OPTIMAL CONTROL OF SYSTEMS WITH DISCONTINUOUS DIFFERENTIAL EQUATIONS

DAVID E. STEWART* AND MIHAI ANITESCU†

Abstract. In this paper we discuss the problem of verifying and computing optimal controls of systems whose dynamics is governed by differential systems with discontinuous right hand side. In our work, we are motivated by optimal control of mechanical systems with Coulomb friction, which exhibit such right hand side. Notwithstanding the impressive development of non-smooth and set-valued analysis, these systems have not been closely studied either computationally or analytically. First, we show that even when the solution crosses and does not stay on the discontinuity, differentiating the results of a simulation gives gradients that have errors of a size independent of the step-size. This means that the strategy of "optimize the discretization" will usually fail for problems of this kind.

We approximate the discontinuous right-hand side for the differential equations or inclusions by a smooth right-hand side. For these smoothed approximations, we show that the resulting gradients approach the true gradients provided the start and end points of the trajectory do not lie on the discontinuity, and that using Euler's method where the step size is "sufficiently small" in comparison with the smoothing parameter. Numerical results are presented for a crude model of car racing which involves Coulomb friction and slip showing that this approach is practical and can handle problems of considerable complexity.

Manuscript date: *May 24, 2005*

1. Introduction. Consider, for example, a block on a table subject to Coulomb friction on the contacting surface, pulled by a force $g(t)$ [13, 31, 28]: The differential equation for this system is

$$m\frac{dv}{dt} \in -\mu N \operatorname{Sgn}(v) + g(t), \tag{1.1}$$

where Sgn is a *set-valued* function given by

$$\operatorname{Sgn}(z) = \begin{cases} \{+1\}, & z > 0, \\ [-1,+1], & z = 0 \\ \{-1\}, & z < 0 \end{cases} \tag{1.2}$$

The quantity N is the normal contact force ($= mg$ for a block of mass m) and μ the coefficient of Coulomb friction.

It should be noted that Sgn is a maximal monotone set-valued map [2, 4]: a set-valued map $F\colon \mathbb{R}^n \to \mathcal{P}(\mathbb{R}^n)$ is maximal monotone if F is monotone ($y_i \in F(x_i)$ for $i = 1, 2$ implies that $(y_2 - y_1)^T (x_2 - x_1) \geq 0$), and for any set-valued map G where $\operatorname{graph} F \subseteq \operatorname{graph} G$ and G is monotone, $F = G$.

A differential inclusion

$$\frac{dx}{dt} \in F(x), \qquad x(0) = x_0$$

with F and x_0 given has unique solutions if F satisfies a one-sided Lipschitz condition: there is a constant $L \geq 0$ where

$$y_i \in F(x_i) \text{ for } i = 1, 2 \quad \text{implies} \quad (y_2 - y_1)^T (x_2 - x_1) \leq L\, \|x_2 - x_1\|^2. \tag{1.3}$$

---

*Corresponding author, Department of Mathematics, University of Iowa, Iowa City, IA 52242, USA. Part of this work was carried out while visiting Argonne National Laboratory.

†Mathematics and Computer Science Division, Argonne National Laboratories, Argonne, IL , USA.

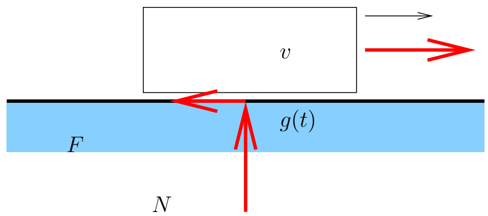

Fig. 1.1. *Block sliding on a table*

If F is upper semi-continuous [3] with closed, convex values and satisfies this one-sided Lipschitz condition, then $x \mapsto F(x) + L\,x$ is a maximal monotone map. In fact the solution operator $S_t \colon \mathbb{R}^n \to \mathbb{R}^n$ where $S_t(x_0) = x(t)$ for given x_0 is Lipschitz with Lipschitz constant $e^{L\,t}$ [4].

Prior work has been done on theoretical aspects non-smooth optimal control problems, including [8, 9, 18, 17, 15]. However, none of this work deals with discontinuous dynamics. The work of Clarke [8, 9] deals with non-smooth but Lipschitz dynamics and objective functions, while Frankowska [18, 17, 15] deals with set-valued but Lipschitz dynamics. Both approaches develop a *maximum principle* generalizing the well-known Pontryagin maximum principle [26] for optimal control.

Thus, this previous work of Clarke, Frankowska *et al.*, cannot be directly applied to systems which have Coulomb friction. Indeed, there might not be a natural maximum principle which can be applied to Coulomb friction problems. Furthermore, we give examples to show that the strategy of "optimizing the discretization" is unlikely to work for such problems.

Numerical work on optimizing systems with dynamics like (1.1) includes [22, 11, 36, 34]. Of these, Glowinski and Kearsley [22] used pattern search to carry out the optimization, with the simulation carried out using a regularized version of a discrete variational inequality obtained via a time-discretization of their multi-dimensional version of (1.1). Driessen and Sadegh [11] set the entire dynamics up as a mixed integer–linear program after using a standard time-discretization. The integer variables were used to represent the values of the "Sgn" function at each time-step. Ventura and Martinez [36] used a hybrid neural network/evolutionary computation approach to computing optimal controls. It should perhaps be noted that the examples considered by Ventura and Martinez had very small friction forces (e.g., $|F| \leq \mu N = 5 \times 10^{-4}$). Van Willigenburg and Loop [34] used adjoint equations to compute gradients so that a conventional constrained optimization routine could be applied (BCPOL from IMSL in this paper). However, there is reason to believe that the adjoint functions computed by Van Willigenburg and Loop are not, in fact, correct, as they did not take into account that the discontinuity in the right-hand side (1.1) causes a discontinuity in the adjoint functions. This phenomenon of discontinuous adjoint functions has been noticed by Driessen and Sadegh [11], and is discussed in depth below.

We mention two examples of analytical investigation of optimal control problems with Coulomb friction. The first is the work of Lipp on the brachistochrone problem with Coulomb friction [24], although the slip is assumed to always be in a fixed direction so that the dynamics is continuous, although not smooth. The second is the work of Kim and Ha [23] who investigate a specific two-dimensional problem with Coulomb friction and find, for their simple system, that the adjoint variables have a jump, and they compute the size of that jump.

We mention that there has been some success with optimizing *static* systems with Coulomb friction. In particular, [25, Ch. 11] discuss using a bundle method of Lemarechal to optimize the friction coefficients for a contact problem.

As can be noticed in all of the above examples for optimal control of (1.1), gradient information is either not used or probably incorrect.

Some authors have considered the problem of computing correct parametric sensitivities. These include the work of Barton *et al.* [20, 33], which develops a "jump formula" for the sensitivities as the trajectory crosses a discontinuity. However, our work is different in the three important ways from that work.

1. The models considered by Barton *et al.* implicitly assume that the trajectory does not stay on the discontinuity for any length of time. This is commonly not true for discontinuous systems such as arise with Coulomb friction, and we analyze such systems in depth in Section 6.
2. The same references contain the observation, that we also emphasize here, that in the case of numerical simulation, the derivatives are not computed correctly if the switching time is not accurately identified. However, we take this observation further in the context of optimal control, by showing that systems of the type described here whose derivative is computed by a fixed step time stepping procedure, may exhibit local minima that accumulate to arbitrary points in the neighborhood of the actual minimum.
3. The models considered by Barton also refer to differential algebraic equation that are index one on the smooth portions, whereas the differential algebraic equations that are equivalent to our model are index two.

Furthermore, in this paper we show that adjoints computed by smoothing the right-hand side of the differential equation will converge to the true adjoints, satisfying the relevant "jump conditions".

1.1. Organization of the paper. In Section 2 we look at the "optimize the discretization" strategy and show that it fails for problems of the same kind as (1.1) whether explicit, implicit, or partly implicit time-discretizations are used. In Section 4 a smoothing approach is introduced, and some general properties of this approaches is developed. This class of systems contains systems of type (1.1). As a result we develop a rule for computing the jumps in the adjoint functions for systems of this type. In Section 6 we show that provided the step-size goes to zero faster than the smoothing parameter, then the gradients and adjoints computed for Euler's method converges to the exact gradients for the discontinuous system. In Section 7 a crude model of a racing car is developed involving Coulomb friction, which is used as a test model. Numerical results are obtained via a smoothing approach which shows the practicality of the approach for a problem of moderate complexity.

Regarding the notation for gradients and Jacobians: Most vectors are considered to be column vectors unless otherwise specified. For a function $f\colon \mathbb{R}^m \to \mathbb{R}^n$, $\nabla f(x)$ is an $n \times m$ matrix so that $(\nabla f(x))_{ij} = \partial f_i / \partial x_j(x)$. This means that for scalar functions ($n = 1$), $\nabla f(x)$ is a *row* vector. However, if f is a function of one variable ($m = 1$), $\nabla f(x) = f'(x)$ is a column vector. This means that $f(x + \Delta x) = f(x) + \nabla f(x)\Delta x + o(\|\Delta x\|)$ for any differentiable f, regardless of whether f is scalar- or vector-valued.

Note that we use C to denote a quantity that depends only on the data of the problem (that is, it does **not** depend on the other parameters introduced, such as the smoothing parameter σ, the step-size h, the time t or step number k). These quantities can differ on each appearance. Since we use asymptotic notation, we remind the reader that $f(s) = O(g(s))$ (as $s \downarrow 0$) means that there are constants $C > 0$ and $s_0 > 0$ where for $0 < s < s_0$, $|f(s)| \leq C\, g(s)$. Also $f(s) = o(g(s))$ means that $\lim_{s\downarrow 0} f(s)/g(s) = 0$. Furthermore, $f(s) = \Omega(g(s))$ means that there are $C > 0$ and $s_0 > 0$ where for $0 < s < s_0$, $f(s) \geq C\,|g(s)|$. Finally, $f(s) = \omega(g(s))$ means that $\lim_{s\downarrow 0} g(s)/f(s) = 0$.

2. "Optimize the discretization" strategy. Consider first the simple differential inclusion

$$\frac{dx}{dt} \in -\mathrm{Sgn}(x), \qquad x(0) = 1. \tag{2.1}$$

The exact solution is unique and is easily checked to be $x(t) = (1-t)_+$ where $z_+ = \max(z,0)$ is the positive part of z. We can discretize this equation using the explicit Euler method, or a partially explicit Euler method. If we set $t_k = t_0 + k\,h$ where $h > 0$ is the time-step, and x^k is our approximation to $x(t_k)$, the discrete-time trajectories will satisfy

$$x^{k+1} \in x^k + h\,F(x^k + \theta(x^{k+1} - x^k)). \tag{2.2}$$

The parameter $\theta \in [0,1]$ indicates how implicit the method is: $\theta = 0$ corresponds to the explicit Euler method; $\theta = \frac{1}{2}$ corresponds to the mid-point rule; $\theta = 1$ corresponds to the fully implicit Euler method [1]. Solutions of the discretized problem are known to converge to solutions of the continuous time differential inclusion (2.1) (see [32, 31, 30]). However, we will shortly see that even though the numerical trajectories converge ($S_{t,h}(x_0) \to S_t(x_0)$ as $h \downarrow 0$), the gradients do not ($\nabla S_{t,h}(x_0) \not\to \nabla S_t(x_0)$) even where S_t is smooth.

Note that if $L\,h < 1$, then there is only one solution to (2.2).

For $\theta < 1$, the main problem is one of "chattering": the numerical solutions will jump from one side of the discontinuity in $dx/dt \in -\mathrm{Sgn}(x)$. For $x^k > h$, $x^{k+1} = x^k - h$, and for $x^k < -h$, $x^{k+1} = x^k - h$. But if $|x^k| \leq h$, then $x^k + \theta(x^{k+1} - x^k) = 0$. That is,

$$x^{k+1} = \frac{1-\theta}{\theta}\,x^k.$$

For $0 < \theta \leq \frac{1}{2}$, this results in oscillation around $x = 0$ that does not go to zero as $k \to \infty$. This is chattering. A similar process occurs at $\theta = 0$, but then the choice of x^{k+1} is not unique.

For $\frac{1}{2} < \theta < 1$ there is still oscillation, but it decays exponentially in k. Thus $\nabla S_{t,h}(1) \to \nabla S_t(1)$ as $h \downarrow 0$ for t *strictly greater or strictly less than one*. This is arguably acceptable, since $S_t(1)$ is not differentiable at $t = 1$.

For $\theta = 1$ there is no oscillation, and the computed gradient of zero would be correct.

We will now look at another example, where even choosing $\theta = 1$ will result in large errors in the gradient, even far from the time when the discontinuity is reached.

Consider the differential inclusion

$$\frac{dx}{dt} \in (1+\alpha) - \mathrm{Sgn}(x), \qquad x(0) = -1, \tag{2.3}$$

with $\alpha > 0$. The exact solution has $x(t) = -1 + (2+\alpha)t$ for $0 \leq t \leq 1/(2+\alpha)$, and $x(t) = \alpha(t - 1/(2+\alpha))$ for $t \geq 1(2+/\alpha)$. For $x(0) = x_0$ with $x_0 \approx -1$, the solution is nearly as simple: $x(t) = x_0 + (2+\alpha)t$ for $0 \leq -x_0/(2+\alpha)$, and $x(t) = \alpha(t + x_0/(2+\alpha))$ for $t \geq -x_0/(2+\alpha)$. This means that $\partial x(2)/\partial x_0 = \alpha/(2+\alpha)$ at $x_0 = -1$.

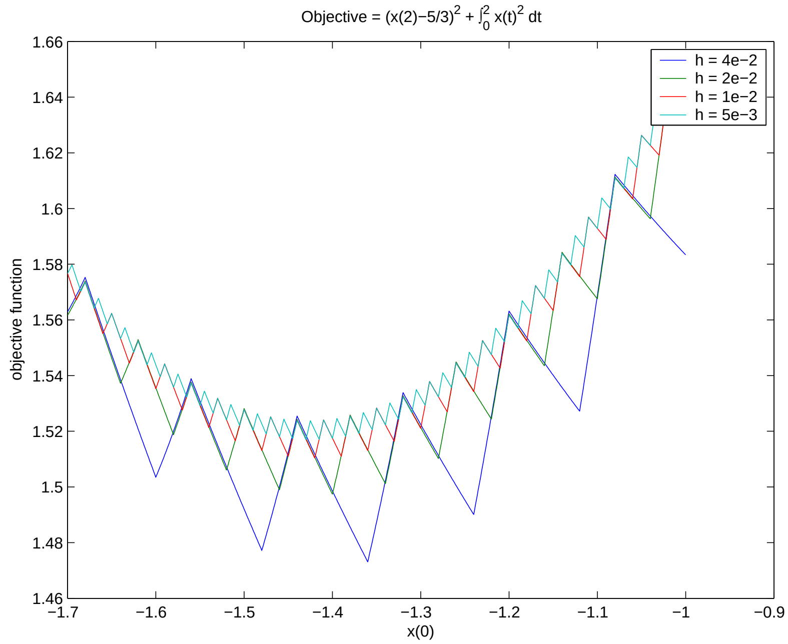

FIG. 2.1. *Computed value of* $(x(2) - 5/3)^2 + \int_0^2 x(t)^2\,dt$ *against* x_0 *for various step-sizes*

This differential inclusion should be easier to handle since it crosses the discontinuity, rather than staying on it as occurs in (2.1).

The discretization (2.2) for (2.3) is

$$x^{k+1} \in x^k + h\,(1+\alpha) - h\,\mathrm{Sgn}(x^k + \theta(x^{k+1} - x^k)). \tag{2.4}$$

If $x^k + \theta(x^{k+1} - x^k) < 0$ then $x^{k+1} = x^k + (2+\alpha)h$; if $x^k + \theta(x^{k+1} - x^k) > 0$ then $x^{k+1} = x^k + \alpha\, h$; if $x^k + \theta(x^{k+1} - x^k) = 0$ then $x^{k+1} = -((1-\theta)/\theta)x^k$. Inserting these formulas for $x^k + 1$ into the first two conditions gives: If $x^k + \theta(2+\alpha)h < 0$ then $x^{k+1} = x^k + (2+\alpha)h$; if $x^k + \theta\,\alpha\,h > 0$ then $x^{k+1} = x^k + \alpha\,h$. Neither of these occurs if $x^k \in -\theta\,h\,[\alpha,\, 2+\alpha]$, where $x^{k+1} = -((1-\theta)/\theta)x^k$. Note that $\partial x^{k+1}/\partial x^k$ is either $-(1-\theta)/\theta$ or one. If $\theta = 1$ (for fully implicit Euler), then $\partial x^{k+1}/\partial x^k$ is either zero or one. This gives the approximations to $\partial x(2)/\partial x_0$ computed from differentiating the numerical solutions of either zero or one.

Now consider $\frac{1}{2} < \theta < 1$. If $x^k \in -h\theta\,[\alpha,\, 2+\alpha]$, then $x^{k+1} = -(1-\theta)\,x^k/\theta > 0$, and so $x^{k+1} + \theta\,\alpha\,h > 0$ and $x^{k+2} = x^{k+1} + \alpha\,h \geq x^{k+1} > 0$, and so on. Thus there can be at most one k where $x^k \in -h\theta\,[\alpha,\, 2+\alpha]$. This means that the approximations to $\partial x(2)/\partial x_0$ obtained by differentiating the numerical solutions is either $-(1-\theta)/\theta$ or one. Either of these answers is clearly far from the correct answer of $\alpha/(2+\alpha)$. Furthermore, the formulas depend on completely different quantities.

As a more explicit example, consider the following results which involve the above differential inclusion with $\alpha = 1$ and numerical solutions computed using $\theta = 1$ (i.e., the fully implicit Euler method). In Figure 2.1 the objective function is $(x(2) - 5/3)^2 + \int_0^2 x(t)^2\,dt$ is plotted against $x(0) = x_0$ with the integral computed using the trapezoidal rule. The trapezoidal method should contribute only $O(h^2)$ error compared with the now $O(h)$ for the errors due to the implicit Euler method. As can

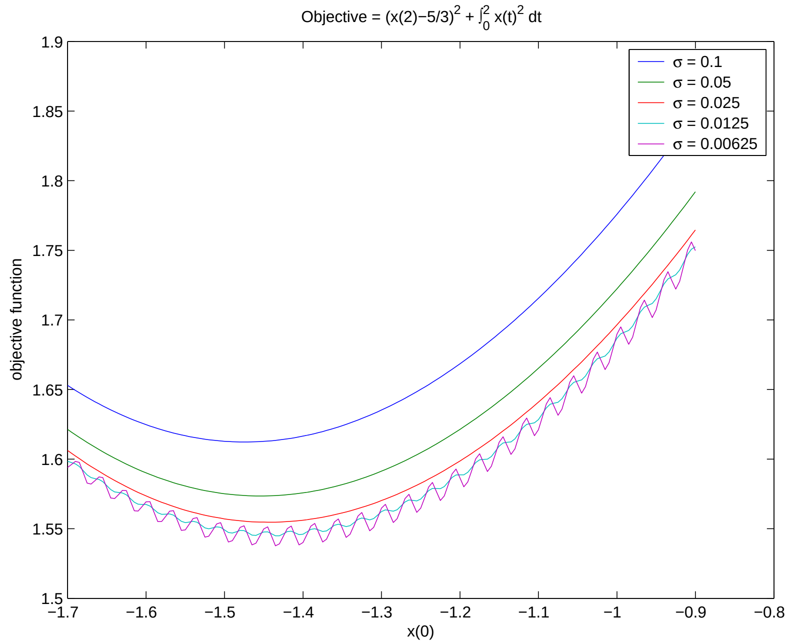

FIG. 2.2. *Computed value of $(x(2) - 5/3)^2 + \int_0^2 x(t)^2\,dt$ against x_0 for various smoothing parameters ($h = 10^{-2}$)*

be clearly seen in Figure 2.1, the values of the computed objective function converge, but the gradients do not. Smoothing the right-hand side of the differential equation (in this case replacing $\mathrm{Sgn}(x)$ with $\mathrm{Sgn}_\sigma(x) := \tanh(x/\sigma)$) improves things greatly regarding the value of gradients, as can be seen in Figure 2.2.

The results in Figure 2.2 also point out the following fact about systems whose dynamics is not necessarily nonsmooth, but that are stiff enough to behave like an almost nonsmooth system, with $\sigma \neq 0$, but $\sigma \approx 0$. Since such systems are stiff, it is likely that the favorite way of simulating them is to use an implicit time stepping scheme with a relatively large time step. If the derivative is computed on the same grid, the resulting optimization problem will have many local minima, not necessarily close to the target minimum. These disappear only when the time step is $o(1/L)$, where the right-hand side has a local Lipschitz constant of L.

3. The model differential inequality. Consider the DI

$$\frac{dx}{dt} \in \begin{cases} \{f_1(x)\}, & \psi(x) < 0, \\ \{f_2(x)\}, & \psi(x) > 0, \\ \mathrm{co}\,\{f_1(x),\, f_2(x)\}, & \psi(x) = 0. \end{cases} \tag{3.1}$$

We assume that this right-hand side satisfies a one-sided Lipschitz condition (1.3). We note that (3.1) can be in one of the following nedegenerate switching cases.

1. Either we have $\nabla\psi(x)\cdot f_1(x)$, $\nabla\psi(x)\cdot f_2(x) > 0$, whenever $\psi(x) = 0$. In this case $d\psi(x(t))/dt$ is strictly increasing before and after the switching time. So the dynamical system described by (3.1) will switch from the set $\psi(x) < 0$ to the set $\psi(x) > 0$.

2. Or we have that $\nabla\psi(x) \cdot f_1(x) > 0$, $\nabla\psi(x) \cdot f_2(x) < 0$, whenever $\psi(x) = 0$. In this case the dynamical system is trapped in the manifold $\psi(x) = 0$, once it reaches it.

Either of these cases will result in jumps in the sensitivities and the adjoint variables, as we show in the following sections. In addition to that case we have the situation where the systems "exits" a singularity, which may occur starting from the second case with $\nabla\psi(x)f_2(x)$ changes sign. In that case, however, there is no discontinuous transition, and the case does not need to be studied separately.

We will address the properties of both cases below.

3.1. Consequences of the one-sided Lipschitz condition. In this subsection we show that the one-sided Lipschitz condition, which is satisfied by Coulomb-friction force laws, enforces certain constraints on the functions f_1 and f_2 which will be used later.

LEMMA 3.1. *Suppose that the right-hand side of (3.1) satisfies the one-sided Lipschitz condition (1.3). Suppose also that ψ is differentiable and $\nabla\psi(x) \neq 0$ for any x where $\psi(x) = 0$. Then on the discontinuity $\Sigma = \{\, x \mid \psi(x) = 0 \,\}$ we must have $(f_2 - f_1) \parallel \nabla\psi^T$ and $\nabla\psi \cdot (f_2 - f_1) \leq 0$.*

Proof. Pick $\eta > 0$, $\epsilon > 0$, and $x \in \Sigma$. Consider any $0 \neq \zeta \perp \nabla\psi(x)^T$. Now put $x_1 = x - \epsilon\nabla\psi(x)^T$ and $x_2 = x + \epsilon\nabla\psi(x)^T + \eta\zeta$. We want to choose ϵ, $\eta > 0$ small enough so that $\psi(x_1) < 0$ and $\psi(x_2) > 0$. Now $\psi(x_1) = \psi(x) - \epsilon\|\nabla\psi(x)\|^2 + o(\epsilon)$, so for any $\epsilon > 0$ sufficiently small, $\psi(x_1) < 0$. Also $\psi(x_2) = \psi(x) + \epsilon\|\nabla\psi(x)\|^2 + o(\epsilon+\eta)$. So for any $\theta > 0$ there is an $\eta_0 > 0$ so that $\epsilon + \eta < \eta_0$ implies that the remainder term $o(\epsilon + \eta)$ is less than $\theta(\epsilon + \eta)$. For such ϵ and η,

$$\psi(x_2) \geq \epsilon\|\nabla\psi(x)\|^2 - \theta(\epsilon + \eta).$$

Provided $0 < \eta < (\|\nabla\psi(x)\|^2 - \theta)\epsilon/\theta$ and ϵ and η are sufficiently small, we have $\psi(x_2) > 0$. If we set $\eta = \frac{1}{2}(\|\nabla\psi(x)\|^2 - \theta)\epsilon/\theta$, then for sufficiently small $\epsilon > 0$, $\psi(x_2) > 0$. Turning this around, if we set $\epsilon = 2\theta\eta/(\|\nabla\psi(x)\|^2 - \theta)$, then for sufficiently small $\eta > 0$ and $0 < \theta \leq 1$, $\psi(x_1) < 0$ and $\psi(x_2) > 0$. Choose $\theta > 0$ sufficiently small so that $\epsilon \leq \eta$.

Now by the one-sided Lipschitz condition, for sufficiently small $\epsilon > 0$ and $\eta > 0$ given as above,

$$\begin{aligned} L\|x_2 - x_1\|^2 &\geq (x_2 - x_1)^T(f_2(x_2) - f_1(x_1)) \\ &= (2\epsilon\nabla\psi(x) + \eta\zeta^T)[f_2(x) - f_1(x) + O(\epsilon + \eta)]. \end{aligned} \tag{3.2}$$

Since $\|x_2 - x_1\| = O(\eta)$, after dividing both sides of (3.2) by η and taking $\eta \downarrow 0$ gives

$$0 \geq (2(\epsilon/\eta)\nabla\psi(x) + \zeta^T)[f_2(x) - f_1(x)]. \tag{3.3}$$

But $\epsilon/\eta = 2\theta/(\|\nabla\psi(x)\|^2 - \theta)$. So

$$0 \geq (2\theta/(\|\nabla\psi(x)\|^2 - \theta)\nabla\psi(x) + \zeta^T)[f_2(x) - f_1(x)],$$

for all $\theta > 0$ sufficiently small. Taking $\theta \downarrow 0$ gives the result that for any $x \in \Sigma$, $\zeta^T[f_2(x) - f_1(x)] \leq 0$. Since $-\zeta$ is also perpendicular to $\nabla\psi(x)^T$, it follows that $f_2(x) - f_1(x)$ is perpendicular to ζ. Noting that ζ is an arbitrary vector perpendicular to $\nabla\psi(x)^T$, we see that $f_2(x) - f_1(x) \parallel \nabla\psi(x)^T$.

To prove the final assertion, set $\zeta = 0$. Then for sufficiently small $\epsilon > 0$, $2\epsilon\nabla\psi(x)[f_2(x) - f_1(x) + O(\epsilon)] \leq 0$. Taking $\epsilon \downarrow 0$ gives $\nabla\psi(x)[f_2(x) - f_1(x)] \leq 0$. ∎

3.1.1. The equivalent equation in the trapped case. We now work under the assumption that $\nabla\psi(x)f_1(x) > 0$ and $\nabla\psi(x)f_2(x) < 0$, that is, whem the problem is "trapped" in the manifold $\psi(x) = 0$, where we must follow the trajectory.

From the third branch in (3.1), the problem of following the trajectory in the discontinuity should be expressed as

$$\frac{dx}{dt} = f^*(x) = (1 - \theta(x))\, f_1(x) + \theta(x) f_2(x).$$

The unknown weighting function $\theta(x)$ is computable from the condition that the mapping $\psi(x)$ is an invariant of the dynamical system defined by $f^*(x)$, that is $\nabla\psi(x)f^*(x) = 0$. In turn, the last equation leads to

$$\nabla\psi(x)\,((1 - \theta(x))\, f_1(x) + \theta(x) f_2(x)) = 0$$

Solving for the unknown weighing function from the last equation we obtain that

$$\theta(x) = \frac{\nabla\psi(x)f_1(x)}{\nabla\psi(x)f_1(x) - \nabla\psi(x)f_2(x)} \tag{3.4}$$

Note that, under the assumption that $\nabla\psi(x)f_1(x) > 0$ and $\nabla\psi(x)f_2(x) < 0$ we must have that

$$0 < \theta(x) < 1$$

at least in a neighborhood of the switching point.

This also shows that the dynamical system will exit the discontinuity manifold only when the weighing function $\theta(x)$ switches to either 0 or 1. From the expression of $\theta(x)$, it follows that such a switch can happen only when either $\nabla\psi(x)f_1(x)$ or $\nabla\psi(x)f_2(x)$ will switch signs.

With this identification, the dynamical system can, in effect, be represented by the following piecewise differential equation of

$$\dot{x} = \begin{cases} f_1(x), & \psi(x) < 0 \\ f^*(x), & \psi(x) = 0 \end{cases} \qquad \dot{x}(0) = x_0$$

3.2. Exact sensitivities with respect to parameters. We are interested in evaluating the sensitivities with respect the parameters, that can be easily incorporated in sensitivities with respect to the initial conditions. Let $x(t, x_0)$ be the value of $x(t)$ where x is the solution of the discontinuous ODE with initial value x_0.

3.2.1. The trapped case. We first compute these sensitivities in the case where the trajectory is trapped in the manifold $\psi(x) = 0$, that is $\nabla\psi(x)f_1(x) > 0$ and $\nabla\psi(x)f_2(x) < 0$.

An important component in computing this sensitivities is the switching time t_s, that can be defined implicitly by the equations

$$\dot{x}(0, x_0) = x_0,\ \dot{x}(t, x_0) = f_1(x),\ \psi(x(t_s, x_0)) = 0.$$

The sensitivity $s(t, x_0)$ satisfies the following equation, before switching.

$$\dot{s} = \nabla f_1(x(t, x_0))s,\ s(0) = I.$$

Using the implicit function theorem, we obtain that

$$\frac{\partial t_s}{\partial x_0} = -\frac{\nabla\psi(x(t_s,x_0))s(t_s,x_0)}{\nabla\psi(x(t_s,x_0))f_1\left(x(t_s,x_0)\right)}$$

From our assumptions, it is immediate that $\nabla\psi(x)f_1(x) \neq 0$.

To determining the equation satisfied by the sensitivity after switching, we proceed in two steps. First we consider the equation that is satisfied by the system once it enters the discontinuity

$$\dot{y}(t,y_0) = f^{*}\left(y(t,y_0)\right), \quad \dot{y}(0,y_0) = y_0$$

and we analyze its sensitivity $s_2(t,y_0) = \partial y/\partial y_0$ with respect to the parameter y_0. Here we have used the identification, that is valid when $t > t_s$,

$$y(t,y_0) = x(t+t_s,x_0).$$

We obtain the following linear differential equation

$$\dot{s}_2 = \nabla f^{*}(y(t,x_0))s_2, \;\; s_2(0) = I.$$

To compute $s_2(t,y_0) = \partial x/\partial x_0$, we glue the solutions before and after reaching to discontinuity by using that

$$y_0 = x(t_s,x_0)$$

We get that

$$\begin{aligned}
\frac{\partial y_0}{\partial x_0} &= \frac{dx(t_s,x_0)}{dx_0} = f_1\left(x(t_s,x_0)\right)\frac{\partial t_s}{\partial x_0} + s(t_s,x_0) \\
&= -f_1\left(x(t_s,x_0)\right)\frac{\nabla\psi(x(t_s,x_0))s(t_s,x_0)}{\nabla\psi(x(t_s,x_0))^T f_1\left(x(t_s,x_0)\right)} + s(t_s,x_0) \\
&= \left[I - \frac{f_1\left(x(t_s,x_0)\right)\nabla\psi(x(t_s,x_0))}{\nabla\psi(x(t_s,x_0))f_1\left(x(t_s,x_0)\right)}\right]s(t_s,x_0)
\end{aligned}$$

The following computation also shows that $x(t,x_0)$ is a differentiable function of x_0 and that its derivative $s(t)$ obeys the following differential equation

$$\dot{s}(t) = \begin{cases} \nabla f_1(x(t,x_0))s(t), & t < t_s \\ \nabla f^{*}(x(t,x_0))s(t), & t > t_s \end{cases}$$

To figure out the jump rule at the switching, we use that $x(t,x_0) = y(t-t_s,y_0)$.

We obtain that, whenever $t > t_s$, the following holds

$$\frac{\partial x(t,x_0)}{\partial x_0} = \frac{dy(t-t_s,y_0)}{dx_0} = -f^{*}(y(t-t_s,y_0))\frac{\partial t_s}{\partial x_0} + s_2(t-t_s,y_0)\frac{\partial y_0}{\partial x_0}$$

As $t \downarrow t_s$, we have that $s_2(t-t_s,y_0)$ approaches the identity. Using our computation for $\partial t_s/\partial x_0$ and $\partial y_0/\partial x_0$, we obtain that, at the switching point, the sensitivity will jump according to the rule

$$s(t_s^{+}) = \left[I + \frac{\left(f^{*}\left(x(t_s,x_0)\right) - f_1\left(x(t_s,x_0)\right)\right)\nabla\psi(x(t_s,x_0))}{\nabla\psi(x(t_s,x_0))f_1\left(x(t_s,x_0)\right)}\right]s(t_s^{-}).$$

If we replace the expression for f^* in the above equation, we obtain that

$$s(t_s^+) = \left[I + \frac{\left(f_2\left(x(t_s,x_0)\right) - f_1\left(x(t_s,x_0)\right)\right) \nabla\psi(x(t_s,x_0))}{\nabla\psi(x(t_s,x_0))\left(f_1\left(x(t_s,x_0)\right) - f_2\left(x(t_s,x_0)\right)\right)} \right] s(t_s^-). \tag{3.5}$$

From Lemma (3.1) we have that $(f_2 - f_1) \parallel \nabla\psi^T$, which, in turn, implies that the matrix in the above relation is an orthogonal projection.

3.2.2. The case where $\nabla\psi(x)f_1(x) > 0$ and $\nabla\psi(x)f_2(x) > 0$. We can immediately see that the previous argument used only the fact that

$$\frac{dx}{dt} \in \begin{cases} \{f_1(x)\}, & t < t_s, \\ \{f^*(x)\}, & t > t_s, \end{cases}$$

Replacing $f^*(x)$ by $f_2(x)$ in the preceding analysis we obtain that

$$s(t_s^+) = \left[I + \frac{\left(f_2\left(x(t_s,x_0)\right) - f_1\left(x(t_s,x_0)\right)\right) \nabla\psi(x(t_s,x_0))}{\nabla\psi(x(t_s,x_0))f_1\left(x(t_s,x_0)\right)} \right] s(t_s^-) \tag{3.6}$$

4. A smoothing approach. We now investigate the approximation of (3.1) by the smoothed system

$$\frac{dx_\sigma}{dt} = \varphi_\sigma(\psi(x_\sigma))\, f_2(x_\sigma) + (1 - \varphi_\sigma(\psi(x_\sigma)))\, f_1(x_\sigma). \tag{4.1}$$

Here

$$\varphi_\sigma(w) = \int_{-\infty}^{w} \theta_\sigma(r')\, dr' \tag{4.2}$$

where $\theta_\sigma(r) = (1/\sigma)\theta(r/\sigma)$ and $\theta \geq 0$, $\operatorname{supp}\theta = [-1, +1]$, and $\int_{-\infty}^{+\infty} \theta(r)\, dr = 1$. Thus, $\varphi_\sigma(w) = 0$ for $w \leq -\sigma$ and $\varphi_\sigma(w) = 1$ for $w \geq +\sigma$. Note that the smoothed differential equation (4.1) is identical with the original discontinuous differential equation (3.1) unless $-\sigma \leq \psi(x) \leq +\sigma$. The set $\{\, x \mid -\sigma \leq \psi(x) \leq +\sigma \,\}$ is called the *transition region.*

Note that in the following convergence results, we will use all five asymptotic order symbols: O, o, Ω, ω and Θ. Recall that $f(s) = O(g(s))$ as $s \to 0$ means that there is a constant C and s_0 where $0 < |s| < s_0$ implies $|f(s)| \leq C\, g(s)$, $f(s) = o(g(s))$ means that $f(s)/g(s) \to 0$ as $s \to 0$, $f(s) = \Omega(g(s))$ means that $g(s) = O(f(s))$, $f(s) = \omega(g(s))$ means that $g(s) = o(f(s))$, and $f(s) = \Theta(g(s))$ means that $f(s) = O(g(s))$ and $f(s) = \Omega(g(s))$.

4.1. One-sided Lipschitz condition for the smoothed system. Lemma 3.1 is useful for showing that the smoothed right-hand side f_σ in (4.1) also satisfies a one-sided Lipschitz condition, although the Lipschitz constant might not be the same as for (3.1). To show this we do need to assume that f_1 and f_2 satisfy an ordinary ("two-sided") Lipschitz condition with constant L_f and that $\nabla\psi$ is also Lipschitz with constant $L_{\nabla\psi}$. As usual, L is the one-sided Lipschitz constant for (3.1). That means that both f_1 and f_2 satisfy the one-sided Lipschitz condition (1.3) with constant L. Since $\Sigma = \{\, x \mid \psi(x) = 0 \,\}$ is a C^1 manifold, there is a continuous "nearest point" map $\pi(x) =$ the nearest point in Σ to x. We can choose a $\sigma_0 > 0$ so that if $0 < \sigma < \sigma_0$, this map is well-defined on the transition region.

We will show that for any $w \in \mathbb{R}^n$, $w^T \nabla f_\sigma(x) w \leq L\|w\|^2$. Note that $w^T \nabla f_1(x) w$, $w^T \nabla f_2(x) w \leq L\|w\|^2$ for all w. Outside the transition region we have $\nabla f_\sigma(x) =$

$\nabla f_1(x)$ or $\nabla f_\sigma(x) = \nabla f_2(x)$, and the desired property of ∇f_σ follows immediately. Inside the transition region, we have

$$\nabla f_\sigma = (1-\varphi_\sigma)\nabla f_1 + \varphi_\sigma \nabla f_2 + (f_2 - f_1)\varphi_\sigma'(\psi)\,\nabla\psi.$$

Thus

$$\begin{aligned} w^T\nabla f_\sigma(x)w &= (1-\varphi_\sigma(\psi(x))w^T\nabla f_1(x)w + \varphi_\sigma(\psi(x))w^T\nabla f_2(x)w \\ &\quad + \varphi_\sigma'(\psi(x))w^T(f_2(x)-f_1(x))\nabla\psi(x)w \\ &\le L\|w\|^2 + \varphi_\sigma'(\psi(x))w^T(f_2(\pi(x)) - f_1(\pi(x)))\nabla\psi(\pi(x))w \\ &\quad + O(1/\sigma)\,O(\|\pi(x)-x\|)\,\|w\|^2. \end{aligned}$$

But the transition region is only $O(\sigma)$ wide, so $\|\pi(x)-x\| = O(\sigma)$. Thus, using Lemma 3.1, and the fact that $\phi_\sigma'(\cdot) \ge 0$, we obtain that

$$w^T\nabla f_\sigma(x)w \le O(1)\,\|w\|^2.$$

Thus f_σ satisfies a one-sided Lipschitz condition, although its one-sided Lipschitz constant may be considerably larger than for f.

5. Convergence of the gradients for the case $\nabla\psi(x)f_1(x) > 0$ and $\nabla\psi(x)f_2(x) > 0$. We now analyze the asymptotic properties, as $\sigma \to 0$ for the case where $\nabla\psi(x)f_1(x) > 0$ and $\nabla\psi(x)f_2(x) > 0$, that is, the case where the trajectory switches from $\psi(x) < 0$ to $\psi(x) > 0$.

5.1. The variational equation of the smoothed differential equation. The variational equation for the smoothed system can be easily written down:

$$\frac{ds_\sigma}{dt} = \{(1-\varphi_\sigma)\nabla f_1 + \varphi_\sigma\nabla f_2 + (f_2-f_1)\varphi_\sigma'(\psi)\,\nabla\psi\}\ s_\sigma. \tag{5.1}$$

For smooth f_1, f_2, the first two terms $\varphi_\sigma\,\nabla f_2$ and $(1-\varphi_\sigma)\,\nabla f_1$ in the braces of equation (5.1) are bounded, but the last term $(f_1 - f_2)\varphi_\sigma'(\psi)\,\nabla g$ might not be bounded. Thus the limiting equation as $\sigma \downarrow 0$ for $\psi(x(t)) \neq 0$ becomes

$$\frac{ds}{dt} = \left\{ \begin{array}{ll} \nabla f_1, & \psi(x) < 0 \\ \nabla f_2, & \psi(x) > 0 \end{array} \right\} s. \tag{5.2}$$

However, this ignores what happens near $\psi(x(t^*)) = 0$. From (3.6) we have determined what happens for the original system (3.1), but to complete a proof of convergence of s_σ to s we must also prove that the jumps match.

Suppose that the limiting solution (which is unique by the one-sided Lipschitz assumption) reaches the surface $\psi(x(t)) = 0$ at time $t = t^*$. By our assumptions that $\psi(x) = 0$ implies $\nabla\psi(x)\cdot f_1(x) > 0$ and $\nabla\psi(x)\cdot f_2(x) > 0$, there can only be one time $t = t^*$ where $\psi(x(t)) = 0$. Put $x^* = x(t^*)$, $f_1^* = f_1(x^*)$, $f_2^* = f_2(x^*)$, $\nabla\psi^* = \nabla\psi(x^*)$. Note that if $t \approx t^*$ and $\sigma \approx 0$, then $x_\sigma(t) \approx x^*$.

Of particular interest to us is the fact that the term $(f_1 - f_2)\varphi_\sigma'(\psi)\nabla\psi$ is unbounded. Although we expect that $\varphi_\sigma'(\psi) \neq 0$ only for a time interval of length $O(\sigma)$, $\varphi_\sigma'(\psi)$ has a magnitude of $O(1/\sigma)$. In the limit as $\sigma \downarrow 0$, this could correspond to a Dirac-δ function. This can be interpreted in the sense of [10]. Since the matrix $(f_2(x_\sigma(t)) - f_1(x_\sigma(t)))\,\nabla\psi(x_\sigma(t))) \to (f_2^* - f_1^*)\,\nabla\psi^*$ as $\sigma \downarrow 0$ in the relevant time interval(s) ($\varphi'_\sigma(x_\sigma(t)) \neq 0$), in the limit the effect of this term is to include a factor of the form

$$\exp\left(\alpha\,(f_2^* - f_1^*)\,\nabla\psi^*\right) \tag{5.3}$$

where α is the limit of $\int \varphi'_\sigma(\psi(x_\sigma(t)))\,dt$. We will show that this limit exists, and will give a simple formula for it and the matrix exponential (5.3).

Now for $-\sigma \leq \psi(x_\sigma(t)) \leq +\sigma$ we have $\|x_\sigma(t) - x^*\| = O(\sigma)$. Then we can write

$$\begin{aligned}\frac{d}{dt}\psi(x_\sigma(t)) &= \nabla\psi(x_\sigma(t)) \cdot \dot{x}_\sigma(t) \\ &= \varphi_\sigma(\psi(x_\sigma(t)))\nabla\psi(x_\sigma(t)) \cdot f_2(x_\sigma(t)) + (1 - \varphi_\sigma(\psi(x_\sigma(t))))\nabla\psi(x_\sigma(t)) \cdot f_1(x_\sigma(t)) \\ &= \varphi_\sigma(\psi(x_\sigma(t)))\nabla\psi^* \cdot f_2^* + (1 - \varphi_\sigma(\psi(x_\sigma(t)))\nabla\psi^* \cdot f_1^* + O(\sigma).\end{aligned}$$

Put $\gamma_i = \nabla\psi^* \cdot f_i^*$, $i = 1, 2$. Then we can write

$$\frac{d\psi}{dt} = \varphi_\sigma(x_\sigma(t))\gamma_2 + (1 - \varphi_\sigma(x_\sigma(t)))\gamma_1 + O(\sigma).$$

For sufficiently small $\sigma > 0$, we have $d\psi/dt > 0$, so we can use a change of variables

Returning to the value of α in (5.3), we consider the integrals

$$\begin{aligned}\int_{-\infty}^{+\infty} \varphi'_\sigma(\psi(x_\sigma(t)))\,dt &= \int_{-\sigma}^{+\sigma} \frac{\varphi'_\sigma(\psi)}{\gamma_1 + \varphi_\sigma(\psi)(\gamma_2 - \gamma_1) + O(\sigma)}\,d\psi \\ &= \frac{1}{\gamma_2 - \gamma_1}\int_{\gamma_1}^{\gamma_2} \frac{dw}{w} + O(\sigma) \qquad (\text{using } w = \gamma_1 + \varphi_\sigma(\psi)(\gamma_2 - \gamma_1)\,) \\ &= \frac{\ln(\gamma_2/\gamma_1)}{\gamma_2 - \gamma_1} + O(\sigma).\end{aligned}$$

Thus we obtain the value for the limit of $\alpha = \ln(\gamma_2/\gamma_1)/(\gamma_2 - \gamma_1)$. To compute the matrix exponential (5.3), we resort to the series definition of the matrix exponential. To simplify notation, put $u = f_2^* - f_1^*$ and $v^T = \nabla\psi^*$. Then

$$\begin{aligned}\exp(\alpha u v^T) &= I + \sum_{k=1}^{\infty} \frac{1}{k!}\alpha^k (uv^T)^k = I + \sum_{k=1}^{\infty} \frac{1}{k!}\alpha^k (v^T u)^{k-1} uv^T \\ &= I + \frac{1}{v^T u}\sum_{k=1}^{\infty} \frac{1}{k!}(\alpha\, v^T u)^k\, uv^T = I + \frac{uv^T}{v^T u}\left[e^{\alpha\, v^T u} - 1\right].\end{aligned}$$

Substituting for u and v we see that the limiting matrix exponential is

$$\begin{aligned}\exp(\alpha(f_2^* - f_1^*)\nabla\psi^*) &= I + \frac{(f_2^* - f_1^*)\nabla\psi^*}{\gamma_2 - \gamma_1}\left[\exp((\gamma_2 - \gamma_1)\ln(\gamma_2/\gamma_1)/(\gamma_2 - \gamma_1)) - 1\right] \\ &= I + \frac{(f_2^* - f_1^*)\nabla\psi^*}{\gamma_2 - \gamma_1}\left[\frac{\gamma_2}{\gamma_1} - 1\right] = I + \frac{(f_2^* - f_1^*)\nabla\psi^*}{\gamma_2 - \gamma_1}\,\frac{\gamma_2 - \gamma_1}{\gamma_1} \\ &= I + \frac{(f_2^* - f_1^*)\nabla\psi^*}{\gamma_1}.\end{aligned}$$

We have thus proved the following result

THEOREM 5.1. *For the case where $\psi(x) = 0$ implies that $\psi(x)f_1(x) > 0$, and $\nabla\psi(x)f_1(x) > 0$, the sensitivity of the solution of the smoothed problem, s_σ, approaches the sensitivity of the solution of the original problem, s, as $\sigma \to 0$.*

Proof. Follows by comparing the right hand side of the last displayed equality with (3.6), as well as our conclusion that the right hand side of (5.1) converges to (5.2) away from the switching time t^*. ∎

5.2. Lagrange multipliers and the jump rule. There is another way to this result via the adjoint equation from the Pontryagin conditions. Since this is a problem without control functions, we consider the problem of minimizing some objective function $g(x(T))$ by varying the initial value x_0. Again we consider using a smoothed right-hand side f_σ (4.1).

From the conventional Pontryagin conditions [26, 7, 21, 5] we have the adjoint equations

$$\frac{d\lambda_\sigma}{dt} = -\nabla f_\sigma(x_\sigma(t))^T \lambda_\sigma, \qquad \lambda_\sigma(T) = \nabla g(x_\sigma(T)). \tag{5.4}$$

As above, we note that

$$\begin{aligned}\nabla f_\sigma(x) = {} & \frac{\varphi'(\psi(x)/\sigma)}{\sigma}[f_2(x) - f_1(x)]\nabla\psi(x) \\ & + \{\varphi(\psi(x)/\sigma)\,\nabla f_1(x) + (1 - \varphi(\psi(x)/\sigma))\,\nabla f_2(x)\}\,,\end{aligned}$$

and that the terms enclosed in $\{\cdots\}$ are bounded as $\sigma \downarrow 0$. Integrating (5.4) backwards in time and using the matrix exponential, we get the approximation around $t = t^*$ where $\psi(x(t^*)) = 0$:

$$\lambda_\sigma(t^* - \epsilon) = \left[I + \frac{e^{\alpha(\gamma_2-\gamma_1)} - 1}{\gamma_2 - \gamma_1}(f_2^* - f_1^*)\nabla\psi^*\right] \lambda_\sigma(t^* + \epsilon) + O(\epsilon) \tag{5.5}$$

where $\sigma = o(\epsilon)$ as $\epsilon \downarrow 0$, and α is some non-negative quantity (possibly dependent on σ). We can simplify (5.5) by setting $\beta := (e^{\alpha(\gamma_2-\gamma_1)} - 1)/(\gamma_2 - \gamma_1)$. Then we get

$$\lambda_\sigma(t^* - \epsilon) = [I + \beta(f_2^* - f_1^*)\nabla\psi^*]\ \lambda_\sigma(t^* + \epsilon) + O(\epsilon). \tag{5.6}$$

The only quantity that is not determined by this approach is β. But this can be computed from the property that the Hamiltonian $H_\sigma(x_\sigma(t), \lambda_\sigma(t)) := \lambda_\sigma(t)^T f_\sigma(x_\sigma(t))$ is a constant function of t.

If we apply this rule around the crossing time t^*, we get $(f_2^*)^T\lambda_\sigma(t^* + \epsilon) = (f_1^*)^T\lambda_\sigma(t^* - \epsilon) + O(\epsilon)$. We can then write an estimate for $\lambda_\sigma(t^* - \epsilon)$ in terms of $\lambda_\sigma(t^* + \epsilon)$:

$$(f_2^*)^T\lambda_\sigma(t^* + \epsilon) = (f_1^*)^T[I + \beta(f_2^* - f_1^*)\nabla\psi^*]\lambda_\sigma(t^* + \epsilon) + O(\epsilon).$$

Solving for β gives $(f_2^* - f_1^*)^T\nabla\psi^* = O(\epsilon)$ or

$$\beta = \frac{1}{\nabla\psi^* f_1^*} + O(\epsilon). \tag{5.7}$$

Note that if $\nabla\psi^*(f_2^* - f_1^*) = O(\epsilon)$, then $\lambda_\sigma(t^* - \epsilon) = \lambda_\sigma(t^* + \epsilon) + O(\epsilon)$. So

$$\lambda_\sigma(t^* - \epsilon) = \left[I + \frac{(f_2^* - f_1^*)\nabla\psi^*}{\nabla\psi^* f_1^*}\right] \lambda_\sigma(t^* + \epsilon) + O(\epsilon). \tag{5.8}$$

Taking $\epsilon \downarrow 0$ and $\sigma = o(\epsilon)$ this gives a simple "jump rule" for the adjoint variables:

$$\lambda(t^{*-}) = \left[I + \frac{(f_2^* - f_1^*)\nabla\psi^*}{\nabla\psi^* f_1^*}\right] \lambda(t^{*+}). \tag{5.9}$$

That the adjoint functions have discontinuities in problems with discontinuous right-hand sides was noted by, for example, Driessen and Sadegh [12], and Kim and Ha [23].

5.3. Convergence results. We now prove the main convergence result for this case.

THEOREM 5.2. *Assume that $\nabla\psi(x)f_1(x) > 0$, $\nabla\psi(x)f_2(x) > 0$, whenever $\psi(x) = 0$. Assume that we integrate the smoothed model equation (4.1) and the corresponding sensitivity equation (5.1) for $\partial x/\partial x_0$ using Euler's method with a time step $h = o(\sigma)$. Then, the numerical sensitivites and the numerical adjoints converge to the sensivities of the original problem as $\sigma \to 0$.*

We separate the proof in the following parts:

1. In Subsection 5.3.1 we prove that the sequence of state variables produced by Euler's method applied to the smoothed equation converge to the one of the model problem (3.1).
2. In Subsection 5.3.2 we prove that the sequence of adjoint variables is convergent.
3. In Subsection 5.3.3 we prove that the sequence of sensitivities is convergent.

5.3.1. Errors in the computed trajectory. Consider using the explicit Euler method for the numerical solution of $dx_\sigma/dt = f_\sigma(x_\sigma)$. Let $t_k = t_0 + k\,h$, where $h > 0$ is the step size.

$$\begin{aligned} x_\sigma^{k+1} &= x_\sigma^k + h\, f_\sigma(x_\sigma^k) \\ x_\sigma(t_{k+1}) &= x_\sigma(t_k) + h\, f_\sigma(x_\sigma(t_k)) + \frac{1}{2}h^2\xi_k \end{aligned}$$

where $\|\xi_k\| \le \frac{1}{2}h^2 \max_{t_k\le t\le t_{k+1}} \|x_\sigma''(t)\|$ by Taylor's theorem to 2nd order. We suppose that $h\,L < 1$. Subtracting the equations for x_σ^{k+1} and $x_\sigma(t_{k+1})$ gives

$$x_\sigma(t_{k+1}) - x_\sigma^{k+1} = (x_\sigma(t_k) - x_\sigma^k) + h\,[f_\sigma(x_\sigma(t_k)) - f_\sigma(x_\sigma^k)] + \frac{1}{2}h^2\xi_k. \tag{5.10}$$

For all k, put $e_{\sigma,k} = x_\sigma(t_k) - x_\sigma^k$. Then

$$e_{\sigma,k+1} = e_{\sigma,k} + h\,[f_\sigma(x_\sigma^k + e_{\sigma,k}) - f_\sigma(x_\sigma^k)] + \frac{1}{2}h^2\xi_k. \tag{5.11}$$

LEMMA 5.3. *Under our standing assumptions, the map $z \mapsto z + h[f_\sigma(x+z) - f_\sigma(x)]$ is Lipschitz with constant $1 + hL + Ch^2/\sigma^2$ for some constant C independent of h and σ.*

Proof. Let $\Phi_{h,\sigma,x}(z) = z + h[f_\sigma(x+z) - f_\sigma(x)]$. Then for any z_1, z_2,

$$\begin{aligned} &\|\Phi_{h,\sigma,x}(z_1) - \Phi_{h,\sigma,x}(z_2)\|^2 \\ &\quad = \|z_1 - z_2 + h[f_\sigma(x+z_1) - f_\sigma(x+z_2)]\|^2 \\ &\quad = \|z_1 - z_2\|^2 + 2h(z_1 - z_2)^T[f_\sigma(x+z_1) - f_\sigma(x+z_2)] + h^2\,\|f_\sigma(x+z_1) - f_\sigma(x+z_2)\|^2 \\ &\quad \le \|z_1 - z_2\|^2 + 2hL\|z_1 - z_2\|^2 + h^2\,\|f_\sigma(x+z_1) - f_\sigma(x+z_2)\|^2 \quad \text{(using (1.3))} \\ &\quad \le (1 + 2hL + Ch^2/\sigma^2)\|z_1 - z_2\|^2 \end{aligned}$$

since f_σ is Lipschitz with constant $C^{1/2}/\sigma$ for some C independent of h or σ. Therefore,

$$\|\Phi_{h,\sigma,x}(z_1) - \Phi_{h,\sigma,x}(z_2)\| \le (1 + 2hL + Ch^2/\sigma^2)^{1/2}\|z_1 - z_2\|.$$

Note that for $w \ge 0$, $(1+w)^{1/2} \le 1 + \frac{1}{2}w$, so

$$\|\Phi_{h,\sigma,x}(z_1) - \Phi_{h,\sigma,x}(z_2)\| \le (1 + hL + Ch^2/(2\sigma^2))\|z_1 - z_2\|,$$

as desired. □

Using Lemma 5.3, we see that

$$\|e_{k+1}\| \le (1 + hL + Ch^2/\sigma^2)\|e_k\| + \frac{1}{2}h^2\|\xi_k\|. \tag{5.12}$$

We can also bound $x''_\sigma(t)$ since

$$\begin{aligned} x''_\sigma(t) &= \frac{d}{dt} f_\sigma(x_\sigma(t)) \\ &= \nabla f_\sigma(x_\sigma(t))\, f_\sigma(x_\sigma(t)), \end{aligned}$$

which immediately gives the bounds

$$\|x''_\sigma(t)\| \le \|\nabla f_\sigma(x_\sigma(t))\|\, \|f_\sigma(x_\sigma(t))\|. \tag{5.13}$$

Using local boundedness of f, we can bound f_σ independently of σ over compact sets. Thus $\|x''_\sigma(t)\| = O(1/\sigma)$.

Thus if $h = o(\sigma^2)$, we can combine these bounds to show that for $L' > L$ and sufficiently small $h > 0$

$$\|e_{k+1}\| \le e^{L'h}\|e_k\| + C\frac{h^2}{\sigma}. \tag{5.14}$$

Using a discrete Gronwall lemma and $e_0 = 0$, we can see that

$$\|e_k\| \le \frac{e^{L'kh} - 1}{L'h} C\frac{h^2}{\sigma} = (e^{L'(t_k - t_0)} - 1)O(h/\sigma). \tag{5.15}$$

This bound can be considerably improved if we know that $\nabla\psi^* f_1^*$, $\nabla\psi^* f_2^* > 0$, as then the time to cross the transition region is $O(\sigma)$. Outside the transition region, the error is $O(h)$ as is well-known [1]. It will take $O(\sigma/h)$ time-steps to cross the transition region under the assumptions that $\nabla\psi^* f_1^*$, $\nabla\psi^* f_2^* > 0$. Suppose we choose $k^* = k^*(\sigma, h)$ so that $x_\sigma(t_{k^*})$ is outside the transition region, but $x_\sigma(t_{k^*+1})$ is inside the transition region, and after $K = K(\sigma, h) = O(\sigma/h)$ steps we find that $x_\sigma(t_{k^*+K})$ is outside the transition region, and the discrete time trajectory remains outside the transition region for a positive time-period (a period whose length does not go to zero as $h \to 0$).

For now we assume only that $h = o(\sigma)$. Then from (5.12) and (5.13) we find that

$$\begin{aligned} \|e_{k^*+K}\| &\le \exp((hL + Ch^2/\sigma^2)K(\sigma, h)) \left[\|e_{k^*}\| + C\,\frac{h^2}{\sigma}K(\sigma, h)\right] \\ &\le \exp(C'(L\sigma + Ch/\sigma))\, [\|e_{k^*}\| + C\,C'\,h] = O(h), \end{aligned}$$

where $K \le C'\sigma/h$, for h, σ small enough.

5.3.2. Errors in the adjoints. Recall that for the Euler method

$$x_\sigma^{k+1} = x_\sigma^k + h\, f_\sigma(x_\sigma^k).$$

Thus a small variation δx_σ^k in x_σ^k results in a variation

$$\delta x_\sigma^{k+1} = [I + h\,\nabla f_\sigma(x_\sigma^k)]\,\delta x_\sigma^k + o(\|\delta x_\sigma^k\|) \tag{5.16}$$

in x_σ^{k+1}. The the discrete sensitivity equations are thus

$$s_\sigma^{k+1} = [I + h\,\nabla f_\sigma(x_\sigma^k)]\, s_\sigma^k,$$

with given $s_\sigma^0 = s^0$.

Alternatively, we can consider the discrete adjoint variables. Suppose we have a function $g\colon \mathbb{R}^n \to \mathbb{R}$ and we wish to determine the gradient of the $g(x^N)$ where $t_f = t_0 + Nh$ with respect to a change in the initial values x^0 where x^N is computed via Euler's method. Let $\Psi_i(x^i) = g(x^N)$ where $x^{k+1} = x^k + h\, f_\sigma(x^k)$ for $k = i, i+1, \ldots, N-1$. Set $\lambda^i = \nabla\Psi_i(x^i)^T$. If $J_k = I + h\,\nabla f_\sigma(x^k)$, for all k, then $J_k = \nabla_{x^k}(x^{k+1})$, so

$$\begin{aligned}\lambda^i &= \nabla\Psi_i(x^i)^T = (\nabla\Psi_{i+1}(x^{i+1})\, J_i)^T \\ &= [I + h\,\nabla f_\sigma(x^i)]^T\, \lambda^{i+1},\end{aligned}$$

and $\lambda^N = \nabla g(x^N)^T$, which are the discrete adjoint equations.

We can investigate the accuracy of either the direct sensitivity equations, or the discrete adjoint equations, to determine the accuracy of the computed gradients.

The adjoint equations for the differential equations are

$$\frac{d\lambda_\sigma}{dt} = -\nabla f_\sigma(x_\sigma(t))^T \lambda_\sigma, \qquad \lambda_\sigma(T) = \nabla g(x_\sigma(T))^T.$$

Thus

$$\lambda_\sigma(t_k) = [I + h\,\nabla f_\sigma(x_\sigma(t_k))^T]\lambda_\sigma(t_{k+1}) + \eta_k, \tag{5.17}$$

$$\lambda_\sigma^k = [I + h\,\nabla f_\sigma(x_\sigma^k)^T]\lambda_\sigma^{k+1}, \tag{5.18}$$

where $\|\eta_k\| \le \frac{1}{2}h^2 \max_{t_k \le t \le t_{k+1}} \|\lambda_\sigma''(t)\|$. We can bound $\lambda_\sigma''(t)$ by differentiating the adjoint equation:

$$\begin{aligned}\lambda_\sigma''(t) &= -\frac{d}{dt}(\nabla f_\sigma(x_\sigma(t))^T \lambda_\sigma(t)) \\ &= -\nabla_x(\nabla f_\sigma(x)^T \lambda_\sigma(t))|_{x=x_\sigma(t)} \frac{dx_\sigma}{dt}(t) - \nabla f_\sigma(x_\sigma(t))^T \frac{d\lambda_\sigma}{dt}(t) \\ &= -\nabla_x(\nabla f_\sigma(x)^T \lambda_\sigma(t))|_{x=x_\sigma(t)} \frac{dx_\sigma}{dt}(t) + [\nabla f_\sigma(x_\sigma(t))^2]^T \lambda_\sigma(t).\end{aligned}$$

Now dx_σ/dt is bounded on finite intervals independently of σ as f_σ is bounded with a one-sided Lipschitz condition. Also λ_σ is bounded independently of σ. Noting that $\nabla f_\sigma(x) = O(1/\sigma)$ and $\nabla\nabla f_\sigma(x) = O(1/\sigma^2)$, we see that $\lambda_\sigma''(t) = O(1/\sigma^2)$ in the transition region. Outside the transition region, $\lambda_\sigma''(t) = O(1)$. Note that for t in a fixed finite interval, the constants implicit in the "O" expressions are independent of t.

Thus η_k is $O(h^2/\sigma^2)$ with a constant independent of h, σ and k where $x_\sigma(t)$ is in the transition region for some $t_k \le t \le t_{k+1}$. Otherwise $\eta_k = O(h^2)$ with a constant independent of h, σ and k.

To obtain bounds on the errors in the adjoints, we first subtract (5.18) from (5.17). This gives

$$\begin{aligned}\lambda_\sigma(t_k) - \lambda_\sigma^k = [I + h\, f_\sigma(x_\sigma(t_k))]^T&[\lambda_\sigma(t_{k+1}) - \lambda_\sigma^{k+1}] \\ &+ h[\nabla f_\sigma(x_\sigma(t_k)) - \nabla f_\sigma(x_\sigma^k)]\lambda_\sigma^{k+1} + \eta_k.\end{aligned}$$

Now $\|\nabla f_\sigma(x_\sigma(t_k)) - \nabla f_\sigma(x_\sigma^k)\| \leq (C/\sigma^2)\|x_\sigma(t_k) - x_\sigma^k\|$ as ∇f_σ is Lipschitz on bounded sets with a Lipschitz constant of $O(1/\sigma^2)$. Assuming that $\nabla\psi^* f_1^*$, $\nabla\psi^* f_2^* > 0$ so that we have $\|x_\sigma(t_k) - x_\sigma^k\| = O(h)$, and furthermore, $\|h[\nabla f_\sigma(x_\sigma(t_k)) - \nabla f_\sigma(x_\sigma^k)]\lambda_\sigma^{k+1} + \eta_k\| = O(h^2/\sigma^2)$ if $x_\sigma(t)$ is in the transition region for some $t_k \leq t \leq t_{k+1}$ or x_σ^k is in the transition region. Otherwise the more usual bounds $\|h[\nabla f_\sigma(x_\sigma(t_k)) - \nabla f_\sigma(x_\sigma^k)]\lambda_\sigma^{k+1} + \eta_k\| = O(h^2)$ hold. Again, assuming that the trajectories $x_\sigma(t)$ or x_σ^k are in the transition region for only $O(\sigma/h)$ many steps, we can apply a Gronwall lemma to obtain a bound

$$\|\lambda_\sigma(t_k) - \lambda_\sigma^k\| = O(h^2/\sigma^2)\, O(\sigma/h) + O(h^2)\, O(1/h) = O(h/\sigma) \tag{5.19}$$

with the implicit constants independent of k, h and σ. Thus if $h = o(\sigma)$, and h, $\sigma \to 0$, the adjoint variables converge to the gradients of $g(x(t_f))$ with respect to the initial conditions for the discontinuous limit.

5.3.3. Convergence of the numerical sequence s_σ^i. The rule of the numerical integration of the sensitivity and of the adjoint variables is

$$\begin{aligned} s_\sigma^{i+1} &= \left(I + h\nabla f_\sigma(x_\sigma^i)\right) s_\sigma^i \\ \lambda_\sigma^i &= \left(I + h\nabla f_\sigma(x_\sigma^i)\right)^T \lambda_\sigma^{i+1} \end{aligned}$$

Multiplying the first equation by λ_σ^{i+1} and the second equation by s_σ^{i+1} we obtain that

$$(s_\sigma^{i+1})^T \lambda_\sigma^{i+1} = (s_\sigma^i)^T \lambda_\sigma^i.$$

Using now the fact that the sensitivity and the adjoint of the exact solution satisfies

$$s(t)^T \lambda(t) = \text{constant},$$

for any change in the initial condition x_0, as well as the fact that

$$\lambda_\sigma^i \xrightarrow{\sigma \to 0} \lambda$$

that was proved above, it follows that

$$s_\sigma^i \xrightarrow{\sigma \to 0} s,$$

which completes the proof of the Theorem 5.2.

6. Convergence of gradients for time-discretizations of the case where $\nabla\psi(x) f_1(x) > 0$ and $\nabla\psi(x) f_2(x) < 0$. In the case treated here, the solution of the smoothed equation (4.1) will stay in the transition region for $\omega(\sigma)$ time. In this case there is also a jump in the limit of the (smoothed) adjoint variables, and there is also a "jump formula" for the limiting adjoint equations. Furthermore, the adjoint variables obtained in the limit are the correct adjoints for the discontinuous system.

Within the transition region, the adjoint variables change most rapidly in directions near to $\nabla\psi^*$. By Lemma 3.1, $f_2^* - f_1^* = -\rho^* \nabla\psi^*$ for some $\rho^* \geq 0$.

If the trajectory stays on the discontinuity for any open interval, then while the trajectory is on the interval, an equivalent right-hand side can be used for the motion on the discontinuity [27].

6.1. The convergence result. In this section, we prove the following result concerning the convergence of the sensitivity of the equation (4.1) to the one of (3.1).

THEOREM 6.1. *Assume that* $\nabla\psi(x)f_1(x) > 0$, $\nabla\psi(x)f_2(x) < 0$, *whenever* $\psi(x) = 0$. *Assume that we integrate the smoothed model equation (4.1) and the corresponding sensitivity equation (5.1) for* $\partial x/\partial x_0$ *using Euler's method with a time step* $h = o(\sigma^2)$. *Then, the numerical sensitivites and the numerical adjoints converge to the sensivities of the original problem as* $\sigma \to 0$.

Note that we will need to take $h = o(\sigma^2)$ in order to resolve the trajectory as it passes through the transition region so that the gradient information will be accurate. It will turn out that this is sufficient for the Euler method to compute approximate gradients that converge to the true gradient as computed in the previous section.

Our initial investigations suggest that the result is true even for the case where $h = o(\sigma)$. Nonetheless, this would require additional complexity to an already very technical proof so we will use the stronger assumption. However, wherever possible in the course of the proof, we will invoke only the weaker assumption $h = o(\sigma)$.

The proof of this result is split into a number of items which appear as subsections in the remainder of this section. We note that the proof of the convergence of the state vectors is identical to the one for the preceding case, from Subsection 5.3.1. In addition, we will use results from Sections 3 and 4 that apply to this case.

6.1.1. Asymptotic behavior of $\varphi_\sigma(\psi(x_\sigma(t)))$. Consider the differential equation

$$\begin{aligned}\frac{d}{dt}\varphi_\sigma(\psi(x_\sigma))) &= \varphi'_\sigma(\psi(x_\sigma))\nabla\psi(x_\sigma)^T[(1-\varphi_\sigma(\psi(x_\sigma)))f_1(x_\sigma) + \varphi_\sigma(\psi(x_\sigma))f_2(x_\sigma)] \\ &= \varphi'_\sigma(\psi(x_\sigma))[(1-\varphi_\sigma(\psi(x_\sigma)))\gamma_1(x_\sigma(t)) + \varphi_\sigma(\psi(x_\sigma))\gamma_2(x_\sigma)]\end{aligned}$$

where $\gamma_i(x) = \nabla\psi(x)^T f_i(x)$, $i = 1, 2$.

We consider this differential equation for a time interval $[t^*_{\sigma,-}, t^*_{\sigma,-} + \epsilon]$ where $t^*_{\sigma,-}$ is the first time when we reach the transition zone. In an interval of this size, $\gamma_i(x_\sigma(t)) = \gamma_i + O(\epsilon+\sigma)$. We will consider $\epsilon = \omega(\sigma)$, so $\gamma_i(x_\sigma(t)) = \gamma_i + O(\epsilon)$. Setting $w(t) = \gamma_1 - (\gamma_1 - \gamma_2)\varphi_\sigma(\psi(x_\sigma(t)))$, we see that

$$\frac{dw}{dt} = -\mu(t)[w(t) + g(t)] \tag{6.1}$$

where $\mu(t) = (\gamma_1 - \gamma_2)\varphi'_\sigma(\psi(x_\sigma(t)))$ and $g(t) = O(\epsilon/\sigma)$. Note that in a time interval of size $O(\sigma)$ the trajectory will reach a point where $\varphi_\sigma(\psi(x_\sigma(t)))$ is halfway between zero and $\gamma_1/(\gamma_1 - \gamma_2)$. Call this time $t^*_{\sigma,-1/2}$. We can bound $\varphi'(\psi(x_\sigma(t)))$ away from zero, at least for a time interval of length bounded away from zero.

On our time interval of length ϵ, then, $\mu(t) = \Theta(1/\sigma)$ and $\gamma_i(x_\sigma(t)) = \gamma_i + O(\epsilon)$. We can solve the differential equation for w starting from $t^*_{\sigma,-1/2}$:

$$\begin{aligned}w(t^*_{\sigma,-1/2} + t) = \exp\left(-\int_0^t \mu(t^*_{\sigma,-1/2} + \tau)\,d\tau\right) w(t^*_{\sigma,-1/2}) \\ + \int_0^t \exp\left(-\int_\tau^t \mu(t^*_{\sigma,-1/2} + s)\,ds\right) g(t^*_{\sigma,-1/2} + \tau)\,d\tau.\end{aligned}$$

Thus for $0 \le t \le \epsilon$ we get

$$w(t^*_{\sigma,-1/2} + t) = \int_0^t e^{-C(t-\tau)/\sigma}\,O(\epsilon/\sigma)\,d\tau = O(e^{-Ct/\sigma}) + O(\epsilon). \tag{6.2}$$

So in a time $\geq \text{const}\,\sigma \log(1/\epsilon)$ the difference between $\varphi_\sigma(\psi(x_\sigma(t)))$ and $\gamma_1/(\gamma_1 - \gamma_2)$ is $O(\epsilon)$.

Set $\theta(x) = \gamma_1(x)/(\gamma_1(x) - \gamma_2(x))$. More careful analysis shows that for t large compared with $\sigma \log(1/\sigma)$, the difference between $\varphi_\sigma(\psi(x_\sigma(t)))$ and $\theta(x_\sigma(t))$ is $O(\sigma)$. To see this, construct the solution to the differential equation

$$\frac{d}{dt}(w - g) = -\mu(t)(w - g) - g'(t) \tag{6.3}$$

as

$$w(t) - g(t) = \exp\left(-\int_0^t \mu(\tau)\, d\tau\right)(w(0) - g(0)) - \int_0^t \exp\left(-\int_\tau^t \mu(s)\, ds\right) g'(\tau)\, d\tau$$

and substitute $w(t) = \varphi_\sigma(\psi(x_\sigma(t^*_{\sigma,-} + t)))$ and $g(t) = \theta(x_\sigma(t^*_{\sigma,-} + t))$.

6.1.2. Asymptotic behavior of $\varphi_\sigma(\psi(x^k_\sigma))$. Note that if $h = O(\epsilon\sigma^2)$, then we can use the global error bound in (5.15). Since $x \mapsto \varphi_\sigma(\psi(x))$ is Lipschitz with constant of $O(1/\sigma)$, for $t_k \geq t^*_{\sigma,-}$ we have $(\gamma_1 - \gamma_2)\varphi'_\sigma(\psi(x^k_\sigma)) = O(\epsilon)$. In addition, we have that $\varphi_\sigma(\psi(x^k_\sigma)) \to \theta(x(t))$ for t above the switching point.

This can be improved to merely requiring $h/\sigma = O(\epsilon)$. However, a detailed rigorous demonstration of this would require improved error bounds that take into account the exponential damping of order $O\left(1/\sigma\right)$ in the direction perpendicular to the manifold $\Sigma = \{\, x \mid \psi(x) = 0 \,\}$. Since this would result in a substantial additional complexity to what is already very technical proof, and we will not follow it in the context of this paper.

6.1.3. Sensitivity equations in the transition region, and their discretization. If we apply Euler's method (with $h = o(\sigma)$) to the sensitivity equations in the transition region, we get

$$\begin{aligned} s^{k+1}_\sigma &= [I + h\,\nabla f_\sigma(x^k_\sigma)]s^k_\sigma \\ &= \big[I + h\varphi'_\sigma(\psi(x^k_\sigma))(f_2(x^k_\sigma)) - f_1(x^k_\sigma))\nabla\psi(x^k_\sigma) \\ &\qquad + h\;(1 - \varphi^k_\sigma)\nabla f_1(x^k_\sigma) + h\varphi^k_\sigma \nabla f_2(x^k_\sigma)\big]\, s^k_\sigma. \end{aligned}$$

where $\varphi^k_\sigma = \varphi(s^k_\sigma)$. Note that away from the boundaries of the transition zone, $\varphi'_\sigma(x) = \Theta(1/\sigma)$. Let $u^k_\sigma = f_1(x^k_\sigma) - f_2(x^k_\sigma)$, $v^k_\sigma = \nabla\psi(x^k_\sigma)^T$, and $F^k_\sigma = (1 - \varphi^k_\sigma)\nabla f_1(x^k_\sigma) + \varphi^k_\sigma \nabla f_2(x^k_\sigma)$. Then we can write the discrete sensitivity equation as

$$s^{k+1}_\sigma = [I - h\varphi'_\sigma(\psi(x^k_\sigma))u^k_\sigma(v^k_\sigma)^T + hF^k_\sigma]s^k_\sigma. \tag{6.4}$$

If x^k_σ was on $\Sigma = \{\, x \mid \psi(x) = 0 \,\}$, then since f satisfies a one-sided Lipschitz condition, $u^k_\sigma \parallel v^k_\sigma$ and $(v^k_\sigma)^T u^k_\sigma \geq 0$. However, while in the transition zone, the distance of x^k_σ from Σ is $O(\sigma)$. Thus the angle between u^k_σ and v^k_σ is $O(\sigma)$ by Lipschitz continuity of f_1, f_2 and $\nabla\psi$, and the assumption that $f_2 - f_1 \neq 0$ and $\nabla\psi \neq 0$ anywhere on Σ.

Let $\alpha^k_\sigma = \varphi'_\sigma(\psi(x^k_\sigma))/((v^k_\sigma)^T u^k_\sigma)$.

For the remainder of this subsection we will drop the σ subscripts.

Then we can write

$$s^{k+1} = \left[I - h\alpha_k \frac{u^k(v^k)^T}{(v^k)^T u^k} + hF^k\right] s^k. \tag{6.5}$$

Note that $\alpha_k = \Theta(1/\sigma)$ and $F^k = O(1)$. Since the angle between u^k and v^k is $O(\sigma)$ and $\alpha_k = O(1/\sigma)$,

$$\alpha_k \frac{u^k (v^k)^T}{(v^k)^T u^k} = \alpha_k \frac{u^k (u^k)^T}{(u^k)^T u^k} + O(1). \tag{6.6}$$

Thus

$$s^{k+1} = \left[I - h\alpha_k \frac{u^k (u^k)^T}{(u^k)^T u^k} + h\widehat{F}^k \right] s^k, \tag{6.7}$$

where $\widehat{F}^k = O(1)$.

Let $\widehat{u}^k = u^k/\|u^k\|_2$. Choose a family of orthogonal matrices Q_k where $Q_k \widehat{u}^k = \widehat{u}^{k+1}$. Since $\|u^{k+1} - u^k\| = O(h)$, and $\|u^k\|$ is bounded away from zero, we can choose Q_k so that $\|Q_k - I\|_2 = O(h)$. Put $R_k = Q_{k-1} Q_{k-2} \cdots Q_1 Q_0$, with $Q_j = I$ if x_σ^j is not in the transition zone. With this in mind, the discrete sensitivity equations can be re-written as

$$\begin{aligned} s^{k+1} &= \left[I - h\alpha_k R_k \widehat{u}^0 (\widehat{u}^0)^T R_k^T + h\widehat{F}^k \right] s^k, \\ &= R_k [I - h\alpha_k \widehat{u}^0 (\widehat{u}^0)^T + h R_k^T \widehat{F}^k R_k] R_k^T s^k. \end{aligned}$$

For large k, we can write

$$\begin{aligned} s^k &= \left[\prod_{j=0}^{k-1} (R_j [I - h\alpha_j \widehat{u}^0 (\widehat{u}^0)^T + h R_j^T \widehat{F}^j R_j] R_j^T) \right] s^0 \\ &= R_k \left[\prod_{j=0}^{k-1} (R_{j+1}^T R_j [I - h\alpha_j \widehat{u}^0 (\widehat{u}^0)^T + h R_j^T \widehat{F}^j R_j]) \right] R_0^T \, s^0 \\ &= R_k \left[\prod_{j=0}^{k-1} (Q_j [I - h\alpha_j \widehat{u}^0 (\widehat{u}^0)^T + h R_j^T \widehat{F}^j R_j]) \right] R_0^T \, s^0 \end{aligned} \tag{6.8}$$

where $\prod_{j=0}^{k-1} A_j = A_{k-1} A_{k-2} \cdots A_1 A_0$. Noting that $Q_j = I + O(h)$, we can absorb the difference $Q_j - I$ in the product into the $O(1)$ term of the above product. This gives

$$s^k = R_k \left[\prod_{j=0}^{k-1} (I - h\alpha_j \widehat{u}^0 (\widehat{u}^0)^T + h G_j) \right] R_0^T \, s^0 \tag{6.9}$$

with $G_j = O(1)$. Provided $0 \le h\alpha_j \le 1$ for all j, it is easy to show that this product is uniformly bounded as $h \downarrow 0$ with kh bounded.

We want to go further, and show that this converges to a matrix of the form $R(I - \widehat{u}^0(\widehat{u}^0)^T) G (I - \widehat{u}^0(\widehat{u}^0)^T)$ with R orthogonal.

In addition to supposing that $h = o(\sigma)$, we choose p, an integer, so that $\sigma = o(ph)$, and write $\epsilon := ph$.

6.1.4. Lower bounds for $\varphi_\sigma(\psi(x_\sigma(t)))$. In the following, we will make certain assumptions about the properties of the function $\rho(x)$ that defines the smoothing function $\varphi_\sigma(x)$. Recall, we have defined

$$\varphi_\sigma(x) = \int_{-\infty}^{x} \rho\left(\frac{r}{\sigma}\right) \frac{dr}{\sigma} = \int_{-\infty}^{x/\sigma} \rho(r')dr' = \int_{-1}^{x/\sigma} \rho(r')dr'$$

Namely, we will assume that there exists a positive parameter c such that

$$\begin{aligned}&\int_{-1}^{x} \rho(r')dr' \le c\rho(x),\ x \in [-1,0];\\ &\int_{x}^{1} \rho(r')dr' \le c\rho(x),\ x \in [0,1].\end{aligned}$$

An immediate consequence of this assumption is that

$$\min(\varphi_\sigma(x), (1-\varphi_\sigma(x)) \le c\rho\left(\frac{x}{\sigma}\right),\ x \in [-\sigma,\sigma].$$

Since we aimed to prove certain properties of the solution while in the transition region and while the solution follows the discontinuity manifold, it is important to define the transition region for the situation where $\sigma \neq 0$. In our case, we will simply define it as the point $x_\sigma(t)$ that satisfies $\rho_\sigma(\psi(x_\sigma(t))) \ge k_1$ where $k_1 > 0$ is sufficiently small. In addition, since we have shown that $\rho_\sigma(\psi(x_\sigma(t))) \overset{\sigma\to 0}{\longrightarrow} \theta(x(t))$, and since our assumption that $\nabla\psi(x)^T f_1(x) > 0$ and $\nabla\psi(x)^T f_2(x) < 0$ implies that $\theta(x(t))$ must be bounded away from 1, it follows that in the transition region and in the region that follows the discontinuity manifold, we will have the following inequality

$$k_1 \le \varphi_\sigma(\psi(x_\sigma(t))) \le 1 - k_2$$

In turn, this implies that whenever $x_\sigma(t)$ is in the transition region or follows the discontinuity, we will have that

$$\min(k_1,k_2) \le \min(\varphi_\sigma(\psi(x_\sigma(t))), (1-\varphi_\sigma(\psi(x_\sigma(t)))) \le c\rho\left(\frac{\psi(x_\sigma(t))}{\sigma}\right),$$

Therefore, in the same régime we will have that

$$\varphi_\sigma'(\psi(x_\sigma(t)) = \frac{1}{\sigma}\rho\left(\frac{\psi(x_\sigma(t)}{\sigma}\right) \ge \frac{\min(k_1,k_2)}{c\sigma} \tag{6.10}$$

As a result, we have that

$$h\sum_{t_k\in I_\sigma} \varphi_\sigma'(\psi(x_\sigma(t_k)) \to \infty$$

as soon as $m(I_\sigma) \ge \sigma^p,\ p \in (0,1)$.

6.1.5. Results on products of nearby projections. Consider the product with $\|\widehat{u}_i\|_2 = 1$ for all i and $\widehat{u}_{i+1} - \widehat{u}_i = O(h)$:

$$P := \prod_{i=1}^{p} (I - h\alpha_i \widehat{u}_i \widehat{u}_i^T).$$

Let $Q_{i,i+1}\widehat{u}_i = \widehat{u}_{i+1}$ by an orthogonal matrix: $Q_{i,i+1} = I + (\widehat{u}_{i+1} - \widehat{u}_i)\widehat{u}_i^T - \widehat{u}_i(\widehat{u}_{i+1} - \widehat{u}_i)^T +$h.o.t. Then put $Q_{r,s} = Q_{r,r+1}Q_{r+1,r+2}\cdots Q_{s-1,s}$ for $s > r$. Note that $Q_{r,s}\widehat{u}_r = \widehat{u}_s$. Also $Q_{r,s}^T\widehat{u}_s = \widehat{u}_r$ so we put $Q_{r,s}^T = Q_{s,r}$. Then

$$\begin{aligned}P &= \prod_{i=1}^{p} (I - h\alpha_i \widehat{u}_i \widehat{u}_i^T)\\ &= \prod_{i=1}^{p} (I - h\alpha_i Q_{0,i}\widehat{u}_0 \widehat{u}_0^T Q_{0,i}^T)\end{aligned}$$

$$
\begin{aligned}
&= \prod_{i=1}^{p} (Q_{0,i}(I - h\alpha_i \widehat{u}_0 \widehat{u}_0^T) Q_{0,i}^T) \\
&= Q_{0,j+1} \left[\prod_{i=1}^{p} (Q_{0,i+1}^T Q_{0,i}(I - h\alpha_i \widehat{u}_0 \widehat{u}_0^T)) \right] Q_{0,1}^T
\end{aligned}
$$

Note that $Q_{0,i+1} = Q_{i,i+1} Q_{0,i}$, so

$$
\begin{aligned}
Q_{0,i+1}^T Q_{0,i} &= Q_{0,i}^T Q_{i,i+1}^T Q_{0,i} \\
&= Q_{0,i}^T \left[I + \widehat{u}_i (\widehat{u}_{i+1} - \widehat{u}_i)^T - (\widehat{u}_{i+1} - \widehat{u}_i) \widehat{u}_i^T + \text{h.o.t.} \right] Q_{0,i} \\
&= I + Q_{0,i}^T \widehat{u}_i (\widehat{u}_{i+1} - \widehat{u}_i)^T Q_{0,i} - Q_{0,i}^T (\widehat{u}_{i+1} - \widehat{u}_i) \widehat{u}_i^T Q_{0,i} + \text{h.o.t.} \\
&= I + \widehat{u}_0 z_i^T - z_i \widehat{u}_0^T + \text{h.o.t.}
\end{aligned}
$$

Note that "h.o.t." means "higher order terms". Note that $z_i = O(h)$ and h.o.t. $= O(h^2)$. If $\widetilde{G}_i := Q_{0,i+1}^T Q_{0,i} - I = \widehat{u}_0 z_i^T - z_i \widehat{u}_0^T + \text{h.o.t.}$ then

$$
P = Q_{0,j+1} \left[\prod_{i=1}^{p} ((I + \widetilde{G}_i)(I - h\alpha_i \widehat{u}_0 \widehat{u}_0^T)) \right] Q_{0,1}^T.
$$

Expanding the factors with $I + \widetilde{G}_i$ we get

$$
\begin{aligned}
P = {} & Q_{0,j+1} \left[\prod_{i=1}^{p} (I - h\alpha_i \widehat{u}_0 \widehat{u}_0^T) \right] Q_{0,1}^T \\
& + Q_{0,j+1} \left[\sum_{i=1}^{p} \left\{ \prod_{j=i+1}^{p} (I - h\alpha_i \widehat{u}_0 \widehat{u}_0^T) \right\} \widetilde{G}_i \left\{ \prod_{j=1}^{i-1} (I - h\alpha_i \widehat{u}_0 \widehat{u}_0^T) \right\} \right] Q_{0,1}^T + O(h^2 p^2).
\end{aligned}
$$

Assume that all $\alpha_i = \Theta(1/\sigma)$ with $\sigma > 0$ small (bounds independent of i) and $h = o(\sigma)$. This assumption is satisfied in our case, once we will specify α_i, by using (6.10). Then

$$
\begin{aligned}
\prod_{j=r}^{s} (I - h\alpha_i \widehat{u}_0 \widehat{u}_0^T) &= I - \left[1 - \prod_{j=r}^{s} (1 - h\alpha_j) \right] \widehat{u}_0 \widehat{u}_0^T \\
&= P_0 + O(\prod_{j=r}^{s} (1 - h\alpha_j)), \qquad P_0 = I - \widehat{u}_0 \widehat{u}_0^T.
\end{aligned}
$$

Thus

$$
\begin{aligned}
\sum_{i=1}^{p} \left\{ \prod_{j=i+1}^{p} (I - h\alpha_i \widehat{u}_0 \widehat{u}_0^T) \right\} \widetilde{G}_i \left\{ \prod_{j=1}^{i-1} (I - h\alpha_i \widehat{u}_0 \widehat{u}_0^T) \right\} = \\
\sum_{i=1}^{p} P_0 \widetilde{G}_i P_0 + O(h\,(\sigma/h) \log(\sigma/h)).
\end{aligned}
$$

But $P_0 \widetilde{G}_i P_0 = (I - \widehat{u}_0 \widehat{u}_0^T)(\widehat{u}_0 z_i^T - z_i \widehat{u}_0^T + O(h^2))(I - \widehat{u}_0 \widehat{u}_0^T) = O(h^2)$, so

$$
\sum_{i=1}^{p} \left\{ \prod_{j=i+1}^{p} (I - h\alpha_i \widehat{u}_0 \widehat{u}_0^T) \right\} \widetilde{G}_i \left\{ \prod_{j=1}^{i-1} (I - h\alpha_i \widehat{u}_0 \widehat{u}_0^T) \right\} = O(\sigma \log(\sigma) + p h^2).
$$

Thus $P = Q_{0,j+1} P_0 Q_{0,1}^T + O(\sigma + ph^2 + p^2h^2)$. In fact, noting that $Q_{0,1} = I + O(h)$, we get

$$P = Q_{0,j+1}(I - \widehat{u}_0 \widehat{u}_0^T) + O(\sigma + ph^2 + p^2h^2).$$

Taking $ph \to \epsilon$ we get

$$P = I - \widehat{u}_0 \widehat{u}_0^T + O(\epsilon). \tag{6.11}$$

6.1.6. Jump conditions for the sensitivity. We wish to use the above result to show that if $G_i = O(1)$ for all i, then

$$\prod_{i=1}^{p} (I - h\alpha_i \widehat{u}_i \widehat{u}_i^T + hG_i) = I - \widehat{u}_0 \widehat{u}_0^T + O(\epsilon) \tag{6.12}$$

provided $ph = O(\epsilon)$.

Now

$$\begin{aligned}
\prod_{i=1}^{p} (I - h\alpha_i \widehat{u}_i \widehat{u}_i^T + hG_i) &= \prod_{i=1}^{p} (I - h\alpha_i \widehat{u}_i \widehat{u}_i^T) \\
&\quad + h \sum_{i=1}^{p} \left\{ \prod_{j=i+1}^{p} (I - h\alpha_i \widehat{u}_i \widehat{u}_i^T) \right\} G_i \left\{ \prod_{j=1}^{i-1} (I - h\alpha_i \widehat{u}_i \widehat{u}_i^T) \right\} \\
\text{(from (6.11))} \qquad &= I - \widehat{u}_0 \widehat{u}_0^T + O(\epsilon) \\
&\quad + h \sum_{i=1}^{p} (I - \widehat{u}_0 \widehat{u}_0^T + O(\epsilon)) G_i (I - \widehat{u}_0 \widehat{u}_0^T + O(\epsilon)) \\
&= I - \widehat{u}_0 \widehat{u}_0^T + O(\epsilon).
\end{aligned}$$

Taking limits as $h, \sigma \to 0$ with $h = o(\sigma^2)$ gives $s(t^* + \epsilon) = (I - \widehat{u}_0 \widehat{u}_0^T) s(t^* - \epsilon) + O(\epsilon)$. That is, $s(t^{*+}) = (I - \widehat{u}_0 \widehat{u}_0^T) s(t^{*-})$, which is the required jump rule for the sensitivities (3.5).

6.1.7. Convergence to the differential equation for the sensitivity on the discontinuity. We assume that $x(t^*)$ (the exact solution) lies on Σ. Then $s(t^{*+}) \perp \nabla \psi(x(t^*))$ by Section 6.1.6.

Consider a time interval $[t^*, t^* + \epsilon]$ with $0 < \epsilon \ll 1$ and we take $ph \to \epsilon$, $h = o(\sigma)$, $\sigma = O(\epsilon^2)$. Then the limit of the computed sensitivities can be computed from

$$s^{k^*+p} = \prod_{i=1}^{p} \left(I - h\alpha_i \frac{u^i (v^i)^T}{(v^i)^T u^i} + h\, F_i \right) s^{k^*}$$

where $F_i = (1 - \varphi_\sigma(\psi(x^{k^*+i})) \nabla f_1(x^{k^*+i}) + \varphi_\sigma(\psi(x^{k^*+i})) \nabla f_2(x^{k^*+i})$, etc. Note that $\varphi_\sigma(\psi(x^{k^*+i})) - \theta(x^{k^*+i}) = O(\sigma)$ with a constant independent of i, $1 \le i \le p$. From the above computations,

$$\begin{aligned}
s^{k^*+p} &= Q_{0,p}(I - \widehat{u}^0 (\widehat{u}^0)^T)\, s^{k^*} \\
&\quad + h \sum_{i=1}^{p} \left\{ \prod_{j=i+1}^{p} (I - h\alpha_{j+k^*} \widehat{u}^j (\widehat{u}^j)^T) \right\} F_i \left\{ \prod_{j=1}^{i} (I - h\alpha_{j+k^*} \widehat{u}^j (\widehat{u}^j)^T) \right\} s^{k^*} \\
&\quad + O(h^2p^2 + \sigma)
\end{aligned}$$

Since $F_i - [(1-\theta(x^{k^*+i})\nabla f_1(x^{k^*+i}) + \theta(x^{k^*+i})\nabla f_2(x^{k^*+i})] = O(\sigma)$, and the trajectories converge, we can take the limit as $h \to 0$, $ph \to \epsilon$. This gives the solution of the smoothed problem. We can also (simultaneously) take $\sigma \to 0$ with $h = o(\sigma)$ to get the limit of the computations with Euler's method. Without taking the limit as $\sigma \to 0$, we get

$$\begin{aligned}
s^{k^*+p} &= Q_{0,p}(I - \widehat{u}^0(\widehat{u}^0)^T)\, s^{k^*} \\
&\qquad + h\sum_{i=1}^{p} Q_{i+1,p}(I - \widehat{u}^{i+1}(\widehat{u}^{i+1})^T) F_i Q_{1,i-1}(I - \widehat{u}^1(\widehat{u}^1)^T)\, s^{k^*} \\
&\qquad + O(h^2p^2 + \sigma) \\
&= Q_{0,p}(I - \widehat{u}^0(\widehat{u}^0)^T)\, s^{k^*} \\
&\qquad + h\sum_{i=1}^{p} Q_{i+1,p}(I - \widehat{u}^{i+1}(\widehat{u}^{i+1})^T) F_i (I - \widehat{u}^{i-1}(\widehat{u}^{i-1})^T) Q_{1,i-1}\, s^{k^*} \\
&\qquad + O(h^2p^2 + \sigma).
\end{aligned}$$

Now

$$\begin{aligned}
Q_{0,p} - I &= \prod_{i=1}^{p}(I + (\widehat{u}^{i+1} - \widehat{u}^i)(\widehat{u}^i)^T - \widehat{u}^i(\widehat{u}^{i+1} - \widehat{u}^i)^T) - I + O(ph^2) \\
&= \sum_{i=1}^{p}\left[(\widehat{u}^{i+1} - \widehat{u}^i)(\widehat{u}^i)^T - \widehat{u}^i(\widehat{u}^{i+1} - \widehat{u}^i)^T\right] + O(p^2h^2) \\
&= \sum_{i=1}^{p}\left[(\widehat{u}^{i+1} - \widehat{u}^i)(\widehat{u}^0)^T - \widehat{u}^0(\widehat{u}^{i+1} - \widehat{u}^i)^T\right] + O(p^2h^2) \\
&= (\widehat{u}^{p+1} - \widehat{u}^0)(\widehat{u}^0)^T - \widehat{u}^0(\widehat{u}^{p+1} - \widehat{u}^0)^T + O(p^2h^2).
\end{aligned}$$

Assuming $(s^{k*})^T\nabla\psi(x(t^*)) = O(\sigma)$, we get

$$\begin{aligned}
\frac{s^{k^*+p} - s^{k^*}}{ph} &= [(\widehat{u}^{p+1} - \widehat{u}^0)(\widehat{u}^0)^T - \widehat{u}^0(\widehat{u}^{p+1} - \widehat{u}^0)^T] s^{k^*}/(ph) \\
&\qquad + \frac{1}{p}\sum_{i=1}^{p} Q_{i+1,p}(I - \widehat{u}^{i+1}(\widehat{u}^{i+1})^T) F_i (I - \widehat{u}^{i-1}(\widehat{u}^{i-1})^T) Q_{1,i-1}\, s^{k^*} \\
&\qquad + O(hp + \sigma/(hp)) \\
&= -\widehat{u}^0(\widehat{u}^{p+1} - \widehat{u}^0)^T s^{k^*}/(ph) \\
&\qquad + \frac{1}{p}\sum_{i=1}^{p} Q_{i+1,p}(I - \widehat{u}^{i+1}(\widehat{u}^{i+1})^T) F_i (I - \widehat{u}^{i-1}(\widehat{u}^{i-1})^T) Q_{1,i-1}\, s^{k^*} \\
&\qquad + O(hp + \sigma/(hp)).
\end{aligned}$$

Now $\widehat{u}^{p+1} = u^{p+1}/\|u^{p+1}\|$ and $u^{p+1} = f_1(x^{k^*+p+1}) - f_2(x^{k^*+p+1})$. Since f_1 and f_2 are C^1 and $f_1 - f_2$ is non-zero on Σ, so $x \mapsto (f_1(x) - f_2(x))/\|f_1(x) - f_2(x)\|$ is a smooth map in a neighborhood of Σ. Thus $(\widehat{u}^{p+1} - \widehat{u}^0)/(ph) \to \nabla[(f_1 - f_2)/\|f_1 - f_2\|](x(t^*))\, f^*(x(t^*)) =: z$ as $ph \to 0$. Since $F_i = (1-\theta(x))\nabla f_1(x(t^*)) + \theta\,\nabla f_2(x(t^*)) + O(ph)$ and $Q_{i+1,p}$, $Q_{1,i-1} = I + O(ph)$, it follows that if $P(t) = I - \widehat{u}(t)\widehat{u}(t)^T$ where $\widehat{u}(t) = u(t)/\|u(t)\|$ and $u(t) = f_1(x(t)) - f_2(x(t))$,

$$\frac{s^{k^*+p} - s^{k^*}}{ph} = \left[-\widehat{u}^0 z^T + P(t^*)\left\{(1-\theta(x))\nabla f_1(x(t^*)) + \theta\,\nabla f_2(x(t^*))\right\} P(t^*)\right] s^{k^*}$$

$$+ O(hp + \sigma/(hp)).$$

Noting that $\sigma = o(\epsilon)$ and taking $ph = \epsilon \to 0$ gives

$$\begin{aligned}\frac{ds}{dt}(t^*) = &\left[-\frac{f_1(x(t^*) - f_2(x(t^*))}{\|f_1(x(t^*) - f_2(x(t^*))\|} z(t^*)^T\right] s(t^*) \\ &+ [P(t^*)\{(1-\theta(x(t^*)))\nabla f_1(x(t^*)) + \theta(x(t^*))\,\nabla f_2(x(t^*))\}\, P(t^*)]\, s(t^*).\end{aligned}$$

with $z(t) = \nabla[(f_1 - f_2)/\|f_1 - f_2\|](x(t))\, f^*(x(t))$. But $s(t^*) \perp u(t^*)$, so

$$\begin{aligned}\frac{ds}{dt}(t^*) = &\left[-\frac{f_1(x(t^*) - f_2(x(t^*))}{\|f_1(x(t^*) - f_2(x(t^*))\|} z(t^*)^T\right] s(t^*) \\ &+ [P(t^*)\{(1-\theta(x(t^*)))\nabla f_1(x(t^*)) + \theta(x(t^*))\,\nabla f_2(x(t^*))\}]\, s(t^*).\end{aligned}$$

Note that $\nabla f^*(x) = (1-\theta(x))\nabla f_1(x) + \theta(x)\nabla f_2(x) + (f_2(x) - f_1(x))\nabla\theta(x)$. Since the equation on the discontinuity is

$$\frac{dx}{dt} = f^*(x),$$

the associated variational equations

$$\frac{ds}{dt} = \nabla f^*(x(t))\, s$$

must keep the tangent plane of the disconuity invariant. Note that

$$\begin{aligned}P(t^*)\frac{ds}{dt}(t^*) &= P(t^*)\,[(1-\theta(x))\nabla f_1(x(t^*)) + \theta\,\nabla f_2(x(t^*))]\, s(t^*) \\ &= P(t^*)\nabla f^*(x(t^*))\, s(t^*).\end{aligned}$$

In order to obtain the correct sensitivity in the limit, it suffices to that the component of $ds/dt(t^*)$ in the direction of $u(t^*)$ is correct. But, $(\widehat{u}^{p+1})^T s^{k^*+p} = O(\sigma)$, so in the limit as h, $\sigma \to 0$ with $h = o(\sigma)$, $u(t) \perp s(t)$ for all $t \approx t^*$. Thus the component of $ds/dt(t^*)$ in the direction of $u(t)$ must also be correct, and so

$$\frac{ds}{dt}(t^*) = \nabla f^*(x(t^*))\, s(t^*).$$

Since this is true for all t^* in the interior of the set $\{\, \tau \mid \psi(x(\tau)) = 0 \,\}$, and since s is Lipschitz on this set (provided $u(t)^T s(t)$ for some t in any interval in $\{\, \tau \mid \psi(x(\tau)) = 0 \,\}$), the limit of the numerically computed sensitivities with $h = o(\sigma)$ satisfy the correct sensitivity equation on the discontinuity:

$$\frac{ds}{dt} = \nabla f^*(x(t))\, s.$$

6.1.8. The jump rule for λ. We have so far been able to obtain the jump rule for the sensitivities when the discontinuity manifold is reached:

$$s(t^{*+}) = \left(I - \frac{u(t^*)u(t^*)^T}{u(t^*)^T u(t^*)}\right) s(t^{*-}). \tag{6.13}$$

The corresponding jump rule for λ can be found via the following the following property of the adjoints and sensitivities which is true for any $s(0)$:

$$\frac{d}{dt}(s_\sigma(t)^T \lambda_\sigma(t)) = 0.$$

Thus $s_\sigma(t)^T \lambda_\sigma(t)$ is independent of t. Now $s(t^{*+}) = P(t^*)\, s(t^{*-})$ where $P(t) = I - u(t)u(t)^T/(u(t)^T u(t))$. So taking $\sigma \to 0$ we obtain $s(t^{*+})^T \lambda(t^{*+}) = s(t^{*-})^T \lambda(t^{*-})$. Thus, $s(t^{*-}) P(t^*)^T \lambda(t^{*+}) = s(t^{*-})^T \lambda(t^{*-})$. Since this is true for all values of $s(t^{*-})$, we get the jump rule for λ:

$$\lambda(t^{*-}) = P(t^*)^T \lambda(t^{*+}) = P(t^*) \lambda(t^{*+}), \tag{6.14}$$

since $P(t^*)$ is symmetric.

Similarly, we find that the Lagrange multipliers will satisfy the following adjoint equations on the manifold of discontinuity:

$$\dot\lambda(t) = -\begin{cases} \nabla f_1(x(t,x_0))^T \lambda(t), & t < t_s \\ \nabla f^*(x(t,x_0))^T \lambda(t), & t > t_s \end{cases}, \ \lambda_\sigma(T) = \nabla g(x(T))$$

which satisfies the following jump rule at the discontinuity

$$\lambda(t_s^-) = \left[I + \frac{\left(f^*\left(x(t_s,x_0)\right) - f_1\left(x(t_s,x_0)\right)\right) \nabla\psi(x(t_s,x_0))^T}{\nabla\psi(x(t_s,x_0))^T f_1\left(x(t_s,x_0)\right)} \right] \lambda(t_s^+)$$

7. A model problem and numerical results. We now investigate numerically the benefits of our theoretical results. We use our smoothing approach to investigate an optimal control problem whose discontinuous dynamics originates in the Coulomb friction. While the proofs of our theorems do not include the case with controls, the benefits of our analysis can be extended to that case as well.

Indeed, consider the following problem

$$\begin{aligned} \min_{u,x} \quad & g(x(T)) \\ \text{subject to} \quad & \dot x = f(x,u), \\ & x(0) = x_0, \\ & u(t) \in K. \end{aligned}$$

where K is a given convex set.

We construct the Lagrangian $\mathcal{L}(x,u) = g(x(T)) - \int_0^T \lambda^T \left(\dot x - f(x,u)\right)$. We compute its first-order variations with respect to feasible $\delta x(t)$ and $\delta u(t) \in \mathcal{T}_K(u(t))$.

$$\delta\mathcal{L} = \nabla_x g(x^T)\delta x(T) - \int_0^T \lambda^T \left(\dot{\delta x} - \nabla_x f(x,u)\delta x - \nabla_u f(x,u)\delta u \right).$$

Choosing the adjoint variable $\lambda(t)$ to satisfy

$$\dot\lambda(t) = -(\nabla_x f(x,u))^T \lambda(t), \ \lambda(T) = \nabla_x g(x(T))$$

as well as doing integration by parts and using $\delta x(0) = 0$, we obtain that

$$\delta\mathcal{L} = \int_0^T \lambda(t)^T \nabla_u f(x(t),u(t))^T \delta u(t).$$

for feasible $\delta u(t) \in \mathcal{T}_K(u(t))$.

Therefore, $\lambda(t)\nabla_u f(x(t),u(t))^T$ is the reduced gradient with respect to u. Using Theorems 6.1 and 5.2 we obtain that, if the problem has discontinuities and we use a smoothing approach with an Euler time-stepping scheme with $h = O(\sigma^2)$, the reduced gradients for the smoothed problem approach the ones of the original problem, if $\nabla_u f(x,u)$ is continuous. Therefore, since the gradients converge, the divergence phenomena described at the beginning of this paper will not occur. Nonetheless, when solving our example we do not have to use the reduced gradient computed in this fashion, we have used it only to argue the convergence of the relevant gradients.

We have used the smoothing approach to compute optimal solutions for a crude approximation of a racing car model. This is a version of the "Michael Schumacher" problem described on CPNET, the Complementarity Problem Network [29]. The differential equations and constraints are different here than in [29], in that aerodynamic drag is ignored here, and the track considered here is a more complex S-bend instead of an ellipse.

The state space consists of a vector $\mathbf{x} \in \mathbb{R}^2$ denoting the position of the center of the vehicle, its velocity $\mathbf{v} \in \mathbb{R}^2$, and the angle in which the vehicle is pointing $\theta \in \mathbb{R}$. The controls consist of the throttle $a(t)$ which accelerates or decelerates the vehicle, and the steering control $s(t)$ which changes the vehicle orientation. The auxiliary functions used are $\mathbf{t}(\theta) = (\cos(\theta), \sin(\theta))$, the unit vector the vehicle is pointing in, and $\mathbf{n}(\theta) = (-\sin(\theta), \cos(\theta))$, a unit normal vector to $\mathbf{t}(\theta)$. The differential equations used here are as follows:

$$\dot{\mathbf{x}} = \mathbf{v} \tag{7.1}$$

$$\dot{\mathbf{v}} = a(t)\,\mathbf{t}(\theta) + F\,\mathbf{n}(\theta), \tag{7.2}$$

$$\dot{\theta} = s(t)\,(\mathbf{t}(\theta)^T\mathbf{v}), \tag{7.3}$$

$$F \in -\mu N \mathrm{Sgn}(\mathbf{n}(\theta)^T\mathbf{v}). \tag{7.4}$$

As usual, μ is the coefficient of friction and N is the normal contact force (assumed constant). These equations are clearly not sufficient to realistically describe a Formula 1 racing car. For example, "spin-outs" and "fish-tailing" where large uncontrolled angular velocities occur, cannot happen in this model. However, (7.1)–(7.4) do provide an interesting control system where friction appears in an essential way. Furthermore, we will see that slip is an essential characteristic of the solutions found for certain optimal control problems.

The initial conditions used were $\mathbf{x}(0) = 0$, $\mathbf{v}(0) = 0$ and $\theta(0) = 0$. These are the conditions for a vehicle initially at rest at the origin, pointing horizontally to the right.

The vehicle is constrained not to leave the track. This introduces state constraints of the form $\mathbf{x}(t) \in C$ where $C \subset \mathbb{R}^2$ is the track. For our particular model problem, we take C to be

$$\{\,(x,y) \mid |y - y_{cl}(x)| \le w/2\,\} \tag{7.5}$$

where w is the "width" of the track, and $\{\,(x, y_{cl}(x)) \mid x \in \mathbb{R}\,\}$ is the curve of the centerline of the track. For an interesting but easily implementable system, we set

$$y_{cl}(x) = \begin{cases} \sin(x), & x \le \pi, \\ \pi - x, & \pi \le x \le 2\pi, \\ -\pi - \sin(x), & 2\pi \le x. \end{cases} \tag{7.6}$$

This generates a C^1, but not C^2, curve for the centerline.

σ	N	objective	T	# iter'ns	CPU time
0.1	250	5.544654	5.53393	303	14.1
	500	5.549708	5.53877	531	53.3
	1000	5.552177	5.54111	349	53.3
	2000	5.553451	5.54239	420	227.0
0.05	500	5.590849	5.58285	1501	153.5
	1000	5.409497	5.39842	977	242.2
	2000	5.409183	5.39813	643	300.8
0.025	1000	5.353886	5.34289	1240	184.5
	2000	5.354256	5.34321	1759	913.2
	4000	5.354451	5.34341	1368	1552.8

TABLE 7.1

Objective, final time and algorithm performance for minimum time problem

The controls are subject to simple bounds constraints:

$$|a(t)| \leq a_{\max} \qquad \text{for all } t, \tag{7.7}$$

$$|s(t)| \leq s_{\max} \qquad \text{for all } t. \tag{7.8}$$

The objective function chosen was a combination of a penalty term for missing a target $\mathbf{x}_{tgt}$, plus the time taken to reach the endpoint:

$$g(T, x(T)) = \alpha \|\mathbf{x}(T) - \mathbf{x}_{tgt}\|^2 + T. \tag{7.9}$$

7.1. Specific parameter values. The following default values were used:

- The penalty parameter for the final target was $\alpha = 10$.
- The target point was $\mathbf{x}_{tgt} = (3\pi, -\pi)^T$, which is on the center line of the track.
- The maximum acceleration and steering controls were $a_{\max} = 2$ and $s_{\max} = 2$.
- The maximum friction force was $\mu N = 4$.

7.2. Numerical results. The discretized optimal control problem was set up using the AMPL modeling language [14], and solved using LOQO [35] under Linux. The baseline problem used a time-step of $h = T/N$ with T the final time and N was fixed at 1000, and the smoothing parameter was $\sigma = 0.1$. A number of runs were carried out using different values of N and σ. A summary of the results is shown in Table 7.1.

For the baseline problem, the value of T computed for this "minimal time" problem was $T = 5.541106358$, making $h \approx 5.5 \times 10^{-3}$. The final objective function value was 5.552177023, which was obtained in 349 iterations which took 53.3 seconds of CPU time on a Pentium 4 running Linux. LOQO reported a dual objective function value of 5.552176961. While the objective function and the feasible region are highly non-convex, this does indicate that the objective function value is likely within about 6×10^{-8} of the value of a local minimum. The objective function value indicates that $\alpha\|\mathbf{x}(T) - \mathbf{x}_{tgt}\|^2 \approx 0.01107$; since $\alpha = 10$, this indicates that $\|\mathbf{x}(T) - \mathbf{x}_{tgt}\| \approx 0.03327$; the target is approached to high accuracy. Similar results are apparent for the other values of N and σ (see Table 7.1).

It should be noted that LOQO, like most software for mathematical programming, can only guarantee that the computed solution is close to a local minimum; guaranteeing a global minimum without convexity is a computationally challenging

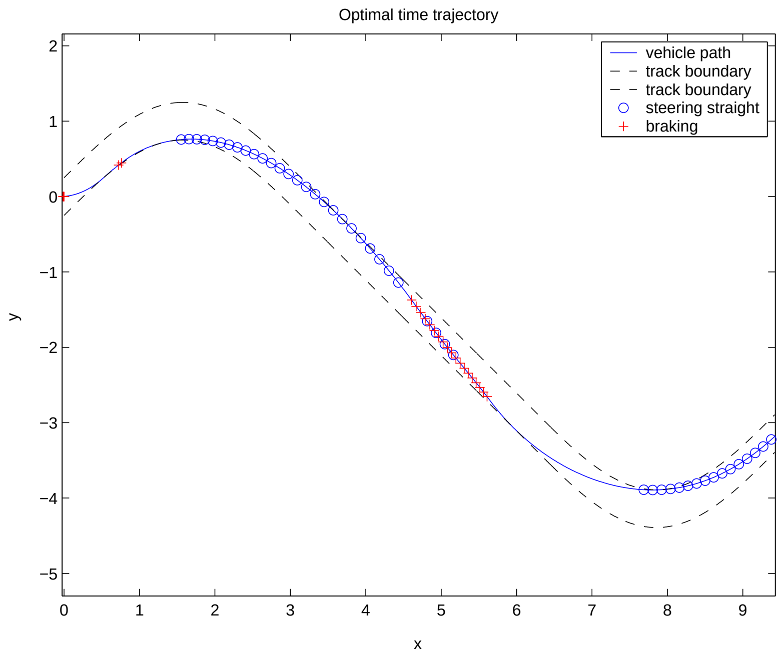

FIG. 7.1. *Trajectory of "race-car"*

task and there is no reason to expect that these optimal control problems do not have many local minima. However, the objective function values reported are remarkably consistent, and the controls used to achieve these local minima are also remarkably similar. This gives us some confidence that the computed objective values and controls are close to a global minimum for the true unsmoothed continuous time problem (7.1)– (7.6).

The computed optimal trajectory for the baseline problem is shown in Figure 7.1. A close-up of the trajectory as it is going around the first half of the first bend is shown in Figure 7.2.

The computed optimal control functions, along with the normal and tangential velocities, are shown in Figure 7.3. A complex maneuver takes place near $t = 3$; this is shown in more detail in Figure 7.4.

Note from Figure 7.1 that the vehicle appears to be braking near the origin. This is because it is at first reversing with the steering set to turn the vehicle clockwise; after a short time the vehicle is accelerating forwards. This can be seen in the control functions in Figure 7.3. When the vehicle comes to rest, the steering wheel is set to the opposite direction to turn the car counter-clockwise in order to stay on the track. After touching the boundary of the track, the vehicle starts turning clockwise in order to turn around the first bend. Since the aim is to minimize the time taken, there is a tendency to keep the tangential velocity as high as possible. Because slip then limits the curvature of the turn, the trajectory can no longer stay in contact with the track boundary (see Figure 7.2).

There is a brief period of deceleration as the vehicle goes around the first bend; this is probably to prevent the vehicle from going outside the track when it touches the north boundary of the track. About when the vehicle reaches the northern-most

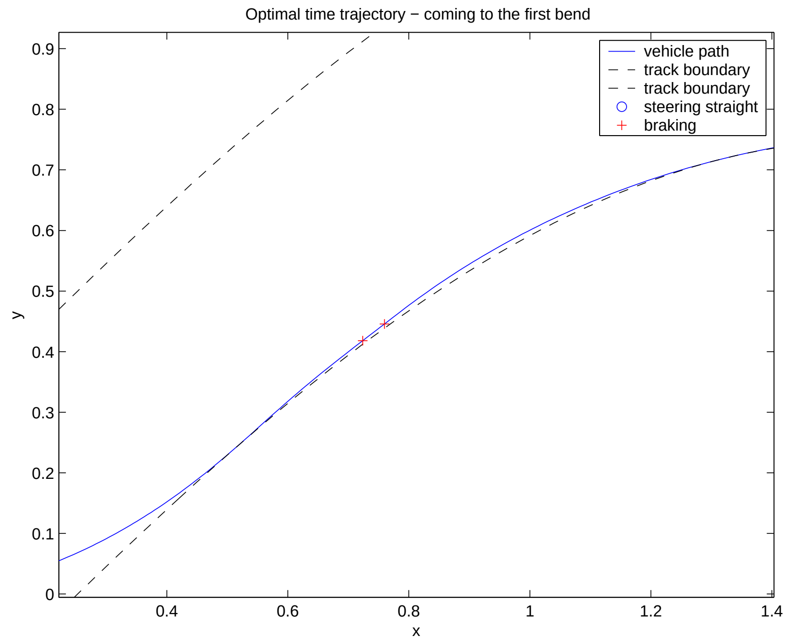

FIG. 7.2. *Trajectory of "race-car" (close-up)*

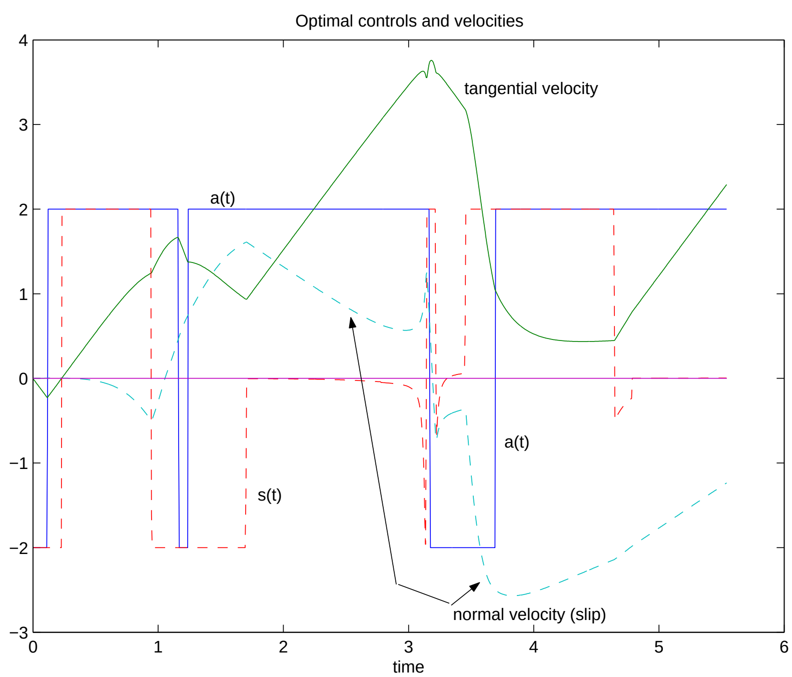

FIG. 7.3. *Computed optimal control functions and velocities*

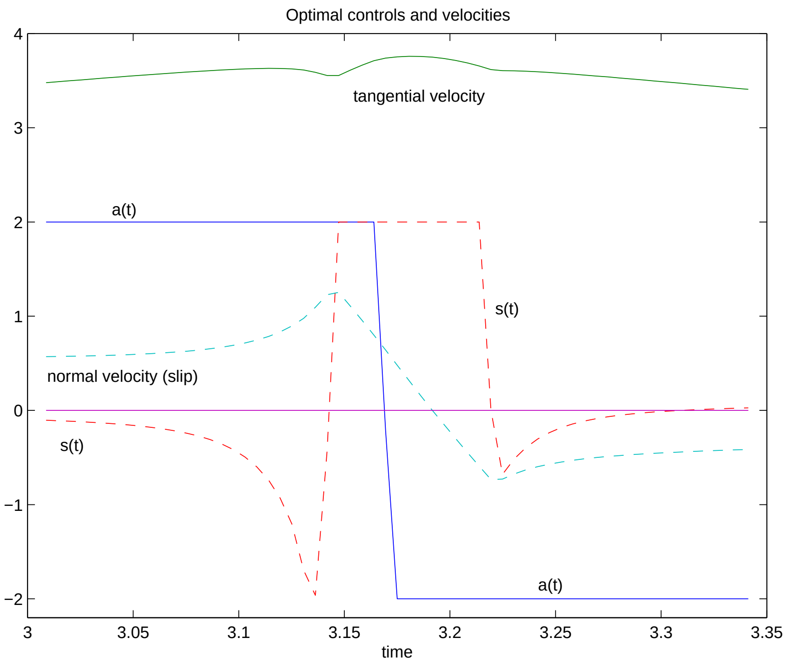

FIG. 7.4. *Computed optimal control functions and velocities (zoom)*

point on the bend the vehicle is already pointing in a direction more than $\pi/4$ radians from east-west (the x-axis). This can be seen in Figure 7.3 near $t = 1.7$ where the ratio of the normal velocity to the tangential velocity is over 1.5. The steering control is set to be straight ($s(t) \approx 0$) until well into the straight segment of the track.

Around $t = 3.2$ there is a complex maneuver which is shown in more detail in Figure 7.4. This occurs at the start of the "braking" part of the trajectory in the straight segment of track, which can be seen in Figure 7.1. First, the steering control reverses direct, which changes the direction of the slip velocity. Then the tangential acceleration control $a(t)$ reverses sign and starts to decelerate the vehicle. Thereafter the steering control returns to near zero. There is some "overshoot" in the steering control. This appears not to be a numerical artefact, as will be discussed below.

Shortly after the above maneuver at about $t = 3.5$, the steering control is set to its maximum value to avoid crossing the southern boundary of the track and to turn around the second bend. The continued deceleration helps the vehicle to change its direction of motion. As the vehicle approaches the second bend, the steering control is set at its maximum value for turning counter-clockwise. This is maintained until the friction forces due to slip is sufficient to make the vehicle reach the final destination without crossing the track boundary. Around $t = 4.7$, the steering control is brought to zero; again there is some overshoot, but it is not clear why this is. As might be expected of an optimal control, it just touches the track boundary as it goes around the second bend.

7.3. Comparison with solutions for different N and σ. Different values of N and σ do result in some differences in the control functions and the velocities. This can be seen in Table 7.1, and also in Figure 7.5. Most of the control and velocities are indistinguishable except for a few features.

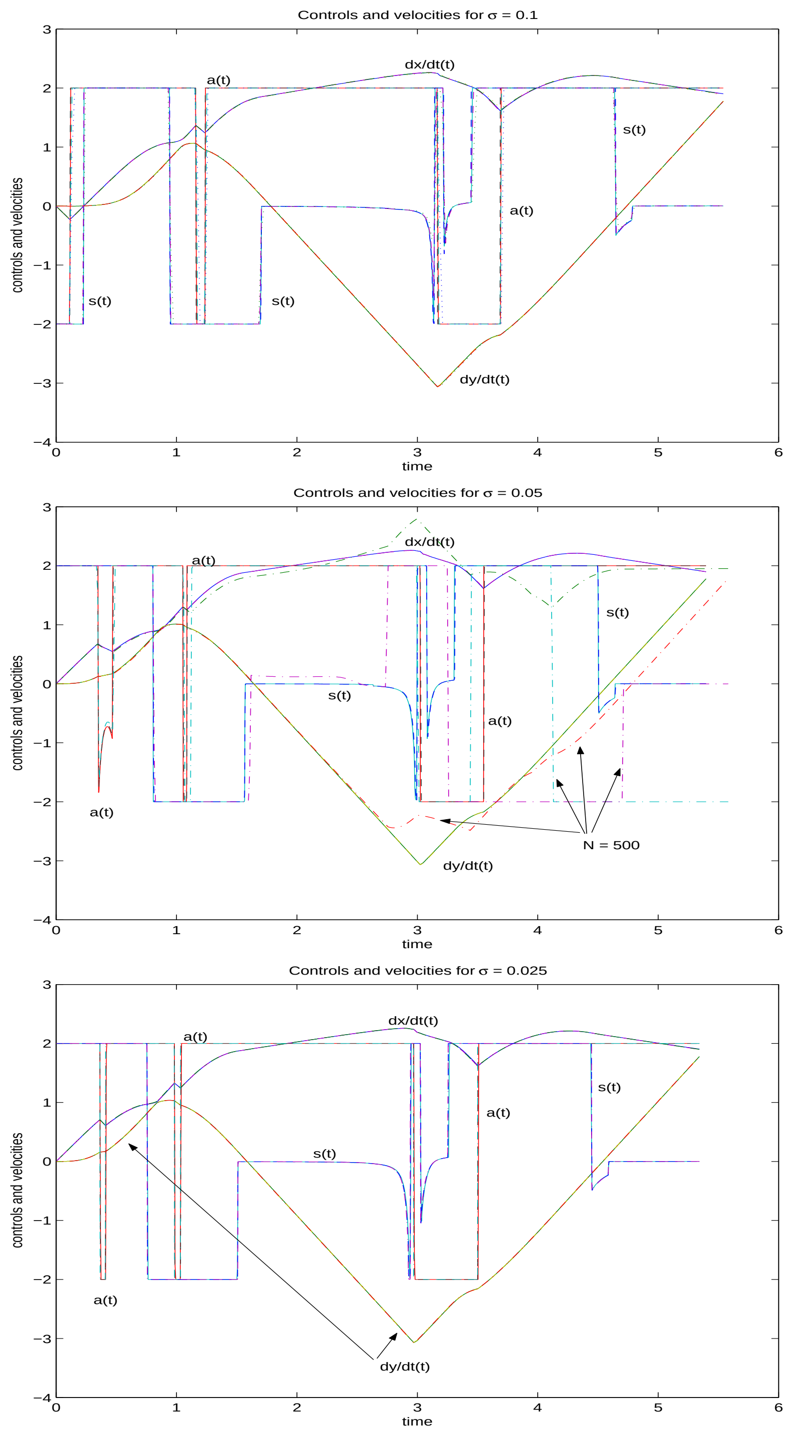

FIG. 7.5. *Controls and velocities for varying N and σ ($\sigma = 0.1$, 0.05, 0.025)*

For $\sigma = 0.1$ there is the initial reversing maneuver, which is not present for smaller values of σ. This is presumably because reducing the value of σ keeps the normal velocity closer to zero and enables the vehicle to make the first counter-clockwise turn without leaving the track. However, for $\sigma = 0.05$ there is a deceleration maneuver near $t = 0.35$ which appears to be of a bang-singular-bang type; for $\sigma = 0.025$ there is there is another deceleration maneuver at about the same time of a bang-bang type.

For $\sigma = 0.05$ and $N = 500$ the computed solution appears to be a local but not global minimum. Note that the objective function for this case is ≈ 5.59 compared with ≈ 5.41 for $N = 1000$ and $N = 2000$ with $\sigma = 0.05$, a difference of 3%. This solution (plotted with "dot-dash" lines) does not have the interesting steering maneuvers near $t = 3$ and near $t = 4.6$, and has quite different control functions near the finish line.

Otherwise the controls are qualitatively and quantitatively very similar. So it would seem that most of the maneuvers observed in the optimal controls are not artefacts, but are truly part of the optimal strategy.

8. Conclusions. Optimal control problems where the dynamics includes discontinuous right-hand sides or differential inclusions can be handled both analytically and computationally by smoothing the right-hand side, at least when the right-hand side satisfies a one-sided Lipschitz condition. Furthermore, not smoothing the right-hand side, and computing the gradients of the discretized system directly, leads to incorrect gradient information with errors comparable to the size of the true gradients even with fully implicit discretization. In order for the computed gradients to converge to the true gradients we need $h = o(\sigma^2)$ where h is the step-size and σ is the smoothing parameter; this is sufficient under non-degeneracy assumptions. Furthermore, we used the computed gradients with modern optimization software to compute optimal controls for moderately complex optimization problems with solutions that could not be easily predicted *a priori*. We expect that such results can be observed for a large class of problems.

The usual adjoint equation or inclusion must be modified to allow for jumps in the adjoint variables if the trajectory crosses the discontinuity. This is quite different to the case of differential inclusions with Lipschitz right-hand sides [6, 16, 19] where an adjoint differential inclusion can be constructed and solved.

A "jump formula" for the adjoint variable in the case where the trajectory crosses a codimension-one discontinuity has been developed which uses only the right-hand side on opposite sides of the discontinuity and the normal vector of the discontinuity manifold. This jump formula opens up the possibility of accurately computing gradients by using a detect/locate/restart method of handling discontinuities in the forward ODE solver.

Acknowledgements. This work was supported by the Mathematical, Information, and Computational Sciences Division subprogram of the Office of Advanced Scientific Computing Research, Office of Science, U.S. Department of Energy, under Contract W-31-109-ENG-38. David Stewart was supported in part by NSF collaborative grant DMS-0139708.

REFERENCES

[1] K. E. Atkinson, *An Introduction to Numerical Analysis*, J. Wiley and sons, 1978. 1st edition.

[2] J.-P. Aubin and A. Cellina, *Differential Inclusions: Set-Valued Maps and Viability Theory*, Springer-Verlag, Berlin, New York, 1984.

[3] J.-P. Aubin and H. Frankowska, *Set-Valued Analysis*, Progress in Systems and Control, #2, Birkhäuser, Boston, Basel, Berlin, 1990.
[4] H. Brézis, *Opérateurs Maximaux Monotones et Semi-groupes de Contractions dans les Espaces de Hilbert*, North-Holland Publishing Co., Amsterdam, 1973. North-Holland Mathematics Studies, No. 5. Notas de Matemática (50).
[5] A. E. Bryson and Y.-C. Ho, *Applied Optimal Control*, Halsted Press, Washington, DC, 1975.
[6] F. H. Clarke, *The maximum principle under minimal hypotheses*, SIAM J. Control Optimization, 14 (1976), pp. 1078–1091.
[7] F. H. Clarke, *Optimal control and the true hamiltonian*, SIAM Rev., 21 (1979), pp. 157–166.
[8] ———, *Methods of dynamic and nonsmooth optimization*, CBMS–NSF Reg. Conf. Ser. #57, SIAM Publ., Philadelphia, PA, 1989.
[9] ———, *Nonsmooth Analysis and Optimization*, SIAM Publ., Philadelphia, PA, 1990. Originally published by the Canadian Math. Soc., 1983.
[10] G. Dal Maso and F. Rampazzo, *On systems of ordinary differential equations with measures as controls*, Differential Integral Equations, 4 (1991), pp. 739–765.
[11] B. J. Driessen and N. Sadegh, *Minimum-time control of systems with Coulomb friction: near global optima via mixed integer linear programming*, Optimal Control Appl. Methods, 22 (2001), pp. 51–62.
[12] ———, *On the discontinuity of the costates for optimal control problems with Coulomb friction*, Optimal Control Appl. Methods, 22 (2001), pp. 197–200.
[13] A. F. Filippov, *Differential Equations with Discontinuous Right-Hand Side*, Kluver Academic, 1988.
[14] R. Fourer, D. M. Gay, and B. W. Kernighan, *AMPL: A Modeling Language for Mathematical Programming*, Brooks/Cole — Thomson Learning, 2nd ed., 2003.
[15] H. Frankowska, *The maximum principle for a differential inclusion problem*, in Analysis and optimization of systems, Part 1 (Nice, 1984), Berlin, 1984, Springer, pp. 517–531.
[16] H. Frankowska, *Adjoint differential inclusions in necessary conditions for the minimal trajectories of differential inclusions*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 2 (1985), pp. 75–99.
[17] H. Frankowska, *Le principe de maximum pour une inclusion différentielle avec des contraintes sur les états initiaux et finaux*, C. R. Acad. Sci. Paris Sér. I Math., 302 (1986), pp. 599–602.
[18] ———, *The maximum principle for an optimal solution to a differential inclusion with end points constraints*, SIAM J. Control Optim., 25 (1987), pp. 145–157.
[19] H. Frankowska and B. Kaškosz, *A maximum principle for differential inclusion problems with state constraints*, Systems Control Lett., 11 (1988), pp. 189–194.
[20] S. Galán, W. F. Feehery, and P. I. Barton, *Parametric sensitivity functions for hybrid discrete/continuous systems*, Appl. Numer. Math., 31 (1999), pp. 17–47.
[21] R. V. Gamkrelidze, *Principles of Optimal Control Theory*, Plenum Press, London, New York, 1978. Orig. in Russian, 1975.
[22] R. Glowinski and A. J. Kearsley, *On the simulation and control of some friction constrained motions*, SIAM J. Optim., 5 (1995), pp. 681–694.
[23] T.-H. Kim and I.-J. Ha, *Time-optimal control of a single-DOF mechanical system with friction*, IEEE Trans. Automat. Control, 46 (2001), pp. 751–755.
[24] S. C. Lipp, *Brachistochrone with Coulomb friction*, SIAM J. Control Optim., 35 (1997), pp. 562–584.
[25] J. Outrata, M. Kočvara, and J. Zowe, *Nonsmooth Approaches to Optimization Problems with Equilibrium Constraints*, vol. 28 of Nonconvex Optimization and Its Applications, Kluwer Academic, Dordrecht, Boston, London, 1998.
[26] L. S. Pontryagin, V. G. Boltjanskij, R. V. Gamkrelidze, and E. F. Mishchenko, *The Mathematical Theory of Optimal Processes*, Interscience Publ., New York, 1962. Original in Russian (1956).
[27] D. Stewart, *A high accuracy method for solving ODEs with discontinuous right-hand side*, Numer. Math., 58 (1990), pp. 299–328.
[28] D. E. Stewart, *Numerical methods for friction problems with multiple contacts*, J. Austral. Math. Soc., Ser. B, 37 (1996), pp. 288–308.
[29] D. E. Stewart, *The "Michael Schumacher" problem*. Available via URL `http://www.cs.wisc.edu/cpnet/cpnetmeetings/iccp99/race-car/race-car.html`, June 1999.
[30] K. Taubert, *Differenz Verfahren für gewöhnliche Anfangswertaufgaben mit unstetiger rechte Seite*, in Numerische Behandlung nichtlinearer Integrodifferential- und Differentialgleichungen, A. Dold and B. Eckmann, eds., 1974, pp. 137–148. Lecture notes ser#395.

[31] ———, *Differenzverfahren für Schwingungen mit trockener und zäher Reibung und für Regelungsysteme*, Numer. Math., 26 (1976), pp. 379–395.

[32] ———, *Converging multistep methods for initial value problems involving multivalued maps*, Computing, 27 (1981), pp. 123–136.

[33] J. E. Tolsma and P. I. Barton, *Hidden discontinuities and parametric sensitivity calculations*, SIAM J. Sci. Comput., 23 (2002), pp. 1861–1874 (electronic).

[34] L. G. van Willigenburg and R. P. H. Loop, *Computation of time-optimal controls applied to rigid manipulators with friction*, Internat. J. Control, 54 (1991), pp. 1097–1117.

[35] R. J. Vanderbei, *LOQO User's Manual, Version 4.05*, Princeton University, Operations Research and Financial Engineering Department, October 2000. Technical Report ORFE-99-??

[36] D. Ventura and T. Martinez, *Optimal control using a neural/evolutionary hybrid system*, in Proc. Int. Joint Conf. on Neural Networks, May 1998, pp. 1036–1041.