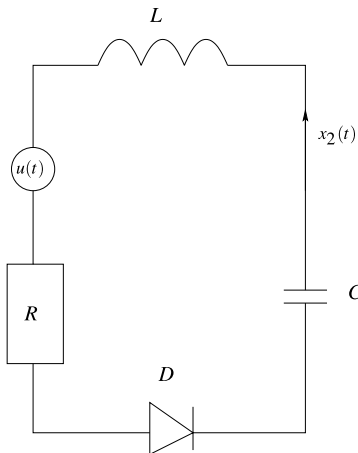
$$v(t^+) = -ev(t^-) + (1+e)\text{proj}_{M(q(t))}(T_\Phi(q(t)); v(t^-)), \quad (2.42)$$

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  $\Phi$ ).

When  $q(t) \in \text{Int}(\Phi)$ , 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  $\Phi$ , and if  $v(t^-)$  belongs to the interior of  $T_\Phi(q(t))$ , then from (2.42) we get  $\text{proj}_{M(q(t))}(T_\Phi(q(t)); v(t^-)) = v(t^-)$  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,  $\lambda(t)$ :

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

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

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

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  $\lambda$  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  $\{x \in \mathbb{R}^4 \mid x_2 \leq 0\}$  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:

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

for some convex, proper and lower semi-continuous function  $\varphi : \mathbb{R}^n \rightarrow \mathbb{R}$ . 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:

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

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:

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

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:

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

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  $x(t) \in \mathbb{R}^n$ :

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

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  $\text{dom}(A)$ , i.e., for all  $x \in \text{dom}(A)$ ,  $y \in \text{dom}(A)$  and all  $x' \in A(x)$ ,  $y' \in A(y)$ , one has

$$(x' - y')^T (x - y) \geq 0. \quad (2.50)$$

- (ii) There exists  $L \geq 0$  such that for all  $t \in [0, T]$ , for all  $x_1, x_2 \in \mathbb{R}^n$ , one has  $\|f(t, x_1) - f(t, x_2)\| \leq L\|x_1 - x_2\|$ .
- (iii) There exists a function  $\Phi(\cdot)$  such that for all  $R \geq 0$ :

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

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

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

It suffices to recall that the subdifferential of a convex proper lower semi-continuous 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:

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

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  $\lambda$  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:

$$\begin{cases} \dot{x}(t) = f(x(t), t, \lambda(t)), \\ w(t) = h(x(t), \lambda(t)), \\ 0 \leq w(t) \perp \lambda(t) \geq 0. \end{cases} \quad (2.52)$$

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:

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

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:

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

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:

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

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:

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

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:

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

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  $\lambda$ . The relative degree is the number of times one needs to differentiate the “output”  $w$  along the dynamics in order to recover the “input”  $\lambda$ . As an example let us consider the following scalar LCS:

$$\begin{cases} \dot{x}(t) = x(t) + \lambda, \\ 0 \leq \lambda \perp w(t) = x(t) \geq 0. \end{cases} \quad (2.58)$$

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  $r \geq 2$  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.<sup>6</sup>

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:

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

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

---

<sup>6</sup>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:

$$\begin{pmatrix} -A^T P + P A & -P B + C^T \\ -B^T P + C & D + D^T \end{pmatrix} \geq 0 \quad \text{and} \quad P = P^T > 0 \quad (2.60)$$

has a solution  $P$ . The quadratic function  $V(x) = \frac{1}{2}x^T P x$  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:

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

Let us define the set  $K = \{z \in \mathbb{R}^n \mid Cz + Fu(t^+) \in Q_D\}$ , with  $Q_D = \{z \in \mathbb{R}^m \mid z \geq 0, Dz \geq 0, z^T Dz = 0\}$ .  $Q_D^*$  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  $V(x) = \frac{1}{2}x^T P x$ ,  $P = P^T > 0$ . 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  $Fu(t) \in Q_D^* + \text{Im}(C)$ . For any  $x(t^-)$  there is a unique solution to:*

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

that is equivalent to:

$$P(x(t^+) - x(t^-)) \in -N_K(x(t^+)) \quad (2.63)$$

and to

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

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  $Fu(t) \in Q_D^* + \text{Im}(C)$  is a sort of constraint qualification condition, which guarantees that the LCP  $0 \leq \lambda \perp Cx + Fu + D\lambda \geq 0$  has a solution (see Sect. 5.2.2 for a similar condition, stated in a different context). Notice that we have implicitly assumed that  $\lambda$  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  $B^T P B p_t = B^T P(x(t^+) - x(t^-))$ . 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  $p_t = (B^T P B)^{-1} B^T P(x(t^+) - x(t^-))$ .

*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  $\lambda^*$  whatever  $u(t)$  and  $x(t)$ . Moreover this  $\lambda^*$  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.

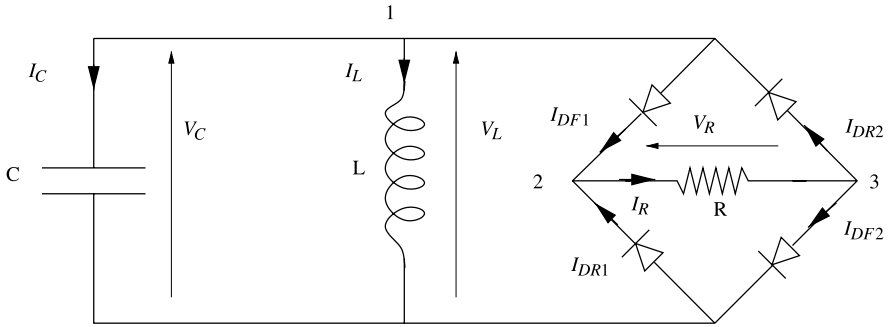
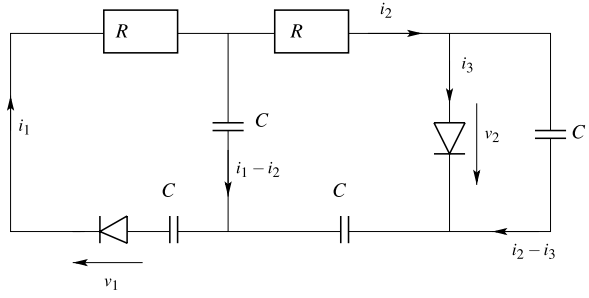
When  $D = 0$ , then one has  $Q_D = \mathbb{R}_+^m$ ,  $Q_D^* = \{0\}$ ,  $K = \{z \in \mathbb{R}^n \mid Cz + Fu(t^+) \geq 0\}$ . Then a state jump may occur depending on the value of  $u(t^+)$  (see Sect. 5.2 for further comments on state jumps).

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  $0 \leq v_1(t) \perp i_1(t) \geq 0$  and  $0 \leq v_2(t) \perp i_3(t) \geq 0$  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:

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

$$0 \leq \lambda(t) \perp w(t) = \begin{pmatrix} 0 & 0 & 1 \\ \frac{-2}{RC} & \frac{0}{RC} & 0 \end{pmatrix} \begin{pmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & \frac{1}{R} \end{pmatrix} \lambda(t) \geq 0.$$

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  $C > 0$ , an inductor  $L > 0$ , a resistor  $R > 0$ . Its dynamics is given by:

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

$$0 \leq w(t) \perp \lambda(t) \geq 0,$$



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

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

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:

$$\begin{aligned} \dot{x}(t) &= -\frac{1}{RC}x(t) + \left(\frac{1}{C} \quad 0 \quad \frac{1}{C} \quad 0\right)\lambda(t), \\ w(t) &= \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} x(t) + \begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 1 & -1 \\ 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \lambda(t), \end{aligned} \quad (2.68)$$

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:

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

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

- (i) the sets  $\chi_i$  are finitely represented as  $\chi_i = \{x \in \mathbb{R}^n \mid C_i x + D_i \geq 0\}$ ,  $C_i \in \mathbb{R}^{m_i \times n}$ ,  $D_i \in \mathbb{R}^{m_i \times 1}$ ,
- (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  $\chi_i$  have an nonempty interior.

Conditions (iii) and (iv) imply that  $\chi_i \cap \text{Int}(\chi_j) = \emptyset$  for all  $i \neq j$ . We denote the set of indices of the partition as  $\mathcal{I}_1$ , i.e. the set of polyhedra is  $\{\chi_i\}_{i \in \mathcal{I}_1}$ . Obviously  $\mathcal{I}_1$  may be finite, or infinite. The properties (ii) and (iii) mean that the polyhedral sets  $\chi_i$  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  $\chi_i$  and  $\chi_j$ : (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:

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

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

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

with  $x(0) \in \mathbb{R}$  and  $|g(t)| \leq \frac{1}{2}$  for all  $t \geq 0$ , 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:

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

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

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

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  $t \geq T$  or  $x(t) = -t - T + \int_T^t g(s)ds$  for  $t \geq T$  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:

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

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

- by construction  $F(x)$  is always non empty, closed and convex for each  $x$ ;<sup>7</sup>
- 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  $\epsilon \rightarrow 0$ . 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 hyper-surface and a polyhedral set. This is particularly true when switching attractive surfaces with co-dimension larger than 2 exist.

---

<sup>7</sup>The boundedness of  $f(x)$  is essential here.

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

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

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

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

and one disregards the possible values on  $\Sigma$  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  $\Omega$  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  $\|F(x)\| \leq c(1 + \|x\|)$  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  $\|F(x)\| \leq c(1 + \|x\|)$  means that for all  $\xi \in F(x)$  one has  $\|\xi\| \leq c(1 + \|x\|)$ : this is a linear growth condition. In view of (2.78) a solution has to satisfy the differential equation

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

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

The proposition says that if the switching surface  $\Sigma$  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  $\Sigma$  is attractive then the solution slides along it (a sliding motion), in the other case it just crosses  $\Sigma$ .

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:

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

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

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

and

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

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  $F : \mathbb{R}^n \rightarrow 2^{\mathbb{R}^n} \setminus \emptyset$  where  $F(t, x)$  is compact for all  $x \in \mathbb{R}^n$  and all  $t \geq 0$ , is called *one-sided Lipschitz continuous* (OSLC) if there is an integrable function  $L : \mathbb{R}^+ \rightarrow \mathbb{R}$  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

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

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  $\varphi : \mathbb{R}^n \rightarrow \mathbb{R}^n$  are multivalued monotone mappings, and  $f(t, x)$  is Lipschitz continuous, are UOSLC. The OSLC constant  $L$  is equal to  $\max(0, \lambda)$ , where  $\lambda$  is the Lipschitz constant of the function  $f(\cdot, \cdot)$ .

*Example 2.73* Consider  $F(x) = \text{sgn}(x)$ , 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  $\langle x_1 - x_2, y_1 - y_2 \rangle \geq 0$ . Therefore the multi-function  $-F(\cdot)$  satisfies  $\langle x_1 - x_2, -y_1 + y_2 \rangle \leq 0$  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  $2(x_1 - x_2) \leq L(x_1 - x_2)^2$  for some  $L$ . A negative  $L$  is impossible, and a nonnegative  $L$  yields  $L \geq \frac{2}{x_1 - x_2}$ . 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  $x_1 : [t_0, +\infty) \rightarrow \mathbb{R}^n$ ,  $x_2 : [t_0, +\infty) \rightarrow \mathbb{R}^n$  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

$$\|x_1(t) - x_2(t)\| \leq \exp(L(t - t_0)) \|x_1(t_0) - x_2(t_0)\| \quad (2.81)$$

for all  $t \geq t_0$ . 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  $\Sigma_{ij}$  between the two cells  $\chi_i$  and  $\chi_j$ , is included into the subspace  $\{x \in \mathbb{R}^n \mid c_{ij}^T x + d_{ij} = 0\}$ . Suppose also that the polyhedron  $\chi_i$  is such that  $c_{ij}^T x + d_{ij} \geq 0$  for all  $x \in \chi_i$  (and consequently  $c_{ij}^T x + d_{ij} \leq 0$  for all  $x \in \chi_j$ ). Then the set:

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

is a first-order (or regular) sliding set for the switching system (2.69) on  $\Sigma_{ij}$ . 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  $\frac{d^2}{dt^2} c_{ij}^T x(t) = c_{ij}^T \ddot{x}(t) < 0$  in  $\chi_i$  and  $\frac{d^2}{dt^2} c_{ij}^T x(t) > 0$  in  $\chi_j$  (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):

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

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:

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

where  $\text{Sgn}(y) = (\text{sgn}(y_1)\text{sgn}(y_2)\cdots\text{sgn}(y_m))^T$ ,  $\text{sgn}(\cdot)$  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:

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

with  $C = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$ ,  $\text{Sgn}(z) = (\text{sgn}(z_1), \dots, \text{sgn}(z_n))^T$  for any vector  $z \in \mathbb{R}^n$ , has a maximal monotone right-hand-side  $x \mapsto C^T \text{Sgn}(Cx)$ . However the set  $C^T \text{Sgn}(Cx)$  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:

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

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  $x \mapsto \text{sgn}(x)$ , where  $\text{sgn}(\cdot)$  is the multivalued sign function. We also have the following for two reals  $y$  and  $z$ :

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

Indeed let  $z > 0$ , then  $\lambda_2 = 0$  and  $\lambda_1 \geq 0$  so that  $\lambda_1 = 2$  and  $y = 1$ . Let  $z < 0$ , then  $\lambda_1 = 0$  and  $\lambda_2 \geq 0$  so that  $\lambda_2 = 2$  and  $y = -1$ . Let  $z = 0$ , then  $\lambda_1 \geq 0$  and  $\lambda_2 \geq 0$ . Since  $\lambda_1 = 2 - \lambda_2$  we get  $y = 1 - \lambda_2$  so  $y \leq 1$ . Similarly  $\lambda_2 = 2 - \lambda_1$  and  $y = \lambda_1 - 1$  so  $y \geq -1$ . 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:

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

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

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

This complementarity system belongs to the class in (2.54) with  $E$  and  $M$  identity matrices of appropriate dimensions,  $F = 1$ ,  $G = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ . 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:

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

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

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

Indeed:

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

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

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

$$PB = C^T. \quad (2.91)$$

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:

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

and

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

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:

$$\text{Rge}(C) - \mathbb{R}_+^m = \mathbb{R}^m, \quad (2.94)$$

where  $\text{Rge}$  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  $y \in \text{Rge}(C)$  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

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

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

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

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 non-smooth evolution problems. The dynamics of electrical circuits with non-smooth 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

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

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

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

### 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:

$$0 \leq -v(t) + R_{\text{off}}i(t) \perp -v(t) \geq 0, \quad (2.99)$$

which we may again rewrite as

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

Using (2.25) we infer that

$$-v(t) = \text{proj}(\mathbb{R}_+; R_{\text{off}}i(t)). \quad (2.101)$$

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

$$\begin{cases} v(t) - R_{\text{off}}i(t) \geq 0, \\ 0 \leq -v(t) \leq +\infty, \\ v(t)(v(t) - R_{\text{off}}i(t)) = 0. \end{cases} \quad (2.102)$$

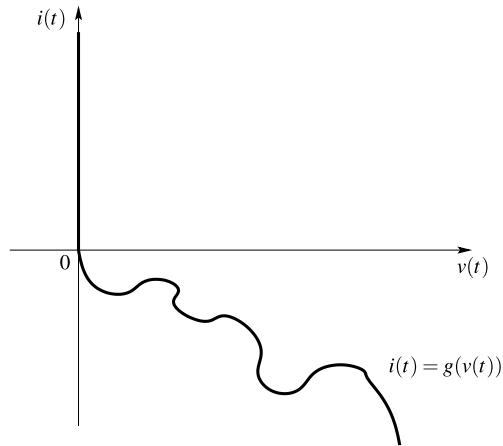
### 2.5.3 A Mixed Nonlinear/Unilateral Diode

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  $v \mapsto g(v)$

---

<sup>8</sup>  $F^+(z) = \max(0, F(z))$ .

**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:

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

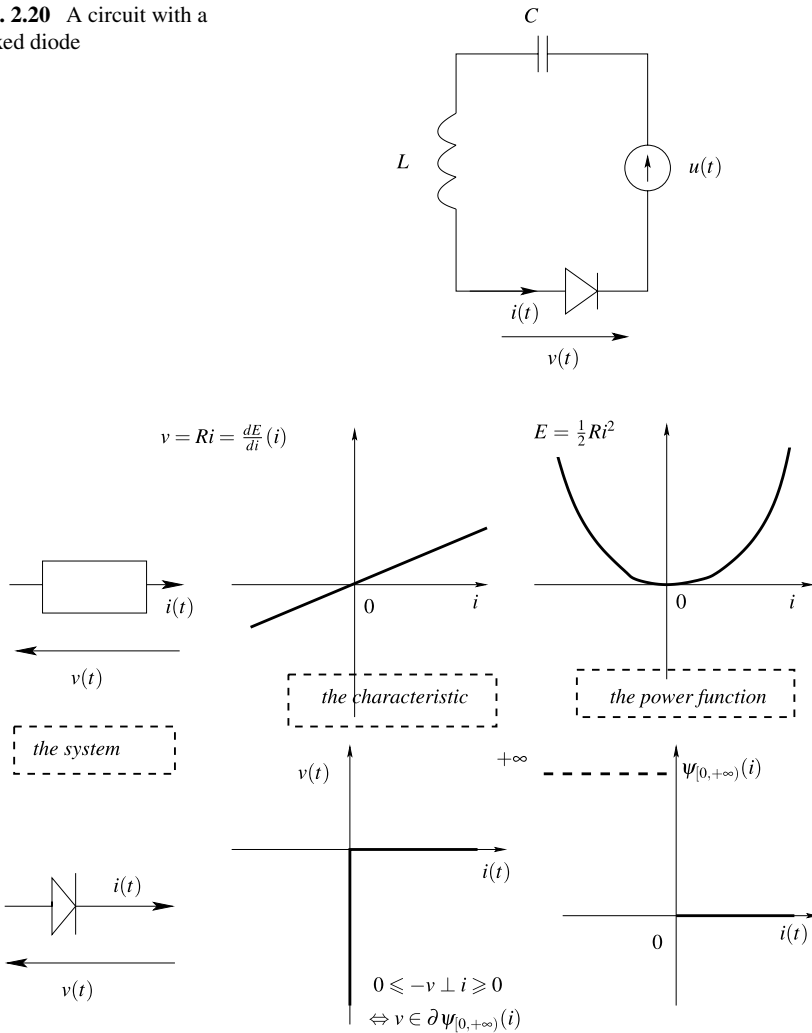
One checks that  $v(t) > 0$  implies that  $i(t) = g(v(t))$ , while  $v(t) = 0$  implies  $i(t) \geq 0$ . Let us now consider the circuit of Fig. 2.20, with a voltage source  $u(t)$ . The state variables are  $x_1(\cdot)$  the capacitor charge, and  $x_2(\cdot)$  the current  $i(t)$  through the circuit.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

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

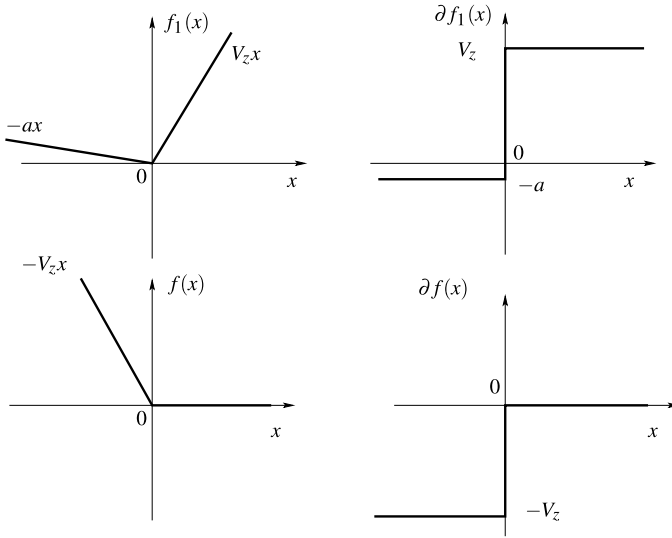
which is an NLCS as in (2.55), letting  $\lambda(t) = v(t)$ . The complementarity conditions in (2.104) define an NLCP (or NCP). If  $x_2(t) < 0$  and  $x_2(t)$  is in the image of  $g(\cdot)$ , then  $-g(v(t)) = -x_2(t) > 0$  for some  $v(t) > 0$ . Uniqueness holds if the function  $g(\cdot)$  is monotone (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  $\varphi : \mathbb{R} \rightarrow \mathbb{R} \cup \{+\infty\}$ . 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  $v \leq 0$  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  $0 \leq w(t) = x(t) - i(t) \perp v(t) \geq 0$  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:

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

When  $i : \mathbb{R} \rightarrow \mathbb{R}$  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.*:

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

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  $\delta_t$ ). At  $t = t_1$  which is an atom of the differential measure  $dx$  one obtains:

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

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

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

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):

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

Using (2.24) this is equivalent to:

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

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

$$0 \leq x_{k+1} - x_k + h \frac{R}{L}x_k \perp x_{k+1} - i_{k+1} \geq 0. \quad (2.111)$$

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:

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

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.

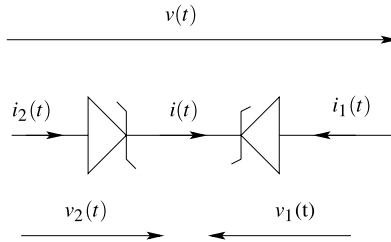
### 2.5.6 The RCD Circuit in (1.3)

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  $v(t) = \text{proj}(\mathbb{R}_+; -u(t) + \frac{1}{C}z(t))$ . Inserting this into the first line of (1.3) one obtains:

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

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  $v(t) \in \mathcal{F}_{z_i}(-i(t))$  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  $f_1(-i) = \begin{cases} ai & \text{if } i \geq 0 \\ -V_z i & \text{if } i \leq 0 \end{cases}$  (see Fig. 2.22). Thus  $\mathcal{F}_{z_i}(-i(t)) = \partial f_1(-i)$ . We infer that the dynamics of this circuit is given by the differential inclusion:

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

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).

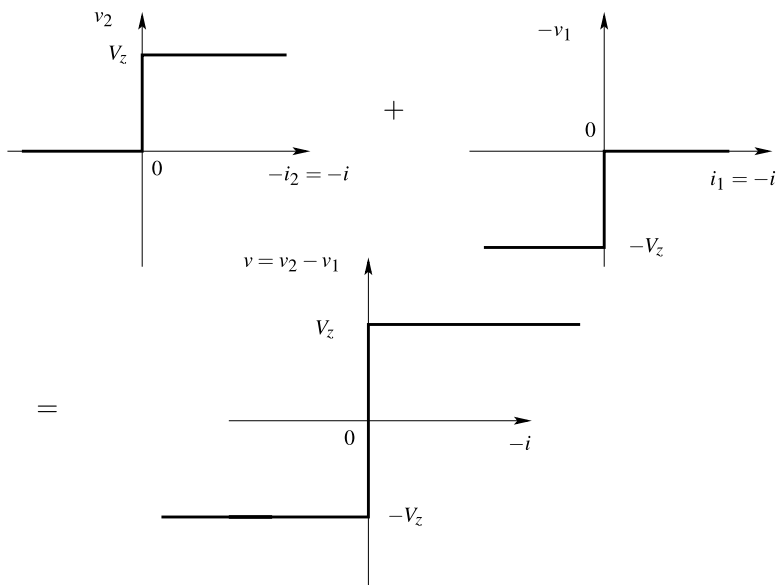
Similar developments hold for the voltage/current law in Fig. 1.6(b). Both characteristics can also be represented in a complementarity formalism.

### 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:

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

where each diode has the voltage/current law  $v_j(t) \in \mathcal{F}_{z_1}(-i_j(t))$  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

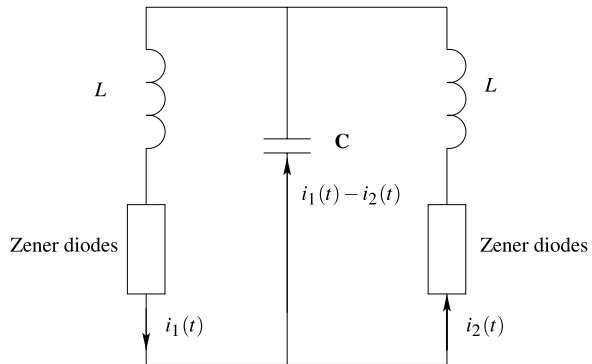
Reversing the sense of the diodes does not change the voltage/current law of the two-diode system, as may be checked. Consider now the circuit of Fig. 2.25, where each Zener box contains two Zener diodes mounted in series as in Fig. 2.23. Its dynamics is given by:

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

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:

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

where we used  $\partial|x| = \operatorname{sgn}(x)$  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,  $\frac{1}{C}$  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:

$$M\ddot{x}(t) + Kx(t) \in -B \operatorname{Sgn}(Cx(t)) \quad (2.118)$$

with  $x^T = (x_1 \ x_2)$ ,  $M = \begin{pmatrix} L & 0 \\ 0 & L \end{pmatrix}$ ,  $K = \begin{pmatrix} \frac{1}{C} & -\frac{1}{C} \\ -\frac{1}{C} & \frac{1}{C} \end{pmatrix}$ ,  $\operatorname{Sgn}(Cx) = (\operatorname{sgn}(x_1) \operatorname{sgn}(x_2))^T$ ,  $B = \begin{pmatrix} V_z & 0 \\ 0 & V_z \end{pmatrix}$ ,  $C = I_2$  the identity matrix. One remarks that the condition (2.91) is trivially satisfied with  $P = B^{-1}$ . The multivalued mapping  $x \mapsto B \operatorname{Sgn}(Cx)$  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  $v(t) \in \mathcal{F}_z(i(t))$  in (1.11) may be rewritten using the subdifferential of the convex lower semi-continuous proper function  $f(i) = \begin{cases} -V_z i & \text{if } i \leq 0 \\ 0 & \text{if } i \geq 0 \end{cases}$  (see Fig. 2.22). We obtain that  $\mathcal{F}_z(i) = \partial f(i)$ . From (1.11) we deduce that

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

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

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

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:

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

Using Proposition 2.37 we deduce that:

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

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:

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

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  $0 \leq v(t) \perp -x_2(t) \geq 0$  into  $-v(t) \in \partial\psi_{\mathbb{R}^+}(-x_2(t))$ . Letting  $f(x_2) \triangleq \psi_{\mathbb{R}^+}(-x_2)$  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  $C$ <sup>9</sup> we may rewrite the system (2.123) as:

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

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

---

<sup>9</sup>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  $K = \{x \in \mathbb{R}^2 \mid Cx \leq 0\}$  is a convex set. Therefore we can rewrite (2.124) as:

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

which fits within (2.49) so that Proposition 2.58 applies. Notice that the condition  $x_0 \in \text{dom}(A)$  of Proposition 2.58 translates into  $x_2(0) \leq 0$  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:

$$dx - Ax(t)dt \in -N_K(x(t)) \quad (2.126)$$

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  $x(t^+) = \text{proj}(K; x(t^-))$ . 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  $0 \leq v(t) \perp -x_2(t) \geq 0$  implies:

$$0 \leq v(t) \perp -\dot{x}_2(t) \geq 0. \quad (2.127)$$

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:

$$0 \leq -L\dot{x}_2(t) - \frac{1}{C}x_1(t) \perp -\dot{x}_2(t) \geq 0, \quad (2.128)$$

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:

$$0 \leq \frac{1}{C}x_1(t) + \frac{1}{L}v(t) \perp v(t) \geq 0. \quad (2.129)$$

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  $\text{bd}\chi$  that separates  $\chi_1$  and  $\chi_2$ . 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  $A_1x(t)$  and  $A_2x(t)$  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

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

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

$$PB_1 = C_1^T, \quad PB_2 = C_2^T, \quad (2.131)$$

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:

$$\begin{cases} \dot{z}(t) = RAR^{-1}z(t) - RBv(t) + REu(t), \\ v_1(t) \in \mathcal{F}_1(C_1x(t)), \quad v_2(t) \in \mathcal{F}_2(C_2x(t)), \end{cases} \quad (2.132)$$

with obvious definitions of  $\mathcal{F}_1(\cdot)$  and  $\mathcal{F}_2(\cdot)$  from (1.42). A key property of the multivalued functions  $\mathcal{F}_i(\cdot)$  is that there exist proper convex lower semi-continuous functions  $\varphi_i(\cdot)$  such that  $\mathcal{F}_i(\cdot) = \partial\varphi_i(\cdot)$ . These functions are given by  $\varphi_1(x) = \begin{cases} -V_2x & \text{if } x < 0 \\ 0 & \text{if } x > 0 \end{cases}$ , and  $\varphi_2(x) = \psi_K(x)$  with  $K = \mathbb{R}^+$ . We may rewrite (2.132) as:

$$\begin{cases} \dot{z}(t) = RAR^{-1}z(t) - RB_1v_1(t) - RB_2v_2(t) + REu(t), \\ v_1(t) \in \partial\varphi_1(C_1R^{-1}z(t)), \quad v_2(t) \in \partial\varphi_2(C_2R^{-1}z(t)). \end{cases} \quad (2.133)$$

Using (2.131) the terms  $RB_1v_1$  and  $RB_2v_2$  may be rewritten as  $R^{-1}C_1^Tv_1$  and  $R^{-1}C_2^Tv_2$ , respectively. Using the inclusions in (2.133) one obtains the two terms  $R^{-1}C_1^T\partial\varphi_1(C_1R^{-1}z)$  and  $R^{-1}C_2^T\partial\varphi_2(C_2R^{-1}z)$ . Now we may use Proposition 2.29 to deduce that  $R^{-1}C_1^T\partial\varphi_1(C_1R^{-1}z) = \partial(\varphi_1 \circ C_1R^{-1})(z)$  and  $R^{-1}C_2^T\partial\varphi_2(C_2R^{-1}z) = \partial(\varphi_2 \circ C_2R^{-1})(z)$ . Let us denote  $\varphi_1 \circ C_1R^{-1}(\cdot) = \phi_1(\cdot)$  and  $\varphi_2 \circ C_2R^{-1}(\cdot) = \phi_2(\cdot)$ , 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:

$$-\dot{z}(t) + RAR^{-1}z(t) + REu(t) \in \partial\Phi(z(t)), \quad (2.134)$$

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:

$$-z_{k+1} + z_k + hRAR^{-1}z_{k+1} + hREu_{k+1} \in h\partial\Phi(z_{k+1}), \quad (2.135)$$

which we rewrite as

$$0 \in (I_3 - hRAR^{-1})z_{k+1} + z_k + hREu_{k+1} + h\partial\Phi(z_{k+1}). \quad (2.136)$$

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

$$0 \in hv_{k+1} + q_k + N_K(v_{k+1}), \quad (2.137)$$

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.

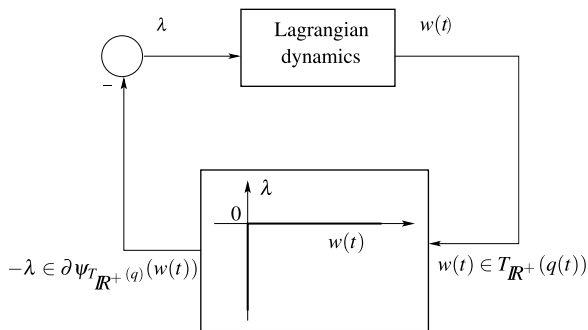
### 2.5.13 The Bouncing Ball

Let us come back on the dynamics in (1.58). This may be recast into the Lagrangian sweeping process (2.38):

$$\begin{cases} mdv + mgdt + u(t)dt = \lambda, \\ -\lambda \in N_{T_{\mathbb{R}^+}(q(t))}(w(t)). \end{cases} \quad (2.138)$$

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  $\lambda$  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.

**Fig. 2.26** Bouncing-ball feedback interconnection with the corner law



Notice that (2.138) is a measure differential inclusion, so that  $\lambda$  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  $\lambda$  thus has a density  $p$  with respect to the Dirac measure  $\delta_t$  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:

$$\begin{cases} \frac{x_{k+1} - x_k}{h} + A(x_{k+1}) \ni f(t_k, x_k), & \text{for all } k \in \{0, \dots, N-1\}, \\ x_0 = x(0), \end{cases} \quad (2.139)$$

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  $\theta$ -method which is used in practical implementations. The next result is proved in Bastien and Schatzman (2002).

**Proposition 2.83** *Under Assumption 2.57,<sup>10</sup> there exists  $\eta$  such that for all  $h > 0$  one has*

$$\text{For all } t \in [0, T], \quad \|x(t) - x^N(t)\| \leq \eta \sqrt{h}. \quad (2.140)$$

Moreover  $\lim_{h \rightarrow 0^+} \max_{t \in [0, T]} \|x(t) - x^N(t)\|^2 + \int_0^t \|x(s) - x^N(s)\|^2 ds = 0$ .

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:

$$\begin{cases} x_{k+1} = x_k + hAx_{k+1} + hB\lambda_{k+1}, \\ w_{k+1} = Cx_{k+1} + D\lambda_{k+1}, \\ 0 \leq \lambda_{k+1} \perp w_{k+1} \geq 0. \end{cases} \quad (2.141)$$

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:

$$0 \leq \lambda_{k+1} \perp C(I_n - hA)^{-1}(x_k + hB\lambda_{k+1}) + D\lambda_{k+1} \geq 0, \quad (2.142)$$

with unknown  $\lambda_{k+1}$  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.

---

<sup>10</sup>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  $\lim_{h \rightarrow 0} \langle x^N(t) - x(t), \varphi(t) \rangle = 0$  for all  $\varphi \in \mathcal{L}^2([0, T]; \mathbb{R}^n)$  and all  $t \in [0, T]$ , which is the weak convergence in  $\mathcal{L}^2([0, T]; \mathbb{R}^n)$ .

**Theorem 2.86** *Consider the LCS in (2.53) with  $D \geq 0$  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  $\|h\lambda_0\| \leq \alpha$  and  $\|\lambda_k^N\| \leq \alpha$  for all  $k \geq 0$ . Then for any sequence  $\{h_k\}_{k \geq 0}$  that converges to zero, one has:*

- (i) *There exists a subsequence  $\{h_{k_l}\} \subseteq \{h_k\}_{k \geq 0}$  such that  $(\{\lambda^N\}_{k_l}, \{w^N\}_{k_l})$  converges weakly to some  $(\lambda, w)$  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  $(\lambda, w)$  and the whole sequence  $\{x^N\}_k$  converges to  $x(\cdot)$ .*

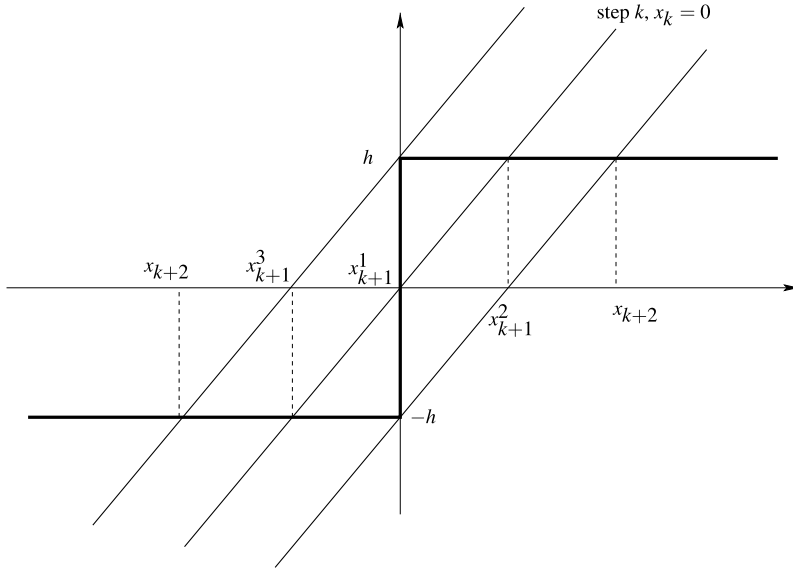
*If the quadruple  $(A, B, C, D)$  is such that Assumption 2.85 holds and is passive, then (iii) holds.*

We emphasize the notation  $x(\cdot)$  since the solutions are functions of time, whereas the notation  $\lambda$  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  $CB = B^T P B \geq 0$ ).

**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:

$$x_{k+1} - x_k \in h \operatorname{sgn}(x_{k+1}). \quad (2.143)$$

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)

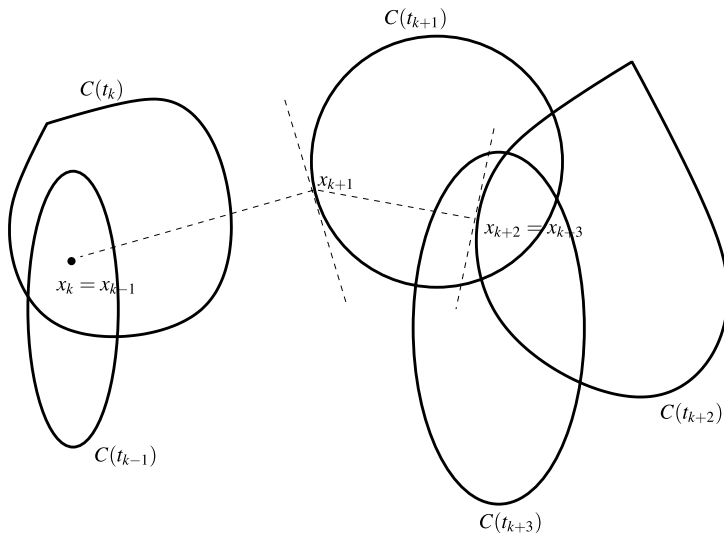
At step  $k$  there are three intersections (solutions of the generalized equation):  $x_{k+1}^1 = 0$ ,  $x_{k+1}^2 = h$  and  $x_{k+1}^3 = -h$ . At step  $k + 1$  starting from  $x_{k+1}^2$  or  $x_{k+1}^3$  there are two solutions:  $x_{k+2}^{2,1} = 0$  or  $x_{k+2}^{2,2} = 2h$ , and  $x_{k+2}^{3,1} = 0$  or  $x_{k+2}^{3,2} = -2h$ . 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:

$$-dx \in N_{C(t)}(x(t)), \quad x(0) = x_0, \quad (2.144)$$

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):

$$-(x_{k+1} - x_k) \in \partial\psi_{C(t_{k+1})}(x_{k+1}), \quad (2.145)$$

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:

$$x_{k+1} = \text{prox}[C(t_{k+1}); x_k]. \quad (2.146)$$

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)  $\text{var}_{[0, T]}(x^N) \leq \|x^N(0)\| + lT$ , for all  $t \in [t_k, t_{k+1}]$  and all  $N \in \mathbb{N}$ ,
- (b)  $\|x^N(t) - x^N(s)\| \leq l(|t - s| + \frac{2}{m})$ , for all  $t, s \in [t_k, t_{k+1}]$ ,
- (c) from which it follows that  $\|x(t) - x(s)\| \leq l|t - s|$  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)  $\|\dot{x}^N(t)\| \leq l$  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  $\varphi(\cdot) \in \mathcal{L}^1([0, T]; \mathbb{R}^n)$  one has

$$\int_0^T \langle \dot{x}^N(t), \varphi(t) \rangle dt \rightarrow \int_0^T \langle \dot{x}^*(t), \varphi(t) \rangle dt,$$

- (f)  $x^N(\cdot) \rightarrow x(\cdot)$  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 usually 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  $Cx(0) + D\lambda(0) \geq 0$ .

- 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  $r \geq 3$  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; Bielek 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.

Nonsmooth Modeling and Simulation for Switched  
Circuits

Acary, V.; Bonnefon, O.; Brogliato, B.

2011, XXIII, 284 p., Hardcover

ISBN: 978-90-481-9680-7