Chapter 3
Discontinuities to Model Missing Knowledge

If the theory of differentiable dynamical systems entered its nonlinear age with the discovery of bifurcations and chaos, then nonsmooth dynamics' nonlinear age will be characterized by tackling issues of determinacy. Loss of determinacy is an inescapable feature of nonsmooth systems, but it requires more explicit expression if it is to be turned to more useful modelling.

When integrating a function that involves something like a step function, the value step(0) at the discontinuity does not affect the value of the integral, provided the argument of the step function is monotonic over the integral path (see, e.g., chapter 1 of [51]). In a dynamic problem of the form $\frac{d}{dt}\mathbf{x} = \mathbf{F}(\mathbf{x}; \text{step}(t))$, for instance, solutions may be written as

$$\begin{aligned}\mathbf{x}(t) &= \mathbf{x}(\tau) + \int_\tau^t ds\ \mathbf{F}(\mathbf{x}(s); \text{step}(s)) \\ &= \mathbf{x}(\tau) + \int_\tau^0 ds\ \mathbf{F}(\mathbf{x}(s); 0) + \int_0^t ds\ \mathbf{F}(\mathbf{x}(s); 1)\ ,\end{aligned}$$

assuming $\tau < 0 < t$, whose existence is provided by Carathéodory's theorem (extending Peano's existence theorem to non-differentiable systems of ordinary differential equations, see, e.g., [28, 51, 67]). The step function simply divides the integral in two, and the value of step(0) does not affect the right-hand side. In a dynamic problem of the form $\dot{\mathbf{x}} = \mathbf{F}(\mathbf{x}; \text{step}(\sigma(\mathbf{x})))$ for some function σ, however, where the discontinuity threshold is the set $\mathcal{D} = \{\mathbf{x} : \sigma(\mathbf{x}) = 0\}$, when seeking solutions of the Carathéodory form,

$$\mathbf{x}(t) = \mathbf{x}(a) + \int_a^t ds\ \mathbf{F}\Big(\mathbf{x}(s); \text{step}\left(\sigma(\mathbf{x}(s))\right)\Big)\ ,$$

the value of the step function becomes essential, because the argument $\sigma(\mathbf{x}(s))$ could remain at zero for a non-vanishing interval of s values. The *existence* of sets of Carathéodory type solutions to problems of the form (2.2) was covered rigorously

M. R. Jeffrey, *Modeling with Nonsmooth Dynamics*,
Frontiers in Applied Dynamical Systems: Reviews and Tutorials 7,
https://doi.org/10.1007/978-3-030-35987-4_3

in [51], but in order to use those solutions for the purposes of modelling, we require a more explicit way to characterize them, to distinguish the different dynamics they make possible.

Mixed up in this problem is the understanding of what 'nonsmoothness' represents as an approximation. Nonsmooth systems are idealized in the sense that they take abrupt transitions, and represent them as discontinuities occurring at definite hypersurfaces in space. In what applied contexts is this a suitable model of abrupt change, and what physical or living processes can it faithfully represent? What are the implications of regularizing a discontinuity to obtain a well-defined system, and does it matter *how* we regularize? Should we obtain similar behaviour whether we smooth the discontinuity out, or numerically calculate the state $x(t)$ by interpolating between time intervals $t, t + \delta_1, t + \delta_2, \ldots$? What happens if the true system actually lies a short distance δx away, or suffers a time lag δt, shifting $x(t)$ to $x(t - \delta t) + \delta x(t - \delta t)$, and does it matter whether these perturbations are simple functions or stochastic processes? Should all such implementations of switching give similar *perturbations* of the nonsmooth model, and if not, how do we make sense of them as approximations of some underlying physics? Various such problems have been discussed in the literature in a somewhat fragmented manner. Filippov's formulation in [51] used differential inclusions, while a recent trend to smooth the discontinuity has followed the approach introduced in [136] (and extended to include hidden dynamics in [31, 111]). The discontinuity's implementation is considered with spatio-temporal delays or 'chatter' in [2, 6, 129, 148], or with stochasticity in [132].

Part of the difficulty of developing a theory of how discontinuities affect dynamics may be in our lack of understanding of what causes them. They tend to be used when we have incomplete knowledge of the underlying processes when some abrupt transition occurs, so we apply different set of equations under different conditions, in different regions of state or parameter space. A way to clarify what a discontinuity represents in a model might be to separate out those that are *passive*, *active*, or *dynamic* in origin.

A *passive* discontinuity is simply a change in the physical parameters defining a system, such as density, conductivity, or reflectivity, which jump across the interface between different materials. A discontinuity in refractive index, for example, is responsible for light bending as it passes between air and water. The jump in density between air and metal allows a hammer to swing freely through the air, and yet impart a force when it contacts with a nail. To precisely understand the frictional and impact contact forces between such media requires a microscale understanding of their interfaces, but such a level of detail can hardly be useful in a large scale model of the dynamics of the bodies themselves, so we approximate using, among other simplifications, discontinuities.

An *active* discontinuity is imposed upon a system as a means of control. Examples might be switches or valves made of mechanical, electronic, biological, or chemical parts, opening and closing different channels that drive or starve different parts of a system. Or they may be decisions made by individuals or groups about how to govern institutions, how to invest, or what causes to support. How we model such discontinuities depends on whether they are imposed instantaneously, grad-

ually, or via a series of sub-processes. With potentially so many complex factors involved, the more detailed the model the less general would be any results obtained from it.

A *dynamic* discontinuity is induced by some small and/or fast scale change in stability, a jump from one stable attractor or pattern to another. The transition may be too abrupt to be easily observed itself, but manifests as a large scale change in behaviour, such as a financial crash, a change in heart rhythm, the collapse of a structure, or onset of turbulence. Of our three categories, these have the most comprehensible origins in the form of bifurcations or phase transitions, for which we have well developed mathematical theory, such as the asymptotic theory discussed in Sect. 1.1, particularly as concerns Stokes discontinuities and shocks [14, 16, 39, 68, 71].

These distinctions are not definitive, nor are they unique. For example, if a person makes a decision *actively*, then this may actually be the result of some *dynamic* phase transition across networks of neurons in the brain. But they illustrate the different features that discontinuities are called upon to approximate, the huge complexity that is disguised by writing $|x|$, $\text{step}(x)$, or other conditional statements, in a systems of differential equations. We must formulate nonsmooth dynamics in a way that is explicit enough to explore the assumptions behind using such terms, and robust enough to start relaxing those assumptions.

Such issues were already in the minds of the pioneers of nonsmooth theory. It is worth recounting Filippov's own thoughts from section 8 of [51] on what nonsmooth models are and what they aspire to:

1. Differential equations with discontinuous right-hand sides are often used as a simplified mathematical description of some physical systems. The choice of one or another way of definition of the right-hand side of the equation on a surface of discontinuity . . . depends on the character of the motion of the physical system near this surface.
2. Suppose that outside a certain neighbourhood of a surface of discontinuity of the function $F(t, x)$ the motion obeys the equation $\dot{x} = F(t, x)$. In this neighbourhood the law of motion may not be completely known. Suppose the motion in this neighbourhood may proceed only in two regimes, and switching over from one regime to the other has a retardation, the value of which is known only to be small.
3. Using these incomplete data, we should choose the way of defining the right-hand side of the equation of the surface of discontinuity, so that a sufficiently small width of the neighbourhood the motions of the physical system differ arbitrarily little from the solutions of the equation $\dot{x} = F(t, x)$ defined in the way we have chosen[1].

The subtle problem Filippov himself poses here has been largely overlooked in the advancing theory of nonsmooth dynamics. Here we will loosen the main concepts of nonsmooth dynamics in a way that allows us to probe these issues more deeply.

[1]In the original text this appears as one continuous passage, but for emphasis we have broken it into three items.

If discontinuities were all known to be of the dynamic kind, then nonsmooth dynamics would be just a direct extension of nonlinear dynamics and singular perturbations. Instead, the dominant influence in the development of nonsmooth dynamics has been of the *active* kind, from electronic and mechanical control, through to biological regulation, where we often have incomplete knowledge of the processes behind the discontinuity. In the next two sections we use these applications to introduce some basic concepts.

A lack of a way to quantify and ultimately reconcile such different approaches to handling discontinuity remains the 'elephant in the room' for nonsmooth dynamicists. Nevertheless, given the successes of these different lines of study we should now be able to form a more general picture. To do so we shall have to slightly loosen some of the standard concepts of nonsmooth dynamics. What we will show here is that the different behaviours that are possible at a discontinuity can be distinguished explicitly. We will also show that in regularizing the discontinuity one may unwittingly single out only one of the many possible behaviours, such that different treatments of the discontinuity giving contradictory dynamics.

The path to a general theory is laid by forming a common framework to describe how solutions 'handle' or *implement* a discontinuity (less restrictive than attempting to 'regularize' the discontinuity), along with recognizing a common set of behaviours that such implementations give rise to.

Our first steps towards formalizing such a framework here culminate in a conjecture that can be paraphrased as:

> solutions of a system of ordinary differential equations with a discontinuity along a threshold $\mathcal{D}$, lie ε-close to solutions of a similar system in which discontinuity occurs in some region $\mathcal{D}^\varepsilon$, such that $\mathcal{D}^\varepsilon \to \mathcal{D}$ as $\varepsilon \to 0$.

Such a conjecture cannot be expected to hold without adequately specifying how switching takes place across $\mathcal{D}^\varepsilon$ between differentiable regimes outside $\mathcal{D}^\varepsilon$. Filippov's own work (see, e.g., sections 8–12 of [51]) considered ε-perturbations away from the ideal discontinuous equations to prove the existence of solutions, and that they travel along the convex hull of the nearby vector fields (we will look at this hull in Sect. 5.1), leading to sets or 'funnels' of possible solutions. Here we will look at how to model specific vector fields and solutions among those possibilities, to find to what extent we can obtain a well-defined and robust model. We shall show that non-idealities, no matter how small, play a crucial and fascinating role.