Nonsmooth optimal value and policy functions for mechanical systems subject to unilateral constraints

Bora S. Banjanin Samuel A. Burden

***Abstract*— State–of–the–art approaches to optimal control of contact–rich robot dynamics use smooth approximations of value and policy functions and gradient–based algorithms for improving approximator parameters. Unfortunately, the dynamics of mechanical systems subject to unilateral constraints—i.e. robot locomotion and manipulation—are generally nonsmooth. We show that value and policy functions generally inherit regularity properties like (non)smoothness from the underlying system's dynamics, and demonstrate this effect in a simple mechanical system. We conclude with a discussion of implications for the use of gradient–based algorithms for optimal control of contact–rich robot dynamics.**

I. Introduction

In the optimization approach to robot control, a policy is sought that extremizes a given performance criterion; the performance achieved by this *optimal policy* is the *optimal value* of the problem. Two widely-applied frameworks for solving such problems are reinforcement learning and trajectory optimization. Although many algorithms are available in either framework, scalable algorithms in both leverage local approximations—gradients of values and policies—to iteratively improve toward optimality. In robotics applications like collision–free motion planning, these gradients are guaranteed to exist and can be readily computed or approximated. We show in this paper that these gradients can fail to exist for contact-rich robot dynamics, precluding application of state–of–the–art algorithms for optimal control.

We begin in Section II by modeling contact–rich robot dynamics using mechanical systems subject to unilateral constraints, and describe how nonsmoothness—discontintinuity or piecewise–differentiability—manifests in trajectory outcomes and (hence) trajectory costs. Then in Section III we provide mathematical derivations that show nonsmoothness in trajectory outcomes and costs gives rise to nonsmoothness in optimal value and (hence) policy functions. Subsequently in Section IV we present numerical simulations that demonstrate discontinuous or merely piecewise–differentiable optimal value and policy functions in a mechanical system subject to unilateral constraints. Finally in Section V we discuss the prevelance of nonsmoothness and how the lack of classical differentiability prevents gradient–based algorithms from converging to optimality.

II. Mechanical systems subject to unilateral constraints

In this section, we formalize a class of models for contact–rich dynamics in robot locomotion and manipulation as *mechanical systems subject to unilateral constraints* and formulate an optimal control problem for these systems.

A. Dynamics

Consider the dynamics of a mechanical system with configuration coordinates $q \in Q = \mathbb{R}^d$ subject to unilateral constraints $a(q) \geq 0$ specified by a differentiable function $a : Q \to \mathbb{R}^n$ where $d, n \in \mathbb{N}$ are finite. We are primarily interested in systems with $n > 1$ constraints, whence we regard the inequality $a(q) \geq 0$ as being enforced componentwise.

Given any $J \subset \{1, \ldots, n\}$, and letting $|J|$ denote the number of elements in the set J, we let $a_J : Q \to \mathbb{R}^{|J|}$ denote the function obtained by selecting the component functions of a indexed by J, and we regard the equality $a_J(q) = 0$ as being enforced componentwise.

It is well–known (cf. [1, Sec. 3] or [2, Sec. 2.4, 2.5]) that, with $J = \{j \in \{1, \ldots, n\} : a_j(q) = 0\}$ denoting the *contact mode*, the system's dynamics take the form

$$M(q)\ddot{q} = f_J(q, \dot{q}, u) + c(q, \dot{q})\dot{q} + Da_J(q)^\top \lambda_J(q, \dot{q}, u), \tag{1a}$$

$$\dot{q}^+ = \Delta_J(q, \dot{q}^-), \tag{1b}$$

where $M : Q \to \mathbb{R}^{d \times d}$ specifies the mass matrix for the mechanical system in the q coordinates, $f_J : TQ \to \mathbb{R}^d$ is termed the *effort map* [1] and specifies[1] the internal and applied forces, $u \in \mathcal{U}$ is an external input, $c : TQ \to \mathbb{R}^{d \times d}$ denotes the *Coriolis matrix* determined by M, $Da_J : Q \to \mathbb{R}^{|J| \times d}$ denotes the (Jacobian) derivative of the constraint function a_J with respect to the coordinates, $\lambda_J : TQ \to \mathbb{R}^{|J|}$ denotes the reaction forces generated in contact mode J to enforce $a_J(q) \geq 0$, $\Delta_J : TQ \to \mathbb{R}^{d \times d}$ specifies the collision restitution law that instantaneously resets velocities to ensure compatibility with the constraint $a_J(q) = 0$, and $\dot{q}^+$ (resp. $\dot{q}^-$) denotes the right– (resp. left–)handed limits of the velocity with respect to time.

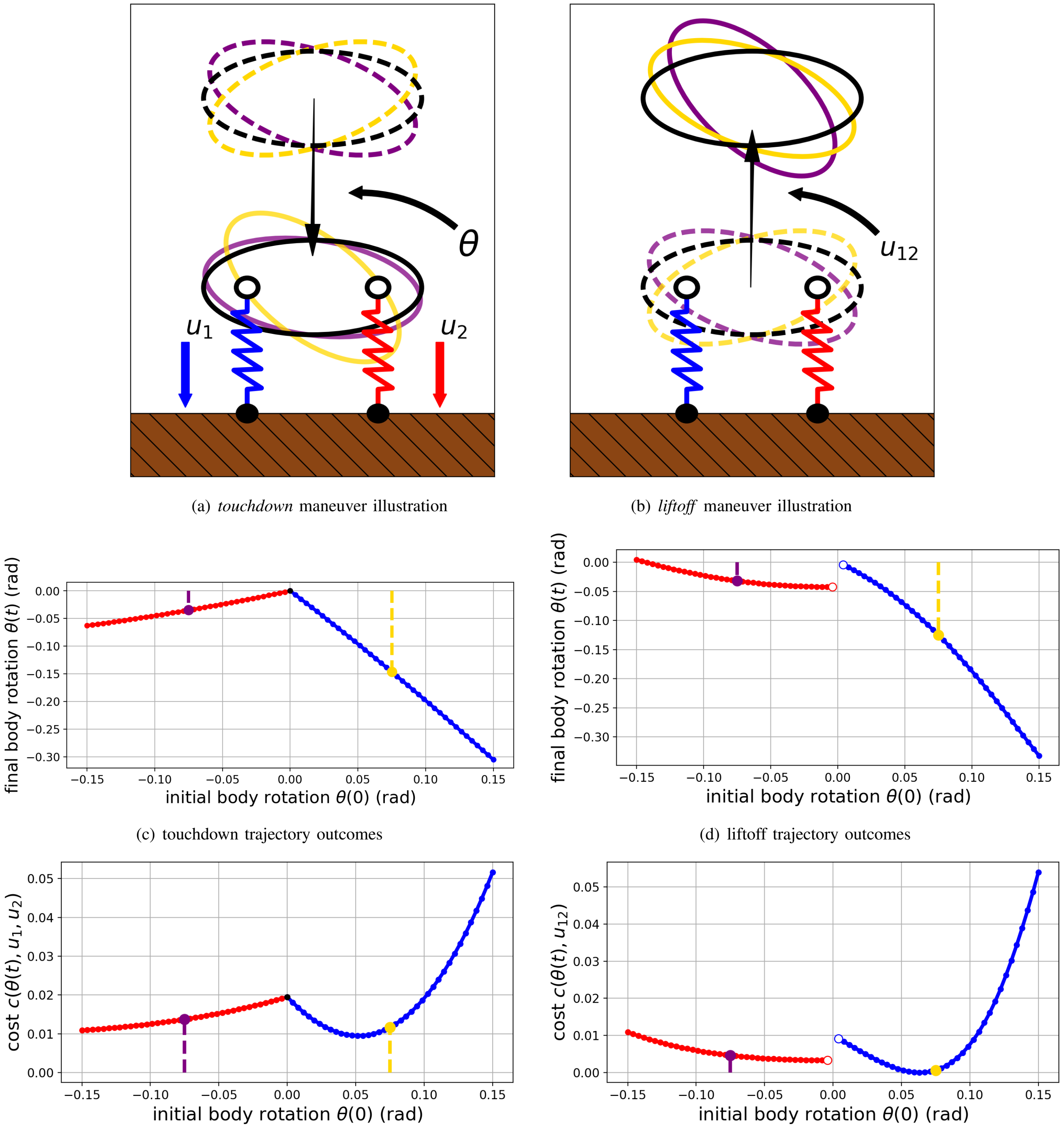
B. Regularity of dynamics

The seemingly benign equations in (1) can yield dynamics with a range of regularity properties. This issue has been thoroughly investigated elsewhere [1], [3], [4]; here we focus specifically on how design choices in a robot's *mechanical* and *control* systems affect regularity of its dynamics.

In what follows, we will frequently refer to the concept of a control system's *flow*, so we briefly review the concept before proceeding. Given a control system (e.g. (1) or (2))

Department of Electrical Engineering, University of Washington, Seattle, WA, USA (`borab,sburden@uw.edu`).

[1] We let $TQ = \mathbb{R}^d \times \mathbb{R}^d$ denote the *tangent bundle* of the configuration space Q; an element $(q, \dot{q}) \in TQ$ can be regarded as a pair of generalized configurations $q \in \mathbb{R}^d$ and velocities $\dot{q} \in \mathbb{R}^d$; we write $\dot{q} \in T_qQ$.



(a) *touchdown* maneuver illustration

(b) *liftoff* maneuver illustration

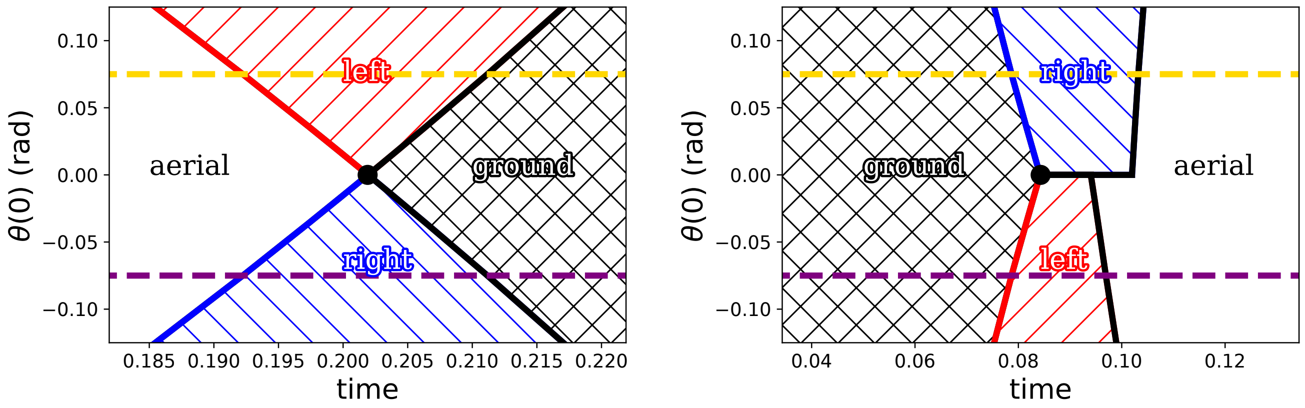
(c) touchdown trajectory outcomes

(d) liftoff trajectory outcomes

(e) touchdown value

(f) liftoff value

Fig. 1. *Piecewise–differentiable and discontinuous trajectory outcomes in saggital–plane biped.* (a,b) Illustration of two maneuvers—*touchdown* and *liftoff*—performed under non–optimal policies that exert different forces depending on which feet are in contact with the ground. In the *touchdown* maneuver, feet are initially off the ground and trajectories terminate when the body height reaches nadir; in the *liftoff* maneuver, feet are initially on the ground and trajectories terminate when the body height reaches apex. (c,d) Trajectory outcomes (final body angle $\theta(t)$) as a function of initial body angle $\theta(0)$. (e,f) Performance of trajectories as measured by the cost functions in (20), (21). Dashed colored vertical lines on (c–f) indicate corresponding colored outcomes on (a,b).



(a) touchdown contact modes

(b) liftoff contact modes

Fig. 2. *Contact modes for touchdown and liftoff maneuvers.* The saggital–plane biped illustrated in Figure 1(a,b) can be in one of four *contact modes* corresponding to which subset $J \subset \{1,2\}$ of the (two) limbs are in contact with the ground; each subset yields different dynamics in (1). (a,b) System contact mode at each time t for a given initial body rotation $\theta(0)$; the body torque input is zero ($u_{12} = 0$) and the leg forces are different ($u_1 \neq u_2$) in mode *left* ($\{1\}$) and *right* ($\{2\}$) than in *aerial* ($\emptyset$) or ground ($\{1,2\}$). Dashed colored horizontal lines indicate corresponding colored trajectories in Figure 1. The increase in force during the transition to modes *left* and *right* in (b) changes the ground reaction force discontinuously, delaying liftoff and causing discontinuous trajectory outcomes in Figure 1(d).

with state space $\mathcal{X}$ and input space $\mathcal{U}$, a *flow* is a function $\phi : [0,t] \times \mathcal{X} \times \mathcal{U}^{[0,t]} \to \mathcal{X}$ such that for all initial states $x \in \mathcal{X}$ and inputs $(u : [0,t] \to \mathcal{U}) \in \mathcal{U}^{[0,t]}$, the function $\phi^{x,u} : [0,t] \to \mathcal{X}$ defined for all $s \in [0,t]$ by $\phi^{x,u}(s) = \phi(s,x,u)$ is a trajectory for the control system. Intuitively, the flow "bundles" all trajectories into a single function. Mathematically, the flow is useful for studying how trajectories vary with respect to initial states and inputs. So long as trajectories exist and are unique for every $x \in \mathcal{X}$ and $u \in \mathcal{U}^{[0,t]}$, the flow is a well–defined function.

It is common to assume that the functions in (1) are continuously–differentiable ($M, f, a, \gamma \in C^r$); however, as illustrated by [1, Ex. 2], this assumption alone does not ensure existence or uniqueness of trajectories. This case contrasts starkly with that of classical control systems, where the equation

$$\dot{x} = F(x,u) \quad (2)$$

yields unique trajectories whose regularity matches the vector field's: if F is continuously differentiable, then there exists a flow for (2) that is continuously differentiable to the same order.

To ensure trajectories for (1) exist uniquely, restrictions must be imposed; we refer the interested reader to [1, Thm. 10] for a specific result and [2] for a general discussion of this issue. Since we are chiefly concerned with how properties of the dynamics in (1) affect properties of optimal value and policy functions, we will assume in what follows that conditions have been imposed to ensure (1) has a flow for states, inputs, and time horizons of interest.

Assuming that a flow ϕ exists for (1) does not provide any regularity properties on the function ϕ; these properties are determined by the design of a robot's mechanical and control systems and their closed–loop interaction with the environment. For instance: when limbs are inertially coupled (e.g. by rigid struts and joints), so that one limb's constraint activation instantaneously changes another's velocity, ϕ is discontinuous at configurations where these two limbs activate constraints simultaneously [5, Table 3] [6]; when limbs are force coupled (e.g. by nonlinear damping), so that one limb's constraint (de)activation instantaneously changes the force on another, ϕ can be piecewise–differentiable at configurations where these two limbs (de)activate constraints simultaneously [3, Fig. 1]. In both instances, mechanical design choices lead to nonsmooth dynamics; Figure 1 provides examples where control design choices lead to nonsmooth dynamics (piecewise–differentiable in Figure 1(a,c,e), discontinuous in Figure 1(b,d,f)). Other nonsmooth phenomena can arise, e.g. *grazing*[2] and *Zeno*[3] trajectories; in what follows we will focus on the case of simultaneous constraint (de)activations due to its prevalence in robot gaits and maneuvers (see Section V-A for a discussion of when this phenomena prevails).

C. Regularity of optimal value and policy functions

A broad class of optimal control problems for the dynamics in (1) can be formulated in terms of *final* ($\ell : \mathcal{X} \to \mathbb{R}$) and *running* ($\mathcal{L} : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$) costs:

$$\nu(x) = \min_{u \in \mathcal{U}^{[0,t]}} \ell(x^u(t)) + \int_0^t \mathcal{L}(x^u(s), u(s))\, ds, \quad (3)$$

where $x^u : [0,t] \to \mathcal{X}$ denotes the unique trajectory obtained from initial state $x^u(0) \in \mathcal{X}$ when input $u \in \mathcal{U}^{[0,t]}$ is applied; in terms of the flow, $x^u(s) = \phi(s, x(0), u)$ for all $s \in [0,t]$. To expose the dependence of the cost in (3) on the flow ϕ, we transcribe the problem in (3) to a simpler form using a

[2] where a constraint function a_j decreases to and then increases from zero without activating constraint j

[3] where a constraint is activated an infinite number of times on a finite time horizon

standard state augmentation technique (cf. [7, Ch. 4.1.2]):

$$\nu(x) = \min_{u \in \mathcal{U}^{[0,t]}} c\left(\phi(t, x, u)\right) \tag{4}$$

As discussed in Section II-B, the continuity and differentiability properties of ϕ are partly determined by a robot's design: it is possible for ϕ and hence $c \circ \phi$ to be discontinuous ($\phi \notin C^0$), continuously–differentiable ($\phi \in C^r$), or piecewise–differentiable and not continuously–differentiable ($\phi \in PC^r \setminus C^r$), depending on the properties of the robot's mechanical and control systems. In the next section, we study how continuity and differentiability properties of $c \circ \phi$ affect the corresponding properties of ν in (4).

III. Continuity and differentiability of optimal value and policy functions

Consider minimization of the cost function $c : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$ with respect to an input $u \in \mathcal{U}$:

$$\nu(x) = \min_{u \in \mathcal{U}} c(x, u); \tag{5}$$

so long as $\mathcal{X}$ and $\mathcal{U}$ are compact and c is continuous, the function $\nu : \mathcal{X} \to \mathbb{R}$ indicated in (5), termed the *optimal value function*, is well–defined. We let $\pi : \mathcal{X} \to \mathcal{U}$ denote an *optimal policy* for (5), i.e.

$$\forall x \in \mathcal{X} : \pi(x) \in \arg\min_{u \in \mathcal{U}} c(x, u) \tag{6}$$

or, equivalently,

$$\forall x \in \mathcal{X} : \nu(x) = c(x, \pi(x)). \tag{7}$$

In this section we study how regularity properties (continuity, differentiability) of the cost function (c) relate to regularity properties of optimal value (ν) and policy (π) functions.

A. Discontinuous cost functions

If the cost ($c : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$) is discontinuous with respect to its first argument, then the optimal policy ($\pi : \mathcal{X} \to \mathcal{U}$) and value ($\nu : \mathcal{X} \to \mathbb{R}$) are generally discontinuous as well. This observation is clear in the trivial case that the cost only depends on its first argument, but manifests more generally.

B. Continuously–differentiable cost functions

This section contains straightforward calculations based on standard results in classical (smooth) Calculus and nonlinear programming; it is provided primarily as a rehearsal for the more general setting considered in the subsequent section.

If c is continuously–differentiable, denoted $c \in C^1(\mathcal{X} \times \mathcal{U}, \mathbb{R})$ or simply $c \in C^1$, then necessarily [7, Ch. 1.1.1]

$$D_2 c(x, \pi(x)) = 0. \tag{8}$$

If c is two times continuously–differentiable (denoted $c \in C^2$) and the second–order sufficient condition [7, Ch. 1.1.2] for strict local optimality for (5) is satisfied at $\pi(x) \in \mathcal{U}$,

$$D_2^2 c(x, \pi(x)) > 0, \tag{9}$$

then the C^1 Implicit Function Theorem (IFT) [8, Thm. C.40] can be applied to (7) to choose π as a C^1 function near x. Note that IFT specifically required the invertibility tacit in (9):

$$\text{the linear function } D_2^2 c(x, \pi(x)) : T_u\mathcal{U} \to T_u\mathcal{U} \text{ is invertible.} \tag{10}$$

If (8) and (9) are satisfied, then applying the C^1 Chain Rule [8, Prop. C.3] to (8) yields

$$D\pi(x) = -D_2^2 c(x, \pi(x))^{-1} \left(D_{12} c(x, \pi(x))\right), \tag{11}$$

and applying the C^1 Chain Rule to (7) yields

$$\begin{aligned} D\nu(x) &= D_x c(x, \pi(x)) \\ &= D_1 c(x, \pi(x)) + D_2 c(x, \pi(x)) D\pi(x), \end{aligned} \tag{12}$$

whence we obtain derivatives of the optimal value and policy functions in terms of derivatives of the cost function.

We conclude that if the cost function is two times continuously–differentiable ($c \in C^2$) and first–order necessary (8) and second–order sufficient (9), (10) conditions for optimality and stability of solutions to (5) are satisfied at $u = \pi(x)$, then the optimal policy and value functions are continuously–differentiable at x ($\pi, \nu \in C^1$) and their derivatives at x can be computed using (11), (12).

Proposition 1: If $c \in C^2(\mathcal{X} \times \mathcal{U}, \mathbb{R})$ satisfies (8), (9), and (10) at $(\xi, \mu) \in \mathcal{X} \times \mathcal{U}$, then there exist neighborhoods $X \subset \mathcal{X}$ of ξ and $U \subset \mathcal{U}$ of μ and a function $\pi \in C^1(X, U)$ such that $\pi(\xi) = \mu$ and, for all $x \in X$, $\pi(x)$ is the unique minimizer for

$$\nu(x) = \min_{u \in U} c(x, u); \tag{13}$$

the derivative of π is given by (11), and the derivative of ν is given by (12).

C. Piecewise–differentiable cost functions

If c is piecewise–differentiable,[4] denoted $c \in PC^1(\mathcal{X} \times \mathcal{U}, \mathbb{R})$ or simply $c \in PC^1$, then necessarily

$$\forall w \in T_u\mathcal{U} : D_2 c(x, \pi(x); w) \geq 0. \tag{14}$$

Here and below, $D_2 c(x, \pi(x)) : T_u\mathcal{U} \to \mathbb{R}$ denotes a continuous and piecewise–linear first–order approximation termed the *Bouligand* (or *B*–)derivative [9, Ch. 3] that exists by virtue of the cost being PC^1 [9, Lem. 4.1.3]; $D_2 c(x, \pi(x); w)$ denotes the evaluation of $D_2 c(x, \pi(x))$ at $w \in T_u\mathcal{U}$.

If c is two times piecewise–differentiable (denoted $c \in PC^2$), and if a sufficient condition [10, Thm. 1] for strict local optimality for (5) is satisfied at $\pi(x) \in \mathcal{U}$,

$$\begin{aligned} \forall w \in \{w \in T_u\mathcal{U} \mid w \neq 0,\ D_2 c(x, \pi(x); w) = 0\} \\ : D_2^2 c(x, \pi(x); w, w) > 0, \end{aligned} \tag{15}$$

and if the piecewise–linear function

$$D_2^2 c(x, \pi(x)) : T_u\mathcal{U} \to T_u\mathcal{U} \text{ is invertible,} \tag{16}$$

[4] We use the notion of piecewise–differentiability from [9, Ch. 4.1]: a function is piecewise–differentiable if it is everywhere locally a continuous selection of a finite number of continuously–differentiable functions.

then a PC^1 Implicit Function Theorem can be applied to choose $\pi \in PC^1$ near x [11, Cor. 3.4].[5] Applying the PC^1 Chain Rule [9, Thm. 3.1.1] to (14) yields (cf. [11, § 3])

$$\begin{aligned} \forall v \in & T_x\mathcal{X} : D\pi(x;v) = \\ & - D_2^2 c(x,\pi(x))^{-1} \left(D_{12}c(x,\pi(x);v) \right), \end{aligned} \tag{17}$$

and applying the PC^1 Chain Rule to (7) yields

$$\begin{aligned} \forall v \in & T_x\mathcal{X} : D\nu(x;v) = D_x c(x,\pi(x);v) \\ & = D_1 c(x,\pi(x);v) + D_2 c(x,\pi(x);D\pi(x;v)), \end{aligned} \tag{18}$$

whence we obtain B–derivatives of the optimal value and policy functions in terms of B–derivatives of the cost.

We conclude that if the cost function is two times piecewise–differentiable ($c \in PC^2$) and first–order necessary (14) and second–order sufficient (15), (16) conditions for optimality and stability of solutions to (5) are satisfied at $u = \pi(x)$, then the optimal policy and value functions are piecewise–differentiable at x ($\pi, \nu \in PC^1$) and their B–derivatives at x can be computed using (17), (18).

Proposition 2: If $c \in PC^2(\mathcal{X} \times \mathcal{U}, \mathbb{R})$ satisfies (14), (15), and (16) at $(\xi, \mu) \in \mathcal{X} \times \mathcal{U}$, then there exist neighborhoods $X \subset \mathcal{X}$ of ξ and $U \subset \mathcal{U}$ of μ and a function $\pi \in PC^1(X, U)$ such that $\pi(\xi) = \mu$ and, for all $x \in X$, $\pi(x)$ is the unique minimizer for

$$\nu(x) = \min_{u \in U} c(x,u); \tag{19}$$

the B–derivative of π is given by (17), and the B–derivative of ν is given by (18).

D. Conclusions regarding regularity of optimal value and policy functions

The results in Sections III-A, III-B, and III-C suggest that we should generally expect regularity of optimal value and policy functions to match that of the cost function: they should be discontinuous when the cost is discontinuous, or piecewise–differentiable when the cost is piecewise–differentiable. In Section IV we provide instances of the class of models described in Section II that exhibit these effects.

IV. OPTIMAL VALUE AND POLICY FUNCTIONS FOR A MECHANICAL SYSTEM SUBJECT TO UNILATERAL CONSTRAINTS

We showed in the previous section that optimal value and policy functions for contact–rich robot dynamics inherit nonsmoothness from the underlying dynamics. To instantiate this result, we crafted the simplest mechanical system subject to unilateral constraints that exhibits the nonsmooth phenomena of interest (piecewise–differentiable or discontinuous trajectory outcomes), yielding the *touchdown* and *liftoff* maneuvers shown in Figure 1(a,b). For the touchdown maneuver, we seek the optimal (constant) force to exert in the left leg (u_1) when the left foot is in contact and the right foot is not; similarly, we seek the optimal choice of force in the right leg (u_2) when the right foot is in contact and the left foot is not: with $a_1, a_2 > 0$ as input penalty parameters,

$$c_{\text{touchdown}}(\theta_{\text{nadir}}, u_1, u_2) = (\theta_{\text{nadir}} - \theta_{\text{desired}})^2 + a_1 u_1^2 + a_2 u_2^2. \tag{20}$$

For the liftoff maneuver, we seek the optimal (constant) torque (u_{12}) to apply to the body while both feet are in contact: with $a_{12} > 0$ as an input penalty parameter,

$$c_{\text{liftoff}}(\theta_{\text{apex}}, u_{12}) = (\theta_{\text{nadir}} - \theta_{\text{desired}})^2 + a_{12} u_{12}^2. \tag{21}$$

We implemented numerical simulations of these models[6] and applied a scalar minimization algorithm[7] to compute optimal policies as a function of initial body rotation.[8]

As expected, the optimal value and policy functions computed for the touchdown and liftoff maneuvers are nonsmooth (Figure 3(c,d,e,f)). This result does not depend sensitively on the problem data; nonsmoothness is preserved after altering parameters of the model and/or cost function. We emphasize that the nonsmoothness in Figure 3 arises from the nonsmoothness in the underlying system dynamics (1); the functions in (20) and (21) are smooth.

V. DISCUSSION

We conclude by discussing how often we expect to encounter the nonsmooth phenomena described above in models of robot behaviors (Section V-A) and what our results imply about the use of smooth tools in this nonsmooth setting (Section V-B).

A. Prevalence of nonsmooth phenomena

In Section IV, we presented two simple optimal control problems where the dynamics of a mechanical system subject to unilateral constraints gave rise to a nonsmooth cost: one where the cost was piecewise–differentiable, and another where it was discontinuous. The reader may have noticed that the nonsmoothness occurred along trajectories that underwent simultaneous constraint (de)activation. This peculiarity was not accidental: the cost is generally continuously–differentiable along trajectories that (de)activate constraints at distinct instants in time.[9]

If the constraint surfaces intersect transversely [8, Ch. 6], then the nonsmoothness presented in Section IV is confined to a subset of the state space with zero Lebesgue measure. In light of this observation, intuition may lead one to ignore these states in practice. However, we believe this intuition will lead the practitioner astray as the complexity of considered behaviors increases. Indeed, since the number of contact mode sequences increases factorially with the number of constraints and exponentially with the number of constraint (de)activations, then the region where the cost function is continuously–differentiable is "carved up" into a

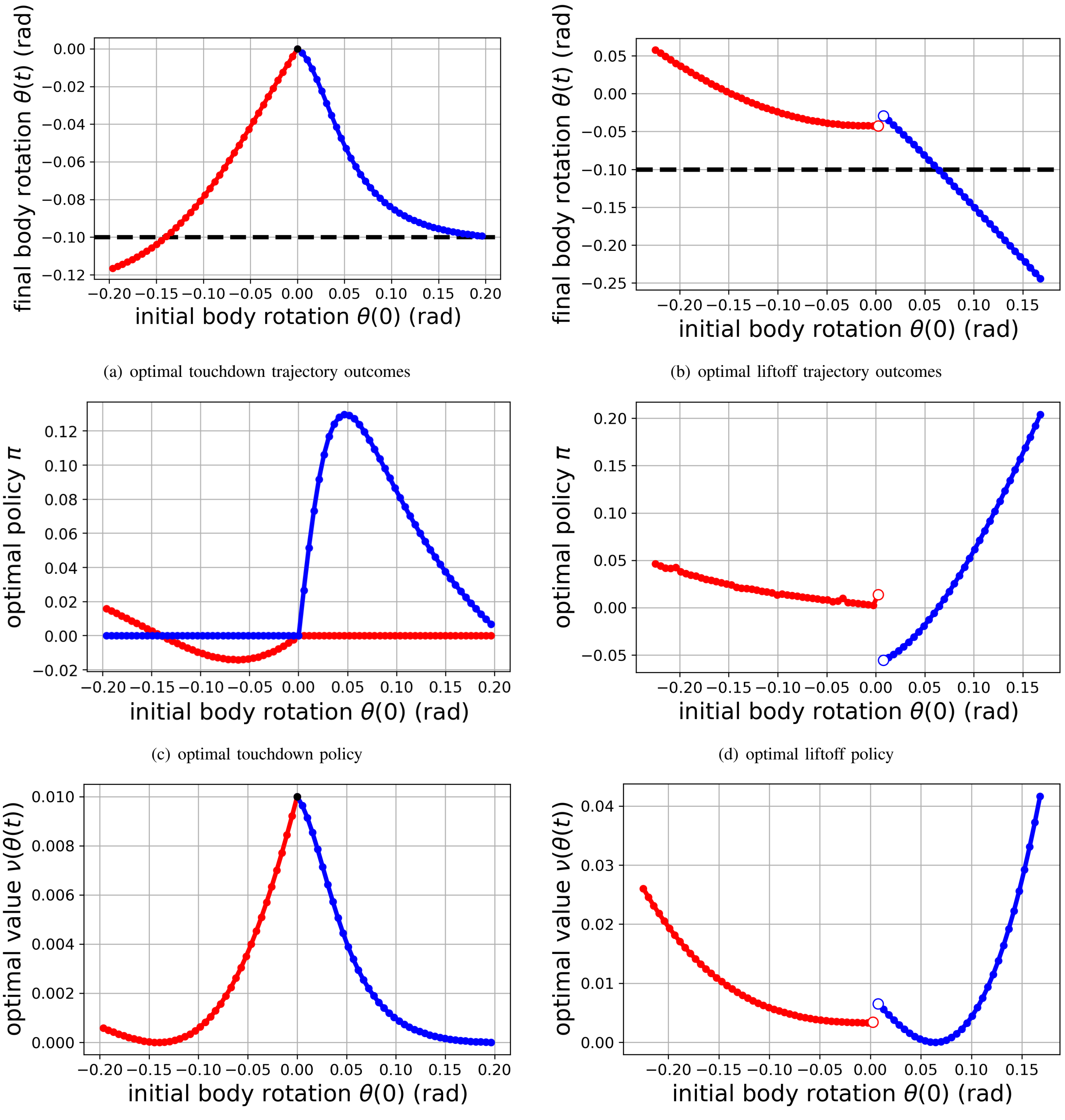
[5] This Implicit Function Theorem requires $D_2 c$ be *strongly* B–differentiable; the costs considered here are not generally strongly B–differentiable, but they are generally PC^r–equivalent to strongly B–differentiable functions [12, Thm. 3.1], whence [11, Cor. 3.4] can be applied indirectly.

[6] using the modeling framework in [2] and simulation algorithm in [13]

[7] `SciPy v0.19.0 minimize_scalar`

[8] We plan to release the software used to generate these results as an `environment` in `OpenAI Gym` [14].

[9] This follows from [15, Eqn. 2.3] so long as the constraint (de)activations are *admissible* [3, Def. 3, Lem. 1].



(a) optimal touchdown trajectory outcomes

(b) optimal liftoff trajectory outcomes

(c) optimal touchdown policy

(d) optimal liftoff policy

(e) optimal touchdown value

(f) optimal liftoff value

Fig. 3. *Optimal trajectories, values and policies for touchdown and liftoff maneuvers.* Optimizing (20), (21) for the biped in Figure 1 yields trajectory outcomes (a,b), policies (c,d), and values (e,f) that are nonsmooth (piecewise–differentiable or discontinuous). Asymmetries in trajectory outcomes are due to unequal input penalty parameters ($a_1 \neq a_2$) in (a) and unequal leg forces ($u_1 \neq u_2$) in (b).

rapidly increasing number of disjoint "pieces" as behavioral complexity[10] increases.

Although we cannot at present comment in general on how these smooth pieces fit together, we note that some important behaviors will reside near a large number of pieces. For instance, periodic behaviors with (near–)simultaneous (de)activation of $n \in \mathbb{N}$ constraints as in [16] could yield up to $(n!)^k$ pieces after $k \in \mathbb{N}$ periods. The combinatorics are similar for tasks that involve intermittently activating (a subset of) n constraints k times as in [17]. Since the dimension of the state space is independent of n and k, these pieces must be increasingly tightly packed as n and/or k increase.

B. Justifying the use of gradient–based algorithms

Suppose a (possibly non–optimal) policy $\pi : \mathcal{X} \to \mathcal{U}$ has an associated value $\nu^\pi : \mathcal{X} \to \mathbb{R}$. If this value admits a first–order approximation with respect to π, then it is natural to improve the policy using steepest descent: with $\alpha > 0$ as a stepsize parameter,

$$\pi^+ = \pi + \alpha \arg \min_{\|\delta\|=1} D_\pi \nu^\pi(\delta). \tag{22}$$

The update in (22) is a *direct policy gradient–based* algorithm [18], [19], and can be interpreted as a *natural* [20] or *trust region* [21] algorithm depending on the norm chosen. In practice, the derivative $D_\pi \nu^\pi$ is not generally available and must be estimated, e.g. using function approximation [22], [23] or sampling [19], [24]. This practice is justified for smooth control systems whose value functions are smooth; it is not generally justified for the mechanical systems subject to unilateral constraints considered here since the value of (optimal or non–optimal) policies can be nonsmooth. To see how nonsmoothness can prevent a gradient–based algorithm from converging to an optimal policy, consider the result of applying one step of the policy gradient algorithm in (22) to the optimal policies in Figure 3(c,d) when ν^π is merely piecewise–differentiable. Since the policy is optimal, the first–order necessary condition for optimality (14) implies that the $\arg\min$ in (22) evaluates to zero, and therefore the optimal policy is a fixed point of the update in (22) when the true (Bouligand) derivative $D_\pi \nu^\pi$ is available. However, an estimate of $D_\pi \nu^\pi$ obtained via sampling or function approximation would be nonzero near $\theta(0) = 0$, causing one step of the policy gradient algorithm in (22) to *diverge* from the optimal policy. This can be seen in Figure 4, where central finite differences was used to compute the change in policy determined by naive application of the *deterministic policy gradient* algorithm [24]. Near discontinuities in the optimal policy, the change from the optimal policy can be arbitrarily large.

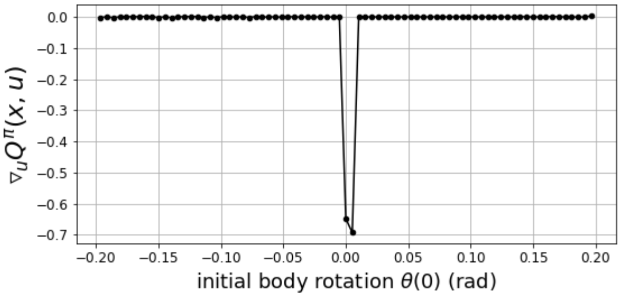


Fig. 4. Change in policy computed using using deterministic policy gradient.

Recent work employs smooth approximations of the contact–rich robot dynamics in (1) to enable application of gradient–based learning [25]–[27] and optimization [17], [28], [29] algorithms. This approach leverages established scalable algorithms, but does not ensure that policies optimized for the smoothed dynamics are (near–)optimal when applied to the original system's nonsmooth dynamics, since the dynamics of the smooth system being optimized differ from those of the original system. As an alternative approach, the framework we introduced in [4] provides design conditions that ensure trajectories of (1) depend continuously–differentiably on initial conditions. Thus in future work it may be possible to justify applying established algorithms for optimal control directly on some mechanical systems subject to unilateral constraints.

Acknowledgements

This material is based upon work supported by the U. S. Army Research Laboratory and the U. S. Army Research Office under contract/grant number W911NF-16-1-0158.

REFERENCES

[1] P. Ballard, "The dynamics of discrete mechanical systems with perfect unilateral constraints," *Archive for Rational Mechanics and Analysis*, vol. 154, no. 3, pp. 199–274, 2000.

[2] A. M. Johnson, S. A. Burden, and D. E. Koditschek, "A hybrid systems model for simple manipulation and self-manipulation systems," *The International Journal of Robotics Research*, vol. 35, no. 11, pp. 1354–1392, 1 Sept. 2016.

[3] A. M. Pace and S. A. Burden, "Piecewise–differentiable trajectory outcomes in mechanical systems subject to unilateral constraints," in *Proceedings of Hybrid Systems: Computation and Control (HSCC)*, 2017.

[4] ——, "Decoupled limbs yield differentiable trajectory outcomes through intermittent contact in locomotion and manipulation," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2017.

[5] C. D. Remy, K. Buffinton, and R. Siegwart, "Stability analysis of passive dynamic walking of quadrupeds," *The International Journal of Robotics Research*, vol. 29, no. 9, pp. 1173–1185, 2010.

[6] Y. Hürmüzlü and D. B. Marghitu, "Rigid body collisions of planar kinematic chains with multiple contact points," *The International Journal of Robotics Research*, vol. 13, no. 1, pp. 82–92, 1994.

[7] E. Polak, *Optimization: Algorithms and Consistent Approximations*. Springer–Verlag, 1997.

[8] J. M. Lee, *Introduction to Smooth Manifolds*, 2nd ed., ser. Graduate texts in mathematics. New York ; London: Springer, 2012.

[9] S. Scholtes, *Introduction to Piecewise Differentiable Equations*. Springer–Verlag, 2012.

[10] R. W. Chaney, "Second-Order sufficient conditions in nonsmooth optimization," *Mathematics of Operations Research*, vol. 13, no. 4, pp. 660–673, 1 Nov. 1988.

[11] S. M. Robinson, "An Implicit-Function theorem for a class of nonsmooth functions," *Mathematics of Operations Research*, vol. 16, no. 2, pp. 292–309, 1991.

[10] as measured by the number of constraints and/or constraint (de)activations

[12] L. Kuntz and S. Scholtes, "Structural analysis of nonsmooth mappings, inverse functions, and metric projections," *Journal of Mathematical Analysis and Applications*, vol. 188, no. 2, pp. 346–386, 1994.

[13] S. A. Burden, H. Gonzalez, R. Vasudevan, R. Bajcsy, and S. S. Sastry, "Metrization and Simulation of Controlled Hybrid Systems," *IEEE Transactions on Automatic Control*, vol. 60, no. 9, pp. 2307–2320, 2015.

[14] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "OpenAI gym," 5 June 2016.

[15] M. A. Aizerman and F. R. Gantmacher, "Determination of stability by linear approximation of a periodic solution of a system of differential equations with discontinuous Right–Hand sides," *The Quarterly Journal of Mechanics and Applied Mathematics*, vol. 11, no. 4, pp. 385–398, 1958.

[16] R. M. Alexander, "The gaits of bipedal and quadrupedal animals," *The International Journal of Robotics Research*, vol. 3, no. 2, pp. 49–59, 1984.

[17] I. Mordatch, E. Todorov, and Z. Popović, "Discovery of complex behaviors through contact-invariant optimization," *ACM Transactions on Graphics*, vol. 31, no. 4, pp. 43:1–43:8, July 2012.

[18] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," *Advances in Neural Information Processing Systems*, vol. 12, pp. 1057–1063, 2000.

[19] J. Baxter and P. L. Bartlett, "Infinite-horizon policy-gradient estimation," *The Journal of Artificial Intelligence Research*, vol. 15, pp. 319–350, 2001.

[20] S. Kakade, "A natural policy gradient," *Advances in Neural Information Processing Systems*, vol. 14, pp. 1531–1538, 2001.

[21] J. Schulman, S. Levine, P. Moritz, M. I. Jordan, and P. Abbeel, "Trust region policy optimization," *CoRR, abs/1502. 05477*, 2015.

[22] K. Doya, "Reinforcement learning in continuous time and space," *Neural Computation*, vol. 12, no. 1, pp. 219–245, Jan. 2000.

[23] V. R. Konda and J. N. Tsitsiklis, "OnActor-Critic algorithms," *SIAM Journal on Control and Optimization*, vol. 42, no. 4, pp. 1143–1166, Jan. 2003.

[24] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic Policy Gradient Algorithms," in *ICML*, Beijing, China, June 2014.

[25] S. Levine and P. Abbeel, "Learning neural network policies with guided policy search under unknown dynamics," in *Advances in Neural Information Processing Systems 27*, 2014, pp. 1071–1079.

[26] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End–to–end training of deep visuomotor policies," *Journal of Machine Learning Research: JMLR*, vol. 17, no. 1, pp. 1334–1373, 2016.

[27] V. Kumar, E. Todorov, and S. Levine, "Optimal control with learned local models: Application to dexterous manipulation," in *IEEE International Conference on Robotics and Automation (ICRA)*, May 2016, pp. 378–383.

[28] T. Erez and E. Todorov, "Trajectory optimization for domains with contacts using inverse dynamics," in *IEEE International Conference on Intelligent Robots and Systems*, Oct. 2012, pp. 4914–4919.

[29] I. Mordatch, K. Lowrey, and E. Todorov, "Ensemble-CIO: Full-body dynamic motion planning that transfers to physical humanoids," in *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, 2015, pp. 5307–5314.