IEEE Control Systems Letters paper presented at
2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Differentiable Sparse Optimal Control

Ryotaro Shima[1], Ryuta Moriyasu[1], Sho Kawaguchi[2] and Kenji Kashima[3]

***Abstract*— This paper develops a framework for differentiating sparse optimal control inputs with respect to cost parameters. The difficulty lies in the non-smoothness induced by a sparsity-enhancing regularizer. To avoid this, we identify the optimal inputs as a unique zero point of a function using the proximal technique. This enables us to characterize the differentiability and employ the implicit function theorem. We also demonstrate the effectiveness of our approach using a numerical example of inverse optimal control.**

I. Introduction

Sparse control, in which smaller supports for control inputs are preferred, has many industrial applications, e.g., engine idling in hybrid vehicles and battery charging. However, it is often implemented using if-then rules and other heuristic methods, which are difficult to interpret. One remedy to this is to identify a control problem for which the manually-tuned inputs are optimal. Such techniques, called inverse optimal control[1] (IOC) in the control field and inverse reinforcement learning (IRL) in the machine learning field, require a sensitivity of optimal inputs with respect to (wrt.) cost parameters. Conventional frameworks apply the implicit function theorem to the first-order optimality equation, such as Karush-Kuhn-Tucker (KKT) condition [1], [2] or Euler-Lagrange equation [3].

However, sparse optimal control enhances sparsity via a nonsmooth penalty, e.g., ℓ_1 norm [4]. Thus, conventional IOC/IRL frameworks cannot be applied—the nonsmoothness changes the first-order optimality condition from an equation to an inclusion relation, as discussed in Section II. A widely used approach to address the nonsmoothness of sparsity involves using a proximal operator [4]. [5] conducted sensitivity analysis using a fixed point equation of proximal gradient descent but is limited to unconstrained optimization and its extension to constrained problem is nontrivial.

This paper develops a framework for differentiating sparse optimal control wrt. the parameters of the cost function. To overcome the difficulty of nonsmoothness induced by the sparsity enhancing regularizer, we focus on the fact that the proximal operator transforms an inclusion relation into an equation. This technique enables us to characterize the sparse optimal input as a unique zero point of a function. Additionally, we conclude that the function satisfies the nonsingularity assumption required for the implicit function theorem under a practical constraint qualification and a convexity assumption.

The remainder of this paper is organized as follows. Section II presents problem setting of this paper and preliminary contents related to nonsmooth optimization. Section III presents our framework for implicit differentiation of sparse optimal control. We also provide sufficient conditions to ensure differentiability. Section IV provides certain implications of the sufficient conditions. Section V demonstrates a numerical example of IOC which verifies the effectiveness of our approach.

A. Notation

Let us denote the set of all positive real numbers by $\mathbb{R}_>$; the i-th element of a vector $v \in \mathbb{R}^n$ by v_i; the (i,j)-th element of a matrix $A \in \mathbb{R}^{n\times m}$ by A_{ij}; the transpose of A by A^T; Hadamard product by $\odot$; the identity function by id; and the ℓ_1 norm of $x \in \mathbb{R}^n$ by $\|x\|_1 := \sum_{i=1}^n |x_i|$. Moreover, let the diagonal matrix whose i-th element is a_i be denoted by $\mathrm{diag}_i(a_i)$ and a soft threshold function with threshold a be $s_a(v) := \mathrm{sgn}(v)\max\{|v|-a, 0\}$. Let $\mathbb{Z}_n := \{1,\dots,n\} \subset \mathbb{Z}$. We denote $[a_1^\mathsf{T} \ \cdots \ a_k^\mathsf{T}]^\mathsf{T}$ by $[a_1;\cdots;a_k]$. S_a acts on $v \in \mathbb{R}^n$ via $S_a(v) = [s_a(v_1);\cdots;s_a(v_n)]$. For $I \subset \mathbb{Z}_m$, $J \subset \mathbb{Z}_n$, $A \in \mathbb{R}^{n\times m}$ and $v \in \mathbb{R}^n$, we denote the sub-matrix consisting of $A_{ij}, i \in I, j \in J$ by A_{IJ} and the sub-vector consisting of $v_i, i \in J$ by v_J. Given the partitions, $\mathbb{Z}_m = I_1 + \cdots + I_k$ and $\mathbb{Z}_n = J_1 + \cdots + J_l$, we express A and v as follows:

$$A = \begin{bmatrix} A_{I_1J_1} & \cdots & A_{I_1J_l} \\ \vdots & \ddots & \vdots \\ A_{I_kJ_1} & \cdots & A_{I_kJ_l} \end{bmatrix} = \begin{bmatrix} A_{:J_1} \\ \vdots \\ A_{:J_l} \end{bmatrix}, \ v = \begin{bmatrix} v_{I_1} \\ \vdots \\ v_{I_k} \end{bmatrix}.$$

II. Preliminaries

A. Sparse optimal control and its sensitivity

This paper considers a discrete-time control system represented by the following state space model:

$$x^{(t+1)} = f(x^{(t)}, u^{(t)}), \quad y^{(t)} = h(x^{(t)}, u^{(t)}), \quad x^{(0)} = x_0 \qquad (1)$$

where $x^{(t)} \in \mathbb{R}^n$ denotes a state, $u^{(t)} \in \mathbb{R}^r$ denotes an input, $y^{(t)} \in \mathbb{R}^m$ denotes an output, $t \in \{0,1,\dots,T-1\}$, $f(x,u) \in \mathbb{R}^n$, and $h(x,u) \in \mathbb{R}^r$. Let $U = [u^{(0)};\cdots;u^{(T-1)}] \in \mathbb{R}^N, Y = [y^{(0)};\cdots;y^{(T-1)}] \in \mathbb{R}^{Tm}$, and $N := Tr$. The model (1) can then be rewritten as $Y = M(U, x_0)$ where M consists of recurrent composition of f and h. We consider the following n_e equality constraints and n_i inequality constraints:

$$LU + b = 0, \quad \hat{g}(U, M(U, x_0)) \le 0,$$

[1]Ryotaro Shima and Ryuta Moriyasu are with Toyota Central R&D Labs., Inc., 41-1, Yokomichi, Nagakute, Aichi, Japan {ryotaro.shima, moriyasu}@mosk.tytlabs.co.jp

[2]Sho Kawaguchi is with Toyota Industries Corporation, 3, Hama-cho, Hekinan, Aichi, Japan sho.kawaguchi@mail.toyota-shokki.co.jp

[3]Kenji Kashima is with Kyoto University, Yoshida-Honmachi, Sakyo-ku, Kyoto City, Japan kk@i.kyoto-u.ac.jp

[1]In the nonlinear control theory, IOC represents designing the controller such that it corresponds to a solution of some optimal control problem. This paper does not use the term IOC in this sense.

where $L \in \mathbb{R}^{n_e \times N}$, $b \in \mathbb{R}^{n_e}$, and $\hat{g}(U,Y) \in \mathbb{R}^{n_i}$. We denote the cost function by $\hat{\ell}(U,Y,\theta) \in \mathbb{R}$ with a parameter $\theta \in \mathbb{R}^p$. The sparse optimal control problem is given as follows:

$$U^*(\theta,x_0) = \arg\min_{U \in \mathscr{C}(x_0)} \ \ell(U,\theta,x_0) + \lambda r(U), \tag{2}$$

$$\mathscr{C}(x_0) := \left\{ U \in \mathbb{R}^N \mid LU + b = 0,\ g(U,x_0) \le 0 \right\}, \tag{3}$$

$$\ell(U,\theta,x_0) := \hat{\ell}(U, M(U,x_0), \theta), \tag{4}$$

$$g(U,x_0) := \hat{g}(U, M(U,x_0)), \tag{5}$$

where $r(U) := \|U\|_1$ and $\lambda \in \mathbb{R}_>$. We assume that f, h, $\hat{\ell}$, and $\hat{g}$ are of class C^2. *The aim of this paper is to calculate the derivative of $U^*(\theta)$ wrt. θ.*

Example 1 (inverse optimal control): Consider N_d sets of initial values $\{x_0^d\}_{d=1}^{N_\mathrm{d}}$ and expert's inputs $\{U_d^\mathrm{e}\}_{d=1}^{N_\mathrm{d}}$. The replication of the expert's inputs using the sparse optimal control (2) is then formulated as follows:

$$\theta^* = \arg\min_\theta \sum_{d=1}^{N_\mathrm{d}} \|U^*(\theta, x_0^d) - U_d^\mathrm{e}\|^2. \tag{6}$$

Gradient-based algorithms, such as stochastic gradient descent, require $\frac{\partial U^*}{\partial \theta}(\theta)$ as well as $U^*(\theta)$. ◁

Remark 1: Because the dependency on x_0 is irrelevant, we omit x_0 from arguments such as $U^*(\theta)$, $\mathscr{C}$, and $g(U)$ if not confusing. Although only the cost function $\hat{\ell}$ depends on θ in (2)–(5), one can trivially extend our method to make f, h, L, b, $\hat{g}$, and λ dependent on θ, as in [3]. ◁

B. KKT condition for sparse optimal control problem

Because $r(U)$ is nondifferentiable, we extend its gradient to a *subgradient* [6], which is a set-valued function defined as $\partial r(U) := \left\{ g \in \mathbb{R}^N \mid r(V) \ge g^\mathsf{T}(V-U),\ \forall V \in \mathbb{R}^N \right\}$. Under a certain constraint qualification, such as the linear independence constraint qualification (LICQ) [7], the KKT condition states that U^* satisfies

$$0 \in \nabla_U \ell(U^*,\theta) + \lambda \partial r(U^*) + z^*, \tag{7}$$

$$z^* = \left(\frac{\partial g}{\partial U}(U^*) \right)^\mathsf{T} \eta^* + L^\mathsf{T} \nu^*,\ LU^* + b = 0, \tag{8}$$

$$g(U^*) \odot \eta^* = 0,\ g(U^*) \le 0,\ \eta^* \ge 0 \tag{9}$$

for some $z^* \in \mathbb{R}^N$, $\eta^* \in \mathbb{R}^{n_i}$, and $\nu^* \in \mathbb{R}^{n_e}$. We emphasize that the condition (7) is *an inclusion relation, and not an equation*. If $r(U)$ is differentiable at U^*, then $\partial r(U^*) = \{\nabla r(U^*)\}$ and the relation (7) can be expressed by an equation. Unfortunately, this is not the case because U^* contains many zeros in sparse optimal control for any θ. This implies conventional IOC frameworks are not applicable to sparse optimal control because they apply the implicit function theorem to the KKT *equation*.

C. Proximal technique [6]

Consider a convex function $f : \mathbb{R}^n \to \mathbb{R}$ and let $\gamma \in \mathbb{R}_>$. The proximal operator prox_γ^f is defined as follows:

$$\mathrm{prox}_\gamma^f(v) := \arg\min_{x \in \mathbb{R}^n} \frac{|x-v|^2}{2\gamma} + f(x) \tag{10}$$

When $f(x) = \|x\|_1$, the operator can be rewritten in closed form as $\mathrm{prox}_\gamma^{\|\cdot\|_1}(v) = S_\gamma(v)$.

Let $x^* := \mathrm{prox}_\gamma^f(v)$. Then, the optimality of x^* in (10) implies $0 \in x^* - v + \gamma \partial f(x^*)$, which reduces to

$$v \in (\mathrm{id} + \gamma \partial f)(x^*). \tag{11}$$

This implies that the inclusion relation (11) can be transformed into an equation $x^* = \mathrm{prox}_\gamma^f(v)$ using the proximal operator.

III. Proposed Framework

This section establishes a new framework for differentiating sparse optimal inputs. To apply the implicit function theorem, the inclusion relation (7) is transformed into an equation using the proximal technique. Sufficient conditions for differentiability are also provided.

A. Implicit differentiation of sparse optimal input

Let $\gamma \in \mathbb{R}_>$ and $d(U,z,\theta) := \nabla_U \ell(U,\theta) + z$. Using the proximal technique, we reduce the inclusion relation (7) to an equation as follows:

$$\begin{aligned} &-d(U^*,z^*,\theta) \in \lambda \partial r(U^*) \\ &\Leftrightarrow U^* - \gamma d(U^*,z^*,\theta) \in (\mathrm{id} + \gamma\lambda \partial r)(U^*) \\ &\Leftrightarrow U^* = S_{\gamma\lambda}(U^* - \gamma d(U^*,z^*,\theta)). \end{aligned}$$

A function $\varphi : \mathbb{R}^2 \to \mathbb{R}$ is called a complementarity function if $\varphi(a,b) = 0 \Leftrightarrow a \ge 0, b \ge 0, ab = 0$. A typical example is the Fischer–Burmeister function:

$$\varphi_\mathrm{fb}(a,b) := a + b - \sqrt{a^2 + b^2}. \tag{12}$$

We transform the complementarity condition (9) using a complementarity function φ as follows:

$$\begin{aligned} \phi(g(U^*), \eta^*) &= 0, \\ \phi(g,\eta) &:= [\varphi(-g_1,\eta_1); \cdots; \varphi(-g_{n_i}, \eta_{n_i})]. \end{aligned}$$

Let $\chi := [U; z; \eta; \nu] \in \mathbb{R}^K$ where $K := 2N + n_i + n_e$. Finally, we obtain a system of K equations which determines χ^* for each θ:

$$F(\chi^*, \theta) = 0, \tag{13}$$

$$F(\chi,\theta) := \begin{bmatrix} U - S_{\gamma\lambda}(U - \gamma d(U,z,\theta)) \\ -z + \left(\frac{\partial g}{\partial U}(U) \right)^\mathsf{T} \eta + A^\mathsf{T} \nu \\ \phi(g(U),\eta) \\ LU + b \end{bmatrix} \in \mathbb{R}^K. \tag{14}$$

Now we apply the implicit function theorem and bridge the derivative of χ^* wrt. θ and that of F wrt. (χ,θ). Because F contains nondifferentiable points, we introduce the concept of the *subderivative*, i.e., a set-valued extension of the derivative. A subderivative is constructed in several ways. In this paper, $\partial F \subset \mathbb{R}^{K \times (K+p)}$ denotes the *Clarke Jacobian* of F at (χ^*,θ), which is a widely used subderivative; see [8] for its definition and Remark 3 for an explicit expression of ∂F in a special case. If F is continuously differentiable

at (χ^*,θ), its Clarke Jacobian ∂F is a singleton consisting of its ordinary derivative:

$$\partial F = \left\{ \left[\frac{\partial F}{\partial \chi}(\chi^*,\theta) \;\; \frac{\partial F}{\partial \theta}(\chi^*,\theta) \right] \right\}.$$

Assumption 1: For every element $[J^\chi \; J^\theta] \in \partial F$, the matrix $J^\chi \in \mathbb{R}^{K\times K}$ is nonsingular. ◁

Proposition 1: Consider the sparse optimal control problem (2). Suppose Assumption 1 holds. Then, the matrix set

$$\partial \chi^* := \left\{ -(J^\chi)^{-1} J^\theta \in \mathbb{R}^{K\times p} \;\middle|\; [J^\chi \; J^\theta] \in \partial F \right\} \tag{15}$$

is a subderivative of $\chi^*(\theta)$ wrt. θ. ◁

Proof: Corollary 1 in [9] directly implies the claim. ■

More specifically, the subderivative in (15) is the so-called *conservative Jacobian*, which is applicable to stochastic gradient descent (see Theorem 3 in [9]).

Remark 2: Even if U^* is sparse, we can show that $\partial\chi^*$ in (15) is a singleton except in the following special cases:

- The inequality constraints do not satisfy strict complementarity, i.e., $(g_i(U^*),\eta_i^*) = (0,0)$ for some $i \in \mathbb{Z}_N$.
- The support of U^* changes around θ, i.e., $d_i(U^*,z^*,\theta) = \pm\lambda$ for some $i \in \mathbb{Z}_N$.

It should be remarked that our method is useful not only in the aforementioned special cases but also whenever U^* is sparse. More specifically, even if $\partial\chi^*$ is a singleton, the KKT condition is still an inclusion relation, and consequently, conventional IOC/IRL methods are not applicable. ◁

B. Sufficient conditions for differentiability

This subsection provides sufficient conditions for the non-singularity assumption. The proof of the theorem is presented in Appendix I.

Assumption 2: ℓ is strongly convex[2] wrt. U and g is convex wrt. U. ◁

Assumption 3: Denote the index set of active constraints by $\mathscr{I} := \{i \in \mathbb{Z}_{n_i} \mid g_i(U^*) = 0\}$ and the support of U^* by $\mathscr{S} := \{j \in \mathbb{Z}_N \mid U_j^* \neq 0\}$. Then, a matrix $\left[(J_g)_{\mathscr{I}\mathscr{S}}; L_{:\mathscr{S}}\right]$ is column full rank where $J_g := \frac{\partial g}{\partial U}(U^*)$. ◁

Theorem 1: Consider the sparse optimal control problem (2) under Assumptions 2 and 3. Then, χ^* satisfying (13) is unique. Furthermore, if φ is the Fischer–Burmeister function φ_{fb} defined as (12), then Assumption 1 holds true. ◁

Remark 3: We denote the Clarke Jacobian of $S_{\gamma\lambda}(\cdot)$ at $\xi := U^* - \gamma d(U^*,z^*,\theta)$ by ∂S and that of ϕ at $(g(U^*),\eta^*)$ by $\partial\phi$. If φ is φ_{fb}, as in Theorem 1, $[J^\chi \; J^\theta] \in \partial F$ is fully parameterized by $D \in \partial S$ and $[\phi^g \; \phi^\eta] \in \partial\phi$ as follows:

$$J^\chi = \begin{bmatrix} I - D + \gamma D H^\ell & \gamma D & 0 & 0 \\ H^g & -I & J_g^\mathsf{T} & L^\mathsf{T} \\ \phi^g J_g & 0 & \phi^\eta & 0 \\ L & 0 & 0 & 0 \end{bmatrix}, \tag{16}$$

$$J^\theta = \begin{bmatrix} \gamma D \dfrac{\partial(\nabla_U \ell)}{\partial\theta};\; 0 \;;\; 0 \;;\; 0 \end{bmatrix} \tag{17}$$

[2]The function $f:\mathbb{R}^n \to \mathbb{R}$ is strongly convex if, for some $\varepsilon \in \mathbb{R}_>$, $f(x) - \varepsilon\|x\|_2^2$ is convex wrt. x.

Algorithm 1 Calculate $U^*(\theta)$ and $\frac{\partial U^*}{\partial\theta}(\theta)$

Require: x_0, θ

1: obtain χ^* satisfying (13)
2: select $D \in \partial S, [\phi^g \; \phi^\eta] \in \partial\phi$ and get J^χ, J^θ by (16)–(17)
3: calculate $J_{\chi^*} = -(J^\chi)^{-1} J^\theta$ and partition it as $J_{\chi^*} = [J_{U^*}; J_{z^*}; J_{\eta^*}; J_{\nu^*}]$ where $J_{U^*} \in \mathbb{R}^{N\times p}$
4: **return** U^*, J_{U^*}

where H^ℓ, H_i^g denote Hessian matrices of ℓ, g_i at (U^*,θ) wrt. U and $H^g := \sum_i \eta_i^* H_i^g$. Moreover, ∂S and $\partial\phi$ are expressed as follows:

$$\partial S = \left\{ \mathrm{diag}_i(v_i) \in \mathbb{R}^{N\times N} \middle| v_i \in \partial s(\xi_i) \right\},$$

$$\partial s(v) := \begin{cases} \{1\} & \text{if } |v| > \gamma\lambda, \\ \{0\} & \text{if } |v| < \gamma\lambda, \\ \{a \in \mathbb{R} | 0 \le a \le 1\} & \text{if } |v| = \gamma\lambda, \end{cases}$$

$$\partial\phi = \left\{ \begin{bmatrix} \mathrm{diag}_i(-a_i) \\ \mathrm{diag}_i(b_i) \end{bmatrix}^\mathsf{T} \middle| [a_i \; b_i] \in \partial\varphi_{\mathrm{fb}}(-g_i(U^*), \eta_i^*) \right\},$$

$$\partial\varphi_{\mathrm{fb}}(a,b) := \begin{cases} \{[0 \; 1]\} & \text{if } a > 0, b = 0, \\ \{[1 \; 0]\} & \text{if } a = 0, b > 0, \\ \mathscr{D} & \text{if } a = b = 0, \end{cases}$$

$$\mathscr{D} := \left\{ [a \; b] \in \mathbb{R}^{1\times 2} \middle| (a-1)^2 + (b-1)^2 \le 1 \right\}.$$

See Proposition 3.1 in [10] for the expression of $\partial\varphi_{\mathrm{fb}}$. ◁

C. Framework of differentiable sparse optimal control

Proposition 1 provides a method for obtaining the sensitivity of sparse optimal inputs. We propose Algorithm 1 to calculate the sparse optimal input $U^*(\theta)$ and its sensitivity $\frac{\partial U^*}{\partial\theta}(\theta)$. χ^* satisfying (13) can be determined numerically using the Newton-Raphson method [11] or by solving (2) using the alternating direction method of multipliers (ADMM) [4]. The computational burden of obtaining $(J^\chi)^{-1}$ can be reduced by eliminating z using the equation

$$z = \left(\frac{\partial g}{\partial U}(U) \right)^\mathsf{T} \eta + L^\mathsf{T} \nu.$$

The implications of Assumptions 2 and 3 are investigated in the following section.

Remark 4: A widely used alternative of $\phi(g(U^*),\eta^*) = 0$ is $g(U^*) \odot \eta^* = 0$ [2]. Let $\tilde{F}(\chi,\theta)$ denote the function obtained when $\phi(g(U),\eta)$ in (14) is altered by $g(U) \odot \eta$. Additionally, let the Clarke Jacobian of $\tilde{F}$ at (χ^*,θ) be denoted by $\partial\tilde{F}$. Then, $[\tilde{J}^\chi \; \tilde{J}^\theta] \in \partial\tilde{F}$ is fully parametrized by $D \in \partial S$, and the relation between $[J^\chi \; J^\theta] \in \partial F$ and $[\tilde{J}^\chi \; \tilde{J}^\theta] \in \partial\tilde{F}$ is given as follows:

$$\tilde{J}^\theta = J^\theta, \quad \tilde{J}^\chi = \mathrm{Diag}(I, I, \Lambda, I) J^\chi,$$

$$\Lambda_{ij} := \begin{cases} g_i(U^*) & \text{if } i = j,\; g_i(U^*) < 0,\; \eta_i^* = 0, \\ -\eta_i^* & \text{if } i = j,\; g_i(U^*) = 0,\; \eta_i^* > 0, \\ 0 & \text{otherwise}, \end{cases}$$

where $\mathrm{Diag}(I, I, \Lambda, I)$ denotes a block diagonal matrix whose diagonal block elements are $I \in \mathbb{R}^{N\times N}$, $I \in \mathbb{R}^{N\times N}$, $\Lambda \in \mathbb{R}^{n_i \times n_i}$, and $I \in \mathbb{R}^{n_e \times n_e}$ in order from the upper left to the lower right. Λ is nonsingular if and only if the inequality constraints satisfy strict complementarity. Thus, if $\tilde{F}(\chi^*, \theta) = 0$ is used instead of (13), the strict complementarity of inequality constraints is required for differentiability as well as the assumptions in Theorem 1. ◁

Remark 5: Smoothing via slack variables [12], [13] also enables us to differentiate sparse optimal control. However, this increases the number of variables by $3N$, which increases the computational load of finding the resulting KKT point and matrix inversion. ◁

IV. Implications of the Assumptions

This section discusses the implications of the previously introduced sufficient conditions for differentiability.

A. Convexity of optimal control problem

This subsection provides two situations in which ℓ and g satisfy Assumption 2.

1) Linear plant case: Suppose that the plant model is linear, i.e., represented by $f(x,u) = A_\mathrm{p}x + B_\mathrm{p}u$ and $h(x,u) = C_\mathrm{p}x + D_\mathrm{p}u$ in (1). In this case, the function M is linear, i.e., $M(U, x_0) = \bar{A}_\mathrm{p}x_0 + \bar{B}_\mathrm{p}U$ for appropriate matrices, $\bar{A}_\mathrm{p} \in \mathbb{R}^{Tm\times n}$ and $\bar{B}_\mathrm{p} \in \mathbb{R}^{Tm\times N}$. Thus, Assumption 2 is fulfilled when the following two conditions hold true:

- $\hat{\ell}$ is convex wrt. (U, Y) and strongly convex wrt. U.
- $\hat{g}$ is convex wrt. (U, Y).

2) Nonlinear plant case: Even when the plant model is nonlinear, we can ensure convexity by adding a nondecreasing property, as in ICRNN [14]. Specifically, Assumption 2 is fulfilled when the following three conditions hold true:

- f and h are convex wrt. (x,u) and nondecreasing wrt. x.
- $\hat{\ell}$ is convex wrt. (U,Y), strongly convex wrt. U, and nondecreasing wrt. Y.
- $\hat{g}$ is convex wrt. (U,Y) and nondecreasing wrt. Y.

B. Constraint qualification and local equivalence

This subsection discusses an implication of Assumption 3 as well as investigates the theoretical properties of the equation system (13).

Consider the solution (χ', θ') to (13). Because $-1 \le v_i \le 1$ for every element $v \in \partial r(U)$ and $i \in \mathbb{Z}_N$, the KKT condition (7) implies $-\lambda \le d_i(U', z', \theta') \le \lambda$. Let us divide the index set $\mathbb{Z}_N$ into the following three subsets:

$$\mathscr{I}^+ := \left\{ i \in \mathbb{Z}_N \mid d_i(U', z', \theta') + \lambda = 0 \right\}, \tag{18}$$

$$\mathscr{I}^- := \left\{ i \in \mathbb{Z}_N \mid d_i(U', z', \theta') - \lambda = 0 \right\}, \tag{19}$$

$$\mathscr{I}^0 := \left\{ i \in \mathbb{Z}_N \mid -\lambda < d_i(U', z', \theta') < \lambda \right\}. \tag{20}$$

We define two vectors $\rho, \sigma \in \mathbb{R}^N$ as follows:

$$\rho_i = \begin{cases} 1 & \text{if } i \in \mathscr{I}^+, \\ -1 & \text{if } i \in \mathscr{I}^-, \\ 0 & \text{if } i \in \mathscr{I}^0, \end{cases} \quad \sigma_i = \begin{cases} 0 & \text{if } i \in \mathscr{I}^+, \\ 0 & \text{if } i \in \mathscr{I}^-, \\ 1 & \text{if } i \in \mathscr{I}^0. \end{cases}$$

Consider a smooth optimization problem

$$\hat{U}^*(\theta) = \underset{U \in \mathscr{C} \cap \hat{\mathscr{C}}}{\arg\min}\ \ell(U, \theta) + \lambda \rho^\mathsf{T} U, \tag{21}$$

$$\hat{\mathscr{C}} := \left\{ U \in \mathbb{R}^N \mid \rho \odot U \ge 0,\ \sigma \odot U = 0 \right\}. \tag{22}$$

Compared with (2), the ℓ_1-regularization term is relaxed to the affine function; instead, the feasible set is limited to $\hat{\mathscr{C}}$.

The KKT condition states that, for the optimal input $\hat{U}^*$, there exist μ^*, $\hat{\eta}^*$, and $\hat{\nu}^*$ such that $\hat{\chi}^* = [\hat{U}^*; \hat{z}^*; \hat{\eta}^*; \hat{\nu}^*]$ satisfies the following equation:

$$\hat{F}(\hat{\chi}^*, \mu^*, \theta) = 0, \tag{23}$$

$$\hat{F}(\chi, \mu, \theta) := \begin{bmatrix} d(U, z, \theta) + \lambda\rho + (\sigma - \rho) \odot \mu \\ \psi(U, \mu) \\ -z + \left(\frac{\partial g}{\partial U}(U)\right)^\mathsf{T} \eta + A^\mathsf{T} \nu \\ \phi(g(U), \eta) \\ LU + b \end{bmatrix}, \tag{24}$$

$$\psi_i(U, \mu) = \begin{cases} \varphi(\rho_i U_i, \mu_i) & \text{if } i \in \mathscr{I}^+ \cup \mathscr{I}^-, \\ U_i & \text{if } i \in \mathscr{I}^0. \end{cases} \tag{25}$$

Theorem 2: Consider (χ', θ') satisfying (13), and consider the smoothed optimization problem (21). We define $\mathscr{I}^+$, $\mathscr{I}^-$, and $\mathscr{I}^0$ by (18)–(20) and $\mathscr{N} \subset \mathbb{R}^{K+p}$ as follows:

$$\mathscr{N} := \left\{ [\chi^*; \theta] \;\middle|\; \begin{array}{ll} d_i(U, z, \theta) < \lambda & (i \in \mathscr{I}^+) \\ d_i(U, z, \theta) > -\lambda & (i \in \mathscr{I}^-) \\ -\lambda < d_i(U, z, \theta) < \lambda & (i \in \mathscr{I}^0) \end{array} \right\}.$$

Then, if $[\chi^*; \theta] \in \mathscr{N}$ and $F(\chi^*, \theta) = 0$, there exists μ^* satisfying $\hat{F}(\chi^*, \mu^*, \theta) = 0$. Conversely, if $[\hat{\chi}^*; \theta] \in \mathscr{N}$ and $\hat{F}(\hat{\chi}^*, \mu^*, \theta) = 0$, then $F(\hat{\chi}^*, \theta) = 0$. ◁

The proof is given in Appendix II. This theorem indicates that the equation (13) is locally equivalent to the KKT condition (24) for the smooth optimization problem (21), which is equivalent to the sparse optimal control problem (2). The LICQ for (21) requires that the matrix $\left[(J_g)_{\mathscr{I}:}; L; I_{\mathscr{I}:}\right]$ is column full rank. In fact, this requirement is equivalent to Assumption 3.

V. Numerical Example

This section demonstrates the IOC of sparse control using our framework of differentiable sparse optimal control.

A. Mimicking sparse control via IOC

We apply the proposed method to Example 1 for the following linear time-invariant system with $r = m = 2$

$$f(x,u) = Ax + Bu,\ h(x,u) = Cx,\ t = 0, \ldots, 30$$

$$A = \begin{bmatrix} 1 & 0.1 & 0.0 & 0.0 \\ -0.12 & 0.9 & 0.08 & 0.04 \\ 0 & 0 & 1 & 0.1 \\ 0.1 & 0.05 & -0.1 & 0.95 \end{bmatrix},\ B = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0.8 \\ 0 & 0 \end{bmatrix},$$

$$C = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

with inequality constraints $-1 \le u_i^{(t)} \le 1$ $(i = 1, 2,\ t = 0, \ldots, 30)$ and equality constraints $\sum_i \sum_t u_i^{(t)} = 0,\ \lambda = 1.$

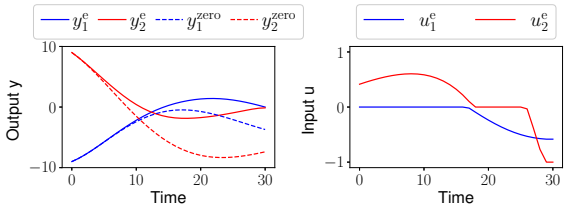

Fig. 1. Expert's sparse control. The expert's trajectory and uncontrolled trajectory are depicted on the left, and the expert's input is depicted on the right.

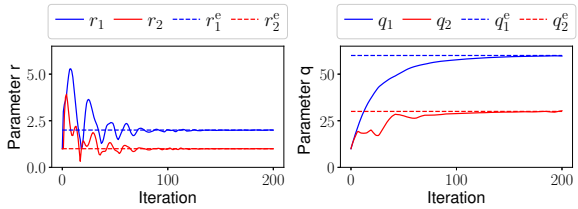

Fig. 2. History of parameters r_1, r_2, q_1, q_2 with expert's parameters $r_1^e, r_2^e, q_1^e, q_2^e$.

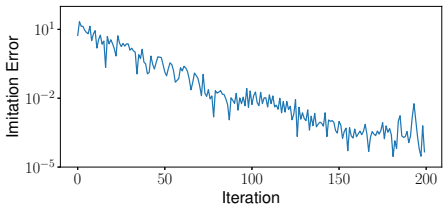

Fig. 3. History of imitation error for each batch.

Let the parameter $\theta = [r_1; r_2; q_1; q_2]$ be nonnegative weights (implemented using absolute values) and $\hat{\ell}$ as follows:

$$\hat{\ell}(U, Y, \theta) = \sum_{t=0}^{30} \sum_{i=1}^{2} r_i \left(u_i^{(t)}\right)^2 + \sum_{i=1}^{2} q_i \left(y_i^{(30)}\right)^2. \tag{26}$$

This cost function expresses the trade-off relationship between reducing the input power and bringing the final output closer to the origin using the control input. It also determines the degree of sparsity of the input because the scale of r_i and q_i determines the ratio of $\hat{\ell}(U, Y, \theta)$ to $r(U)$. We sampled 100 initial states $\{x_0^d\}_{d=1}^{100}$ from a uniform distribution over a rectangular region $[-10, 10]^4$.

The expert's actions are prepared using $U_i^e := U^*(\theta^e, x_0^i)$ with the expert's parameter $\theta^e := [2; 1; 60; 30]$ so that the optimal solution to (6) is θ^e. Fig. 1 illustrates the expert's sparse optimal control $U^*(\theta^e, x_0)$ for the initial state $x_0 = [-9; 5; 9; -8]$, as well as the controlled output Y^e and the uncontrolled output Y^{zero}. The final value of the controlled output is much closer to the origin than that of the uncontrolled output.

We expect that the parameter θ converges to the expert's parameter θ^e using the gradient-based optimization algorithm which solves (6). We calculate the sensitivity $\frac{\partial U^*}{\partial \theta}(\theta)$ as well as sparse optimal control $U^*(\theta)$ using Algorithm 1. Note that ℓ satisfies Assumption 2 if $r_1, r_2 \neq 0$, and that Assumption 3 holds true if $U_i^* \notin \{0, 1, -1\}$ for some $i \in \mathbb{Z}_N$.

Let the initial parameter be set to $\theta = [1; 1; 10; 10]$. Mini-batch learning using the Adam optimizer results in the acquisition of the parameter value $\theta = [2.00; 1.00; 59.7; 30.3]$. Fig. 2 illustrates that the parameter θ asymptotically approaches the expert's parameter θ^e. Fig. 3 depicts the logarithmic scale plot of the imitation error in (6) for each batch, demonstrating that the loss is reduced from approximately 10^1 to approximately 10^{-4}, thereby proving the successful replication of sparse control.

B. Accuracy and computational burden

To supplement the aforementioned IOC result, we compare the sensitivity obtained using Algorithm 1 with that obtained via numerical difference. Let $\theta = [2; 1; 60; 30]$ and $x_0 = [-9; 5; 9; -8]$. The optimal input is the same as that depicted in the right column of Fig. 1, which is sparse. The sensitivity $\frac{\partial U^*}{\partial \theta}(\theta)$ is calculated using Algorithm 1 as well as using centered difference around θ with perturbation $\varepsilon = 1.0 \times 10^{-2}$. When both values are compared elementwise, the absolute value of the difference is observed to be at most 4.0×10^{-5}. This verifies that the sensitivity of $U^*(\theta)$ is calculated correctly. Note that the centered difference requires solving the sparse optimal control problem $2p$ times, where p denotes the dimension of θ, while the proposed framework requires it to be solved only once.

VI. Conclusions

In this paper, a framework for differentiable sparse optimal control was developed. Our method establishes new avenues to differentiating nonsmooth optimal control. Future work includes extending our method to general nonsmooth problems and deriving their differentiability conditions.

Appendix I
Proof of Theorem 1

Proof: Consider the smoothed optimization problem (21) around $\theta' = \theta$, where $[\chi^*; \theta] \in \mathcal{N}$. We take $\alpha \in \mathbb{R}_>$ and modify the expression for the feasible set $\hat{\mathscr{C}}$ as follows:

$$\hat{\mathscr{C}} = \left\{ U \in \mathbb{R}^N \mid \alpha\rho \odot U \geq 0,\ \alpha\sigma \odot U = 0 \right\}.$$

The KKT condition can be expressed as $\hat{F}^{(\alpha)}(\hat{\chi}^*, \mu^*, \theta) = 0$, where $\hat{F}^{(\alpha)}$ is a function replacing ψ with $\psi^{(\alpha)}$ in (24) and $\psi^{(\alpha)}(U, \mu)$ is given as follows:

$$\psi_i^{(\alpha)}(U, \mu) = \begin{cases} \varphi(\alpha\rho_i U_i, \mu_i) & \text{if } i \in \mathscr{I}^+ \cup \mathscr{I}^-, \\ \alpha U_i & \text{if } i \in \mathscr{I}^0. \end{cases}$$

Let $\hat{P} := \alpha \mathrm{diag}_i(\rho_i + \sigma_i)$. Note that $\hat{P}$ is nonsingular. Then, Assumptions 2 and 3 indicate that the smoothed problem admits a unique KKT point. By the equivalence, the solution χ^* of $F(\chi^*, \theta) = 0$ is unique. Theorem 2 implies the existence of μ^* satisfying $\hat{F}^{(\alpha)}(\chi^*, \mu^*, \theta) = 0$.

We now proceed to the proof of the latter claim. We denote the Clarke Jacobian of $\psi^{(\alpha)}$ at (χ^*, μ^*) by $\partial\psi^{(\alpha)}$ and that

of $\hat{F}^{(\alpha)}$ at $(\hat{U}^*, \mu^*, \theta)$ by $\partial \hat{F}^{(\alpha)}$. $\partial \psi$ is expressed as follows:

$$\partial \psi^{(\alpha)} = \left\{ \begin{bmatrix} \mathrm{diag}_i(\alpha(\rho_i + \sigma_i) a_i) \\ \mathrm{diag}_i(\alpha(\rho_i + \sigma_i) b_i) \end{bmatrix}^\mathsf{T} \middle| \begin{matrix} [a_i \ b_i] \in \partial \psi_i(\rho_i U_i, \mu_i) \\ i \in \mathbb{Z}_N \end{matrix} \right\},$$

$$\partial \psi_i(a,b) = \begin{cases} \partial \varphi_{\mathrm{fb}}(a,b) & \text{if } i \in \mathscr{I}^+ \cup \mathscr{I}^-, \\ \{[1 \ 0]\} & \text{if } i \in \mathscr{I}^0. \end{cases}$$

$[\hat{J}^\chi \ \hat{J}^\theta] \in \partial \hat{F}^{(\alpha)}$ is fully parameterized by $[\psi^g \ \psi^\mu] \in \partial \psi^{(\alpha)}$ and $[\phi^g \ \phi^\eta] \in \partial \phi$, and $\hat{J}^\chi$ is expressed as follows:

$$\hat{J}^\chi = \underbrace{\begin{bmatrix} \hat{P} & H^\ell & I & 0 & 0 \\ \psi^\mu & \hat{P}\psi^g & 0 & 0 & 0 \\ 0 & H^g & -I & J_g^\mathsf{T} & L^\mathsf{T} \\ 0 & \phi^g J_g & 0 & \phi^\eta & 0 \\ 0 & L & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 0 & I \\ I & 0 & 0 & 0 & 0 \\ 0 & I & 0 & 0 & 0 \\ 0 & 0 & I & 0 & 0 \\ 0 & 0 & 0 & I & 0 \end{bmatrix}}_{=: \hat{J}'}.$$

The 3rd diagonal block element of $\hat{J}'$ is $-I$, and via appropriate permutations, the Schur complement matrix of $-I$ corresponds to the KKT matrix of the smoothed optimization problem (21), which is nonsingular by Theorem 3.4 in [10]. Therefore, $\hat{J}'$ is nonsingular. The upper-left matrix of $\hat{J}'$ is $\hat{P}$, and we have nonsingularity of its Schur complement matrix, which is denoted by S. The straightforward calculation reveals that S comprises the same components as J^χ, except for the 1st-row block. Because ψ^g, ψ^μ, and $\hat{P}$ are diagonal, the 1st-row block of S, denoted by S_1, is given as follows:

$$S_1 = \hat{P} \begin{bmatrix} R & -\alpha^{-2}\psi^\mu & 0 & 0 \end{bmatrix}, \ R = \psi^g - \alpha^{-2}\psi^\mu H^\ell.$$

We select $\alpha = \gamma^{-\frac{1}{2}}$. For simplicity, we denote $d(U^*, z^*, \theta)$ by d. The claim follows from the existence of $[\psi^g \ \psi^\mu] \in \partial \psi$ which satisfies $D = I - \psi^g = -\psi^\mu$. To prove this existence, we consider four cases, (A)–(D). (A) $U_i - \gamma d_i > \gamma\lambda$, where $D_{ii} = 1$. In this case, the equivalence indicates that $i \in \mathscr{I}^+$, $U_i^* > 0$, and $\mu_i^* = 0$. Thus, $[\psi_{ii}^g \ \psi_{ii}^\mu] = [0 \ -1]$. Therefore, $D_{ii} = 1 - \psi_{ii}^g = -\psi_{ii}^\mu = 1$ holds. (B) $U_i - \gamma d_i < -\gamma\lambda$, where $D_{ii} = 1$. It follows that $i \in \mathscr{I}^-$, $U_i^* < 0$, and $\mu_i^* = 0$. Thus, $[\psi_{ii}^g \ \psi_{ii}^\mu] = [0 \ -1]$. Therefore, $D_{ii} = 1 - \psi_{ii}^g = -\psi_{ii}^\mu = 1$ holds. (C) $|U_i - \gamma d_i| < \gamma\lambda$, where $D_{ii} = 0$. We have $U_i^* = 0$ and $i \in \mathscr{I}^0$. Therefore, $[\psi_{ii}^g \ \psi_{ii}^\mu] = [1 \ 0]$ holds, which implies $D_{ii} = 1 - \psi_{ii}^g = -\psi_{ii}^\mu = 0$. (D) $U_i^* - \gamma d_i = \pm\gamma\lambda$, where $0 \leq D_{ii} \leq 1$. In this case, $U_i^* = 0$ and $\mu_i^* = 0$. Thus, $[\psi_{ii}^g \ \psi_{ii}^\mu] \in \mathscr{D}$. Since $\{[1-a \ \ a] \mid 0 \leq a \leq 1\} \subset \mathscr{D}$, we have $[\psi_{ii}^g \ \psi_{ii}^\mu]$ such that $D_{ii} = 1 - \psi_{ii}^g = -\psi_{ii}^\mu$ holds. ■

Appendix II
Proof of Theorem 2

Lemma 1: $\zeta^{(\gamma)}(a,b) := a - \max\{a - \gamma b, 0\}$ is the complementarity function for any $\gamma > 0$. ◁

Proof: The relationship $\zeta^{(\gamma)}(a,b) = 0 \Leftrightarrow a \geq 0, b \geq 0, ab = 0$ can be derived in a straightforward manner by considering the two cases, $b > 0$ and $b = 0$. ■

Lemma 2: If $v + a > x$, then $x = s_a(v) \Leftrightarrow x = \max\{v - a, 0\}$. If $v - a < x$, then $x = s_a(v) \Leftrightarrow x = \min\{v + a, 0\}$. ◁

Proof: $v + a > x = s_a(v)$ implies $s_a(v) \geq 0$ and $v + a > 0$. Thus $s_a(v) = \max\{v - a, 0\}$. Conversely, $v + a > x = \max\{v - a, 0\}$ implies $x \geq 0$ and thus $v + a > 0$. Therefore $s_a(v) = \max\{v - a, 0\}$. The former implies the latter because $-v + a > -x$ and $s_a(v) = -s_a(-v)$. ■

We now proceed to the proof of Theorem 2.

Proof: If $i \in \mathscr{I}^+$, then $[\hat{\chi}^*; \theta] \in \mathscr{N}$ implies $\hat{U}_i^* - \gamma d_i(\hat{U}^*, \hat{z}^*, \theta) + \gamma\lambda > \hat{U}_i^*$. Thus, by Lemma 2, we have

$$\begin{aligned} & \hat{U}_i^* = s_{\gamma\lambda}\left(\hat{U}_i^* - \gamma d_i(\hat{U}^*, \hat{z}^*, \theta) - \gamma\lambda\right) \\ \Leftrightarrow \ & \zeta^{(\gamma)}\left(\hat{U}_i^*, d_i(\hat{U}^*, \hat{z}^*, \theta) + \lambda\right) = 0 \\ \Leftrightarrow \ & \zeta^{(\gamma)}(\hat{U}_i^*, \mu_i^*) = 0, \ d_i(\hat{U}^*, \hat{z}^*, \theta) + \lambda - \mu_i^* = 0. \end{aligned}$$

Additionally, Lemma 1 implies $\zeta^{(\gamma)}(\hat{U}_i^*, \mu_i^*) = 0 \Leftrightarrow \varphi(\hat{U}_i^*, \mu_i^*) = 0$. If $i \in \mathscr{I}^-$, in a similar manner as before, we have

$$\begin{aligned} & \hat{U}_i^* = s_{\gamma\lambda}\left(\hat{U}_i^* - \gamma d_i(\hat{U}^*, \hat{z}^*, \theta) - \gamma\lambda\right) \\ \Leftrightarrow \ & -\zeta^{(\gamma)}\left(-\hat{U}_i^*, -d_i(\hat{U}^*, \hat{z}^*, \theta) + \lambda\right) = 0 \\ \Leftrightarrow \ & \zeta^{(\gamma)}(-\hat{U}_i^*, \mu_i^*) = 0, \ d_i(\hat{U}^*, \hat{z}^*, \theta) - \lambda + \mu_i^* = 0 \\ \Leftrightarrow \ & \varphi(-\hat{U}_i^*, \mu_i^*) = 0, \ d_i(\hat{U}^*, \hat{z}^*, \theta) - \lambda + \mu_i^* = 0. \end{aligned}$$

If $i \in \mathscr{I}^0$, then both the aforementioned equivalences hold, and thus $\hat{U}_i^* = 0$. In this case, the converse is trivial. ■

References

[1] P. Englert, N. A. Vien, and M. Toussaint, "Inverse KKT: Learning cost functions of manipulation tasks from demonstrations," *The International Journal of Robotics Research*, vol. 36, no. 13-14, pp. 1474–1488, 2017.

[2] B. Amos, I. Jimenez, J. Sacks, B. Boots, and J. Z. Kolter, "Differentiable MPC for end-to-end planning and control," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[3] W. Jin, Z. Wang, Z. Yang, and S. Mou, "Pontryagin differentiable programming: An end-to-end learning and control framework," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7979–7992, 2020.

[4] M. Nagahara, *Sparsity Methods for Systems and Control*. Now Publishers, 2020.

[5] Q. Bertrand, Q. Klopfenstein, M. Massias, M. Blondel, S. Vaiter, A. Gramfort, and J. Salmon, "Implicit differentiation for fast hyperparameter selection in non-smooth convex learning," *J. Mach. Learn. Res.*, vol. 23, no. 1, jan 2022.

[6] N. Parikh, S. Boyd, *et al.*, "Proximal algorithms," *Foundations and Trends® in Optimization*, vol. 1, no. 3, pp. 127–239, 2014.

[7] D. W. Peterson, "A review of constraint qualifications in finite-dimensional spaces," *SIAM Review*, vol. 15, no. 3, pp. 639–654, 1973. [Online]. Available: https://doi.org/10.1137/1015075

[8] F. Clarke, "On the inverse function theorem," *Pacific Journal of Mathematics*, vol. 64, no. 1, pp. 97–102, 1976.

[9] J. Bolte, T. Le, E. Pauwels, and T. Silveti-Falls, "Nonsmooth implicit differentiation for machine-learning and optimization," *Advances in Neural Information Processing Systems*, vol. 34, pp. 13 537–13 549, 2021.

[10] F. Facchinei, A. Fischer, and C. Kanzow, "Regularity properties of a semismooth reformulation of variational inequalities," *SIAM Journal on Optimization*, vol. 8, no. 3, pp. 850–869, 1998. [Online]. Available: https://doi.org/10.1137/S1052623496298194

[11] L. Qi and J. Sun, "A nonsmooth version of Newton's method," *Mathematical Programming*, vol. 58, no. 1-3, pp. 353–367, 1993.

[12] S. Kyochi, S. Ono, and I. Selesnick, "Epigraphical reformulation for non-proximable mixed norms," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 5400–5404.

[13] A. Bemporad, F. Borrelli, and M. Morari, "Model predictive control based on linear programming - the explicit solution," *IEEE Transactions on Automatic Control*, vol. 47, no. 12, pp. 1974–1985, 2002.

[14] Y. Chen, Y. Shi, and B. Zhang, "Optimal control via neural networks: A convex approach," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=H1MW72AcK7