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

Bora Banjanin, Samuel A. Burden
Department of Electrical Engineering
University of Washington, Seattle, WA, USA
borab,sburden@uw.edu

Abstract: State–of–the–art approaches to reinforcement learning for 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 smooth function approximators and gradient–based algorithms for contact–rich robot dynamics arising in locomotion and manipulation.

Keywords: locomotion, manipulation, piecewise–differentiability

1 Introduction and background on 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* $u \in \mathcal{U}$:

$$\nu(x) = \min_{u \in \mathcal{U}} c(x, u); \tag{1}$$

so long as $\mathcal{X}$ and $\mathcal{U}$ are compact and c is continuous, the function $\nu : \mathcal{X} \to \mathbb{R}$ indicated in (1), termed the *optimal value function*, is well–defined. We let $\pi : \mathcal{X} \to \mathcal{U}$ denote an *optimal policy* for (1), i.e.

$$\forall x \in \mathcal{X} : \pi(x) \in \arg\min_{u \in \mathcal{U}} c(x, u) \tag{2}$$

or, equivalently,

$$\forall x \in \mathcal{X} : \nu(x) = c(x, \pi(x)). \tag{3}$$

In this paper we study how regularity properties (continuity, differentiability) of the cost function (c) relate to regularity properties of optimal value (ν) and policy (π) functions, and apply these results to optimal control problems involving contact–rich robot dynamics.

Organization: the remainder of this section concerns regularity properties of optimal value and policy functions; Section 2 describes a class of models for contact–rich robot dynamics; Section 3 presents optimal value and policy functions in a simple mechanical system; Section 4 discusses generality and implications of these results.

1.1 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.

1.2 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.

1st Conference on Robot Learning (CoRL 2017), Mountain View, United States.

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 [23, Ch. 1.1.1]

$$D_2 c(x, \pi(x)) = 0. \tag{4}$$

If c is two times continuously–differentiable (denoted $c \in C^2$) and the second–order sufficient condition [23, Ch. 1.1.2] for strict local optimality for (1) is satisfied at $\pi(x) \in \mathcal{U}$,

$$D_2^2 c(x, \pi(x)) > 0, \tag{5}$$

then the C^1 Implicit Function Theorem (IFT) [16, Thm. C.40] can be applied to (3) to choose π as a C^1 function near x. Note that IFT specifically required the invertibility tacit in (5):

$$\text{the linear function } D_2^2 c(x, \pi(x)) : T_u\mathcal{U} \to T_u\mathcal{U} \text{ is invertible.} \tag{6}$$

If (4) and (5) are satisfied, then applying the C^1 Chain Rule [16, Prop. C.3] to (4) yields

$$D\pi(x) = -D_2^2 c(x, \pi(x))^{-1} \left(D_{12} c(x, \pi(x))\right), \tag{7}$$

and applying the C^1 Chain Rule to (3) yields

$$D\nu(x) = D_x c(x, \pi(x)) \quad = D_1 c(x, \pi(x)) + D_2 c(x, \pi(x)) D\pi(x), \tag{8}$$

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 (4) and second–order sufficient (5), (6) conditions for optimality and stability of solutions to (1) 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 (7), (8).

Proposition 1 *If $c \in C^2(\mathcal{X} \times \mathcal{U}, \mathbb{R})$ satisfies (4), (5), and (6) 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 $\pi \in C^1(X, U)$ such that $\pi(\xi) = \mu$ and*

$$\forall x \in X : \pi(x) \text{ is the unique minimizer for } \nu(x) = \min_{u \in U} c(x, u); \tag{9}$$

the derivative of π is given by (7), and the derivative of ν is given by (8).

1.3 Piecewise–differentiable cost functions

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

$$\forall w \in T_u\mathcal{U} : D_2 c(x, \pi(x); w) \geq 0. \tag{10}$$

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 [26, Ch. 3] that exists by virtue of the cost being PC^1 [26, 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 [7, Thm. 1] for strict local optimality for (1) is satisfied at $\pi(x) \in \mathcal{U}$,

$$\forall w \in \{w \in T_u\mathcal{U} \mid w \neq 0,\ D_2 c(x, \pi(x); w) = 0\} : D_2^2 c(x, \pi(x); w, w) > 0, \tag{11}$$

and if

$$\text{the piecewise–linear function } D_2^2 c(x, \pi(x)) : T_u\mathcal{U} \to T_u\mathcal{U} \text{ is invertible,} \tag{12}$$

then a PC^1 Implicit Function Theorem can be applied to choose π as a PC^1 function near x [25, Cor. 3.4].[2] Applying the PC^1 Chain Rule [26, Thm. 3.1.1] to (10) yields (cf. [25, § 3])

$$\forall v \in T_x\mathcal{X} : D\pi(x; v) = -D_2^2 c(x, \pi(x))^{-1} \left(D_{12} c(x, \pi(x); v)\right), \tag{13}$$

[1] We use the notion of piecewise–differentiability from [26, Ch. 4.1]: a function is piecewise–differentiable if it is everywhere locally a continuous selection of a finite number of continuously–differentiable functions.

[2] This Implicit Function Theorem requires $D_2 c$ be *strongly* B–differentiable; the costs considered below are not generally strongly B–differentiable, but they are generally PC^r–equivalent to strongly B–differentiable functions [15, Thm. 3.1], whence [25, Cor. 3.4] can be applied indirectly.

and applying the PC^1 Chain Rule to (3) yields

$$\forall v \in T_x\mathcal{X} : D\nu(x;v) = D_x c(x,\pi(x);v) \quad = D_1 c(x,\pi(x);v) + D_2 c(x,\pi(x);D\pi(x;v)), \quad (14)$$

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

We conclude that if the cost function is two times piecewise–differentiable ($c \in PC^2$) and first–order necessary (10) and second–order sufficient (11), (12) conditions for optimality and stability of solutions to (1) 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 (13), (14).

Proposition 2 *If $c \in PC^2(\mathcal{X} \times \mathcal{U}, \mathbb{R})$ satisfies (10), (11), and (12) 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 $\pi \in PC^1(X, U)$ such that $\pi(\xi) = \mu$ and*

$$\forall x \in X : \pi(x) \text{ is the unique minimizer for } \nu(x) = \min_{u \in U} c(x,u); \quad (15)$$

the B–derivative of π is given by (13), and the B–derivative of ν is given by (14).

1.4 Conclusions regarding regularity of optimal value and policy functions

The results in Sections 1.1, 1.2, and 1.3 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 3 we provide instances of the class of models described in Section 2 that exhibit these effects.

2 Mechanical systems subject to unilateral constraints

In this section, we formalize a class of models relevant for contact–rich dynamics in robot locomotion and manipulation as *mechanical systems subject to unilateral constraints* and apply results from the preceding section to optimal control problems involving these models.

2.1 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.

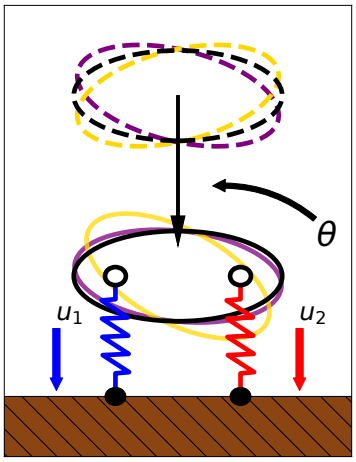
It is well–known (cf. [3, Sec. 3] or [11, Sec. 2.4, 2.5]) that, with $J = \{j \in \{1, \ldots, n\} : a_j(q) = 0\}$ denoting the *contact mode*, the system's dynamics take the form

$$M(q)\ddot{q} = f_J(q,\dot{q},u) + c(q,\dot{q})\dot{q} + Da_J(q)^\top \lambda_J(q,\dot{q},u), \quad (16a)$$

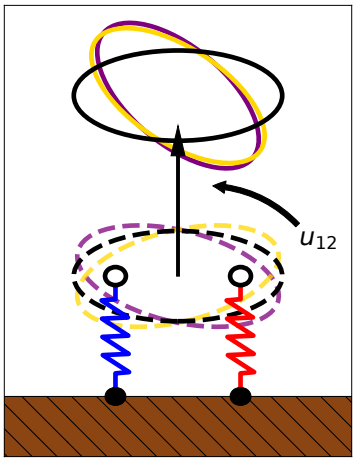
$$\dot{q}^+ = \Delta_J(q,\dot{q}^-), \quad (16b)$$

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* [3] and specifies[3] 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.

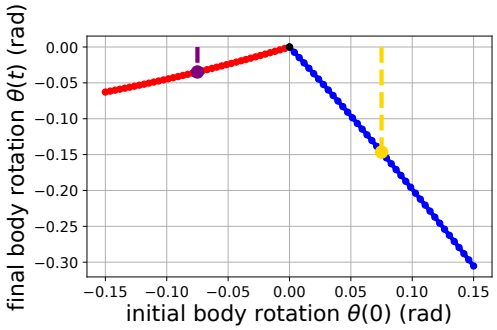
[3] We let $TQ = \mathbb{R}^d \times \mathbb{R}^d$ denote the *tangent bundle* of the configuration space Q; an element $(q, \dot{q}) \in TQ$ can be regarded as a pair containing generalized configurations $q \in \mathbb{R}^d$ and velocities $\dot{q} \in \mathbb{R}^d$; we write $\dot{q} \in T_q Q$.



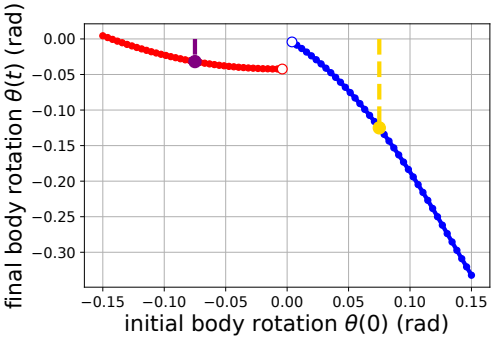
(a) *touchdown* maneuver illustration



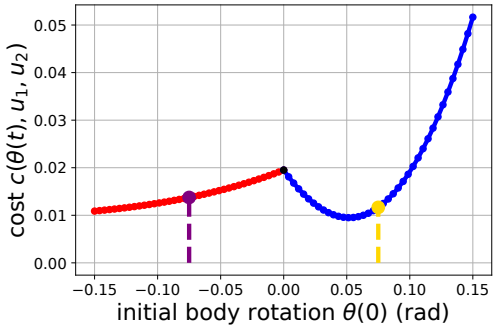
(b) *liftoff* maneuver illustration



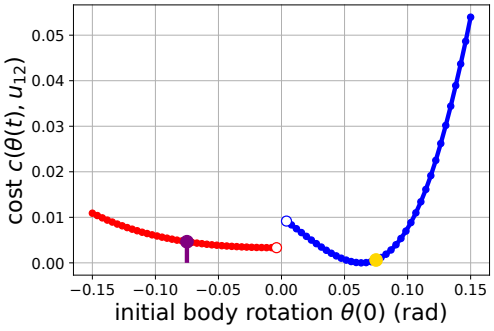
(c) touchdown trajectory outcomes



(d) liftoff trajectory outcomes

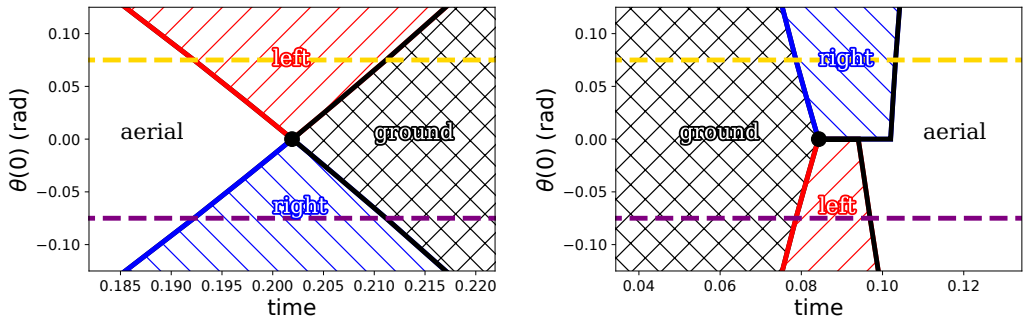


(e) touchdown value



(f) liftoff value

Figure 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

Figure 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 (16). (a,b) System contact mode at each time t for a given initial body rotation $\theta(0)$; the body torque is zero and the leg forces are different 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).

2.2 Regularity of dynamics

The seemingly benign equations in (16) can yield dynamics with a range of regularity properties. This issue has been thoroughly investigated elsewhere [3, 21, 22]; 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. (16) or (17)) 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 (16) are continuously–differentiable ($M, f, a, \gamma \in C^r$); however, as illustrated by [3, 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

$$\dot{x} = F(x, u) \tag{17}$$

yields unique trajectories whose regularity matches the vector field's: if F is continuously differentiable, then there exists a flow for (17) that is continuously differentiable to the same order.

To ensure trajectories for (16) exist uniquely, restrictions must be imposed; we refer the interested reader to [3, Thm. 10] for a specific result and [11] for a general discussion of this issue. Since we are chiefly concerned with how properties of the dynamics in (16) affect properties of optimal value and policy functions, we will assume in what follows that conditions have been imposed to ensure (16) has a flow for states, inputs, and time horizons of interest.

Assuming that a flow ϕ exists for (16) 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 [24, Table 3] [10]; 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 [22,

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*[4] and *Zeno*[5] 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 4.1 for a discussion of when this phenomena prevails).

2.3 Regularity of optimal value and policy functions

Returning to minimization of a cost function as in (1), we now make explicit the dependence of the cost on the robot's dynamics in terms of *final* ($\ell : \mathcal{X} \to \mathbb{R}$) and *running* ($\mathcal{L} : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$) costs:

$$\nu(x) = \min_{u \in \mathcal{U}^{[0,t]}} \ell(x^u(t)) + \int_0^t \mathcal{L}(x^u(s), u(s))\, ds, \tag{18}$$

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 $s \in [0,t]$. Since we seek to expose the dependence of the cost in (18) on the flow ϕ, we transcribe the problem in (18) to a simpler form using a standard state augmentation technique (cf. [23, Ch. 4.1.2]):

$$\nu(x) = \min_{u \in \mathcal{U}^{[0,t]}} c\left(\phi(t, x, u)\right). \tag{19}$$

We focus on how regularity of the flow ϕ affects regularity of the optimal value ν and policy π.

Based on the observations in Section 1, we expect the regularity of the optimal value and policy functions for (19) to be related to the regularity of the cost function in (19), which is a composite of a continuously–differentiable function c and a flow function ϕ. As discussed in Section 2.2, the regularity of ϕ is partly determined by the robot's design; it is possible for ϕ to be discontinuous ($\phi \notin C^0$), continuously–differentiable ($\phi \in C^r$), or piecewise–differentiable (and not continuously–differentiable, $\phi \in PC^r \setminus C^r$). Thus regularity of the composition $c \circ \phi$ is generally determined by regularity of ϕ. We conclude that nonsmooth dynamics should yield nonsmooth optimal value and policy functions in mechanical systems subject to unilateral constraints.

3 Optimal value and policy functions for a mechanical system subject to unilateral constraints

We conjectured in the previous section that optimal value and policy functions for contact–rich robot dynamics are generally nonsmooth. To investigate this conjecture, 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 we compute 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.

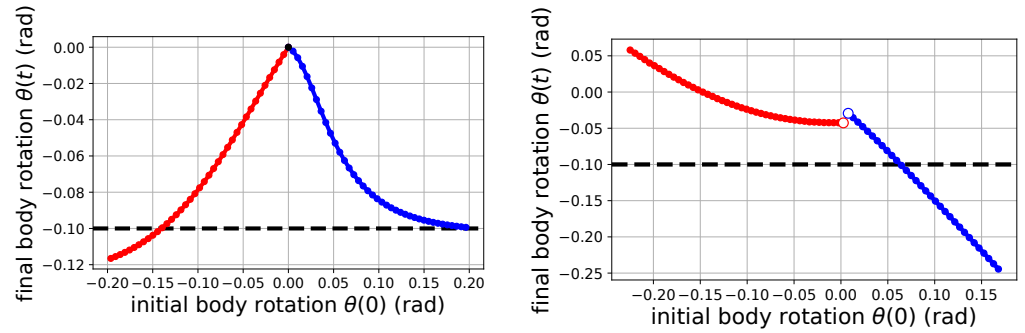
[4] where a constraint function a_j decreases to and then increases from zero without activating constraint j

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

[6] using the modeling framework from [11] and simulation algorithm from [6]

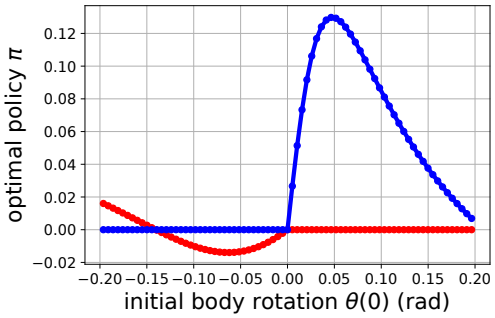
[7] `SciPy v0.19.0 minimize_scalar`

[8] We plan to release the software used to generate these results as an `environment` in `OpenAI Gym` [5].

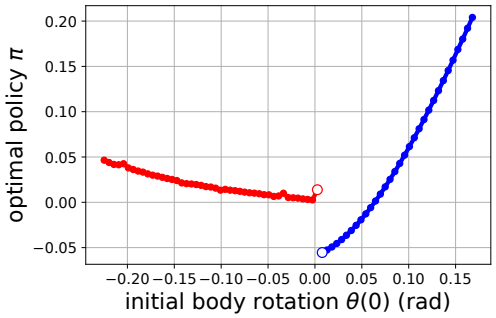


(a) optimal touchdown trajectory outcomes

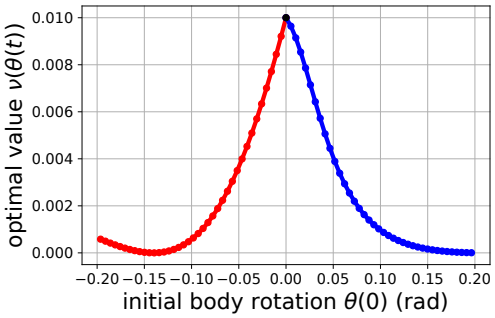
(b) optimal liftoff trajectory outcomes



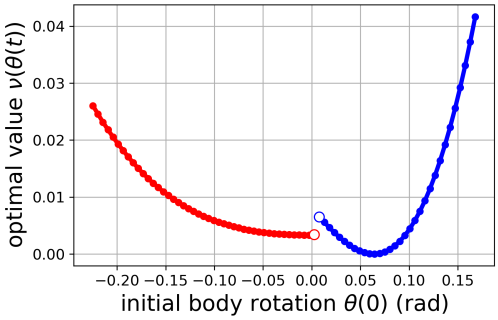
(c) optimal touchdown policy



(d) optimal liftoff policy



(e) optimal touchdown value



(f) optimal liftoff value

Figure 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).

4 Discussion

We conclude by discussing how often we expect to encounter the nonsmooth phenomena described above in models of robot behaviors (Section 4.1) and what our results imply about the use of smooth tools in this nonsmooth setting (Section 4.2).

4.1 Prevalence of nonsmooth phenomena near behaviors of interest

In Section 3, 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 manifested 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 [16, Ch. 6], then the nonsmoothness presented in Section 3 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 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 [2] 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 [19]. 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.

4.2 Function approximation and gradient–based algorithms

It is common practice to use smooth functions to approximate optimal value and policy functions. This practice is justified for finite–state Markov Decision Processes (MDP) and smooth control systems whose value functions (optimal or otherwise) are known to be smooth under mild regularity conditions. Our results imply that this practice is not generally justified for optimal control of mechanical systems subject to unilateral constraints since the value functions are generally discontinuous or only piecewise–differentiable. Future work could investigate the use of nonsmooth function approximators in this setting.

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 [4, 29], and can be interpreted as a *natural* [12] or *trust region* [27] 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 [8, 13] or sampling [4, 28]. Analogously to the conclusions in the preceding paragraph, this practice is justified for classical systems (finite–state MDP, smooth control systems) whose value functions are smooth; it is not generally justified for the systems considered here since the value of (optimal or non–optimal) policies can be nonsmooth. It should be possible in future work to apply nonsmooth analogues of steepest descent (22) to these systems when the value function is piecewise–differentiable so that $D_\pi \nu^\pi$ is piecewise–linear.

Recent work employs smooth approximations of the contact–rich robot dynamics in (16) to enable application of gradient–based learning [14, 17, 18] and optimization [9, 19, 20] algorithms. This approach leverages established scalable algorithms, but does not guarantee that policies optimized for the smoothed dynamics will be (near–)optimal when applied to the original nonsmooth dynamics. As an alternative approach, the framework recently introduced in [21] provides conditions under which the dynamics in (16) yield trajectories that 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.

[9] This follows from [1, Eqn. 2.3] so long as the constraint (de)activations are *admissible* [22, Def. 3, Lem. 1].
[10] as measured by the number of constraints and/or constraint (de)activations

Acknowledgments

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] 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*, 11(4):385–398, 1958. ISSN 0033-5614. doi:10.1093/qjmam/11.4.385.

[2] R. M. Alexander. The gaits of bipedal and quadrupedal animals. *The International journal of robotics research*, 3(2):49–59, 1984. ISSN 0278-3649. doi:10.1177/027836498400300205.

[3] P. Ballard. The dynamics of discrete mechanical systems with perfect unilateral constraints. *Archive for rational mechanics and analysis*, 154(3):199–274, 2000. ISSN 0003-9527. doi:10.1007/s002050000105.

[4] J. Baxter and P. L. Bartlett. Infinite-horizon policy-gradient estimation. *The journal of artificial intelligence research*, 15:319–350, 2001. ISSN 1076-9757.

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

[6] 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*, 60(9):2307–2320, 2015. ISSN 0018-9286. doi:10.1109/TAC.2015.2404231.

[7] R. W. Chaney. Second-Order sufficient conditions in nonsmooth optimization. *Mathematics of Operations Research*, 13(4):660–673, 1 Nov. 1988. ISSN 0364-765X. doi:10.1287/moor.13.4.660.

[8] K. Doya. Reinforcement learning in continuous time and space. *Neural computation*, 12(1):219–245, Jan. 2000. ISSN 0899-7667.

[9] T. Erez and E. Todorov. Trajectory optimization for domains with contacts using inverse dynamics. In *IEEE International Conference on Intelligent Robots and Systems*, pages 4914–4919, Oct. 2012. doi:10.1109/IROS.2012.6386181.

[10] Y. Hürmüzlü and D. B. Marghitu. Rigid body collisions of planar kinematic chains with multiple contact points. *The International journal of robotics research*, 13(1):82–92, 1994. ISSN 0278-3649. doi:10.1177/027836499401300106.

[11] 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*, 35(11):1354–1392, 1 Sept. 2016. ISSN 0278-3649, 1741-3176. doi:10.1177/0278364916639380.

[12] S. Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14:1531–1538, 2001. ISSN 1049-5258.

[13] V. R. Konda and J. N. Tsitsiklis. OnActor-Critic algorithms. *SIAM Journal on Control and Optimization*, 42(4):1143–1166, Jan. 2003. ISSN 0363-0129, 1095-7138. doi:10.1137/S0363012901385691.

[14] 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)*, pages 378–383, May 2016. doi:10.1109/ICRA.2016.7487156.

[15] L. Kuntz and S. Scholtes. Structural analysis of nonsmooth mappings, inverse functions, and metric projections. *Journal of mathematical analysis and applications*, 188(2):346–386, 1994. ISSN 0022-247X. doi:10.1006/jmaa.1994.1431.

[16] J. M. Lee. *Introduction to smooth manifolds*. Graduate texts in mathematics. Springer, New York ; London, 2nd ed. edition, 2012. ISBN 9781441999825.

[17] S. Levine and P. Abbeel. Learning neural network policies with guided policy search under unknown dynamics. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1071–1079. Curran Associates, Inc., 2014.

[18] S. Levine, C. Finn, T. Darrell, and P. Abbeel. End–to–end training of deep visuomotor policies. *Journal of machine learning research: JMLR*, 17(1):1334–1373, 2016. ISSN 1532-4435.

[19] I. Mordatch, E. Todorov, and Z. Popović. Discovery of complex behaviors through contact-invariant optimization. *ACM transactions on graphics*, 31(4):43:1–43:8, July 2012. ISSN 0730-0301. doi:10.1145/2185520.2185539.

[20] 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*, pages 5307–5314, 2015. doi:10.1109/IROS.2015.7354126.

[21] A. M. Pace and S. A. Burden. 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.

[22] 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.

[23] E. Polak. *Optimization: algorithms and consistent approximations*. Springer–Verlag, 1997.

[24] C. D. Remy, K. Buffinton, and R. Siegwart. Stability analysis of passive dynamic walking of quadrupeds. *The International journal of robotics research*, 29(9):1173–1185, 2010. ISSN 0278-3649. doi:10.1177/0278364909344635.

[25] S. M. Robinson. An Implicit-Function theorem for a class of nonsmooth functions. *Mathematics of Operations Research*, 16(2):292–309, 1991. ISSN 0364-765X.

[26] S. Scholtes. *Introduction to piecewise differentiable equations*. Springer–Verlag, 2012. doi:10.1007/978-1-4614-4340-7.

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

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

[29] 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*, 12:1057–1063, 2000. ISSN 1049-5258.