Learning Linear Complementarity Systems

Wanxin Jin JINWX@SEAS.UPENN.EDU
Alp Aydinoglu ALPAYD@SEAS.UPENN.EDU
Mathew Halm MHALM@SEAS.UPENN.EDU
Michael Posa POSA@SEAS.UPENN.EDU
University of Pennsylvania, Philadelphia, PA 19104, USA

Editors: R. Firoozi, N. Mehr, E. Yel, R. Antonova, J. Bohg, M. Schwager, M. Kochenderfer

Abstract

This paper investigates the learning, or system identification, of a class of piecewise-affine dynamical systems known as linear complementarity systems (LCSs). We propose a violation-based loss which enables efficient learning of the LCS parameterization, without prior knowledge of the hybrid mode boundaries, using gradient-based methods. The proposed violation-based loss incorporates both dynamics prediction loss and a novel complementarity - violation loss. We show several properties attained by this loss formulation, including its differentiability, the efficient computation of first- and second-order derivatives, and its relationship to the traditional prediction loss, which strictly enforces complementarity. We apply this violation-based loss formulation to learn LCSs with tens of thousands of (potentially stiff) hybrid modes. The results demonstrate a state-of-the-art ability to identify piecewise-affine dynamics, outperforming the clustering-based piecewise-affine regression methods and the methods which must differentiate through non-smooth linear complementarity constraints.

Keywords: linear complementarity problems, piece-wise affine systems, system identification

1. Introduction

Many physical systems of interest are well captured by multi-modal or hybrid representations. For example, robotics problems which treat contact with the environment (Stewart and Trinkle, 2000; Brogliato, 1999), optimal control problems (Bemporad et al., 2000), and control of networks (Heemels et al., 2002), all exhibit switching or hybrid properties.

In this work, we are interested in system identification/model learning of multi-modal systems. We focus on piecewise-affine (PWA) models as they can sufficiently describe the multi-modal nature of dynamics due to the approximation properties of affine functions (Breiman, 1993; Lin and Unbehauen, 1992) but are tractable enough for control tasks due to their simple (affine) structure over polyhedral regions (Bemporad and Morari, 1999). Even though PWA models are widely used, it is well-known that PWA regression is NP-hard in general (see (Lauer, 2015) for a detailed analysis), because it requires simultaneous classification of the data points into modes and the regression of a submodel for each mode.

In this paper, we consider PWA models in the context of linear complementarity systems (LCSs) (Heemels et al., 2000). We focus on a subclass of LCS models (with P-matrix assumption) that are equivalent to continuous piecewise affine models (Heemels et al., 2001; Camlibel et al., 2007). LCS models are efficient representations of PWA models and an LCS has the ability to represent/approximate a system with large number of hybrid modes compactly, with only few com-

plementarity variables. In some cases, an LCS with n_λ complementarity variables is equivalent to a PWA model with 2^{n_λ} modes. Many robotics problems that involve contact can be efficiently locally approximated via LCS models, e.g., we have exploited the LCS representation to enable contact-aware (Aydinoglu et al., 2021) and real-time control of robotic tasks (Aydinoglu and Posa, 2022). In this work, we propose an approach that learns an LCS from state-input data of a hybrid system, which does not contain any prespecified number of modes. The approach is able to identify LCS models by proposing an implicit loss function.

1.1. Related Work

Many successful approaches in identifying piecewise models have been proposed over the years. See the survey paper (Paoletti et al., 2007) for a detailed overview. Mixed integer formulations that mainly focus on hinging hyperplanes and piecewise affine Wiener models have been proposed (Roll et al., 2004) but as the number of integer variables scale with number of data points such approaches are only applicable in small data regime. On the contrary, researchers have also focused on convex formulations where first they estimate a set of submodels and then select few of them that explains the data (Elhamifar et al., 2014), but the approach relies on restricting the parameter space and can be overly conservative. Many alternate approaches that enable PWA system identification from data exist such as (Ferrari-Trecate et al., 2003; Nakada et al., 2005; Bemporad et al., 2005; Hartmann et al., 2015; Du et al., 2020).

Researchers also suggested recursive PWA identification algorithms (Bako et al., 2011; Breschi et al., 2016). Most of the above methods are clustering-based where a predetermined number of models are identified and each training data point is associated with one of the models. Then, linear separation techniques are used to compute the polyhedral partitions. This iterative nature can lead to overly conservative, suboptimal solutions. Approaches that simultaneously cluster, PWL-separate and fit (Bemporad, 2021) rely heavily on initial assignment of data points to clusters. Unlike our approach, none of the methods above have been tested on identifying PWA functions with thousands of partitions, and most of them have been only tested on functions with less than 30 pieces.

For more expressive models such as deep neural networks, researchers have explored the positive effect of imposing structured knowledge to capture the multi-modality (de Avila Belbute-Peres et al., 2018; Li et al., 2019; Battaglia et al., 2016). Particularly in robotics, special emphasis has been on multi-body systems with frictional contact (Geilinger et al., 2020) and it has been shown that imposing structure leads to accurate, sample efficient strategies (Pfrommer et al., 2020). Similar work has demonstrated the difficulty inherent in learning non-smooth dynamical systems without exploiting particular structures (Parmar et al., 2021). Researchers have also explored learning models as functionals of signed-distance fields (Driess et al., 2021). These methods lead to rich, accurate but complex models that are not amenable to techniques of model-based control. On the contrary, here our focus is on simple models such as PWA models that enable model-based control while sufficiently capturing the hybrid dynamics.

Notation: Regular and bold lowercase letters represent scalar and vectors, respectively. Uppercase letters represent matrices. For vector $\boldsymbol{v} \in \mathbb{R}^n$, $\boldsymbol{v}[i]$ is the i-th entry, $i = 1, 2, ..., n$. $\mathrm{diag}(\boldsymbol{v})$ is to diagonalize a vector $\boldsymbol{v}$ into a matrix. $\mathrm{vec}\,(A)$ denotes the vectorlization of a matrix A into a column vector; $\otimes$ is the Kronecker product. I_n denotes the identity matrix with size of $n \times n$. $A \succ 0$ means symmetric A is positive definite.

2. Problem Statement

Consider the following discrete-time linear complementarity system (LCS), where the state evolution is governed by a linear dynamics in (1a) and a linear complementarity problem (LCP) in (1b):

$$\boldsymbol{x}_{t+1} = A\boldsymbol{x}_t + B\boldsymbol{u}_t + C\boldsymbol{\lambda}_t + \boldsymbol{d}, \tag{1a}$$

$$\mathbf{0} \leq \boldsymbol{\lambda}_t \perp D\boldsymbol{x}_t + E\boldsymbol{u}_t + F\boldsymbol{\lambda}_t + \boldsymbol{c} \geq \mathbf{0}. \tag{1b}$$

Here, $\boldsymbol{x}_t \in \mathbb{R}^{n_x}$ and $\boldsymbol{u}_t \in \mathbb{R}^{n_u}$ are the system state and input at time step t, respectively; and $\boldsymbol{\lambda}_t \in \mathbb{R}^r$ is the complementarity variable at time step t. $A \in \mathbb{R}^{n_x \times n_x}$, $B \in \mathbb{R}^{n_x \times n_u}$, $C \in \mathbb{R}^{n_x \times n_\lambda}$, $\boldsymbol{d} \in \mathbb{R}^{n_x}$, $D \in \mathbb{R}^{n_\lambda \times n_x}$, $E \in \mathbb{R}^{n_\lambda \times n_u}$, $F \in \mathbb{R}^{n_\lambda \times n_\lambda}$, and $\boldsymbol{c} \in \mathbb{R}^{n_\lambda}$ are system matrix/vector parameters. At $(\boldsymbol{x}_t, \boldsymbol{u}_t)$, $\boldsymbol{\lambda}_t$ is solved from the LCP in (1b), written as

$$\boldsymbol{\lambda}_t \in \mathrm{LCP}(F, \boldsymbol{q}_t) \quad \text{with} \quad \boldsymbol{q}_t := D\boldsymbol{x}_t + E\boldsymbol{u}_t + \boldsymbol{c}. \tag{2}$$

It is well-known that $\mathrm{LCP}(F, \boldsymbol{q}_t)$ has a unique solution $\boldsymbol{\lambda}_t$ for every $\boldsymbol{q}_t$ if and only if F is P-matrix (Cottle et al., 2009). We will discuss this in the next section.

In this paper, we consider to learn a LCS from a dataset $\mathcal{D} = \{(\boldsymbol{x}_t^*, \boldsymbol{u}_t^*, \boldsymbol{x}_{t+1}^*)\}_{t=1}^N$. Specifically, we aim to find the system parameter

$$\boldsymbol{\theta} = \{A, B, C, \boldsymbol{d}, D, E, F, \boldsymbol{c}\} \tag{3}$$

by minimizing a loss $L(\boldsymbol{\theta}, \mathcal{D})$. Thus, the problem of interest in this paper is to solve

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \quad L(\boldsymbol{\theta}, \mathcal{D}) + R(\boldsymbol{\theta}). \tag{4}$$

Here, $R(\boldsymbol{\theta})$ can be any regularization term imposed on $\boldsymbol{\theta}$ which will be discussed later.

3. LCP and Prediction-based Formulation

This section will discuss the solution to LCP in (2), and then describe a prediction-based loss formulation $L(\boldsymbol{\theta}, \mathcal{D})$ for (4). To start, we make the following assumption on F in (2).

Assumption 1 *$F \in \mathbb{R}^{n_\lambda \times n_\lambda}$ satisfies $F + F^\mathsf{T} \succ 0$.*

The set of F satisfying Assumption 1 contains all positive matrices of feasible dimension and any asymmetric matrices with definite-positive symmetric part. In robotics applications, Assumption 1 has been widely used in soft contact dynamics such as (Aydinoglu et al., 2021). While restrictive to model standard directional dynamics (Stewart and Trinkle, 2000), this assumption has been similarly used in quite a few works (Anitescu and Hart, 2004; Castro et al., 2021) for solving frictional contact. Any F satisfying Assumption 1 can be shown to be a P-matrix (Tsatsomeros, 2002), thus leading to the existence and uniqueness of $\boldsymbol{\lambda}_t$. In fact, under Assumption 1, $\boldsymbol{\lambda}_t = \mathrm{LCP}(F, \boldsymbol{q}_t)$ can be solved by the following convex optimization due to the fact $\boldsymbol{\lambda}^\mathsf{T} F \boldsymbol{\lambda} = \frac{1}{2}\boldsymbol{\lambda}^\mathsf{T}(F + F^\mathsf{T})\boldsymbol{\lambda}$,

$$\boldsymbol{\lambda}_t = \arg \min_{\boldsymbol{\lambda}} \ \frac{1}{2}\boldsymbol{\lambda}^\mathsf{T}(F + F^\mathsf{T})\boldsymbol{\lambda} + \boldsymbol{\lambda}^\mathsf{T}\boldsymbol{q}_t \qquad \text{s.t.} \quad F\boldsymbol{\lambda} + \boldsymbol{q}_t \geq \mathbf{0}, \qquad \boldsymbol{\lambda} \geq \mathbf{0}, \tag{5}$$

With the above assumption, one natural loss in (4) can be

$$L^{\text{pred}}(\boldsymbol{\theta}, \mathcal{D}) = \sum_{t=1}^{N} \frac{1}{2}\|\boldsymbol{x}_{t+1}^{\boldsymbol{\theta}} - \boldsymbol{x}_{t+1}^*\|^2 \quad \text{with} \quad \begin{array}{l} \boldsymbol{x}_{t+1}^{\boldsymbol{\theta}} = A\boldsymbol{x}_t^* + B\boldsymbol{u}_t^* + C\boldsymbol{\lambda}_t^* + \boldsymbol{d}, \\ \boldsymbol{\lambda}_t^* = \text{LCP}(F, D\boldsymbol{x}_t^* + E\boldsymbol{u}_t^* + \boldsymbol{c}). \end{array} \tag{6}$$

Here, $\boldsymbol{x}_{t+1}^{\boldsymbol{\theta}}$ is the predicted next state, implicitly depending on $\boldsymbol{\theta}$. We call (6) *prediction-based loss*, as it evaluates the difference between the predicted $\boldsymbol{x}_{t+1}^{\boldsymbol{\theta}}$ and observed $\boldsymbol{x}_{t+1}^*$. One can minimize (6) via any gradient-based method by differentiating through LCP (de Avila Belbute-Peres et al., 2018). This requires differentiblity of a LCP, given below.

Lemma 1 *With Assumption 1, $\boldsymbol{\lambda}_t^* = \text{LCP}(F, D\boldsymbol{x}_t^* + E\boldsymbol{u}_t^* + \boldsymbol{c})$ is differentiable with respect to $(F, D, E, \boldsymbol{c})$, if the following strict complementarity holds at $\boldsymbol{\lambda}_t^*$: $\boldsymbol{\lambda}_t^*[i] > 0$ or $(F\boldsymbol{\lambda}_t^* + D\boldsymbol{x}_t^* + E\boldsymbol{u}_t^* + \boldsymbol{c})[i] > 0$, $\forall i = 1, 2, ..., n_\lambda$.*

A sketch of a proof of Lemma 1 is given in Appendix. The above (6) will serve as a benchmark in the following method development.

4. Proposed Method for Learning LCS

This section will develop a new method for learning LCS. As we will show this section and the experiments in next section, the proposed method attains several advantages over the prediction-based loss (6) both in theoretical property and implementation.

4.1. Violation-based Loss

To start, we give the following lemma stating an equivalence of a LCP.

Lemma 2 *Given any $\boldsymbol{q}_t \in \mathbb{R}^{n_\lambda}$ and F satisfying Assumption 1, solving $\boldsymbol{\lambda}_t = \text{LCP}(F, \boldsymbol{q}_t)$ is the equivalent to solving the following strongly-convex quadratic program:*

$$(\boldsymbol{\lambda}_t, \boldsymbol{\phi}_t) = \arg \min_{\boldsymbol{\lambda} \geq \boldsymbol{0},\ \boldsymbol{\phi} \geq \boldsymbol{0}} \quad \boldsymbol{\lambda}^\mathsf{T}\boldsymbol{\phi} + \frac{1}{2\gamma}\|F\boldsymbol{\lambda} + \boldsymbol{q}_t - \boldsymbol{\phi}\|^2, \tag{7}$$

with any constant $0 < \gamma < \sigma_{\min}(F^\mathsf{T} + F)$ ($\sigma_{\min}(\cdot)$ denotes the smallest singular value).

Proof Define $f(\boldsymbol{\lambda}, \boldsymbol{\phi}) := \boldsymbol{\lambda}^\mathsf{T}\boldsymbol{\phi} + \frac{1}{\gamma}\|F\boldsymbol{\lambda} + \boldsymbol{q}_t - \boldsymbol{\phi}\|^2$. By non-negativity, it is obvious that $\boldsymbol{\lambda}_t = \text{LCP}(F, \boldsymbol{q}_t)$ and $\boldsymbol{\phi}_t = F\boldsymbol{\lambda}_t + \boldsymbol{q}_t$ is a global solution to $f(\boldsymbol{\lambda}, \boldsymbol{\phi})$. Further, we need to show that $f(\boldsymbol{\lambda}, \boldsymbol{\phi})$ is strongly convex, because by strong convexity (7) has a unique global solution, which is $(\boldsymbol{\lambda}_t, \boldsymbol{\phi}_t)$. To do so, we compute the Hessian of $f(\boldsymbol{\lambda}, \boldsymbol{\phi})$,

$$\nabla^2 f = \begin{bmatrix} \frac{1}{\gamma}F^\mathsf{T}F & I_{n_\lambda} - \frac{1}{\gamma}F^\mathsf{T} \\ I_{n_\lambda} - \frac{1}{\gamma}F & \frac{1}{\gamma}I_{n_\lambda} \end{bmatrix}. \tag{8}$$

Due to Schur complement, $\nabla^2 f \succ 0$ iff $\frac{1}{\gamma}I_{n_\lambda} \succ 0$ and $\frac{1}{\gamma}F^\mathsf{T}F - (I_{n_\lambda} - \frac{1}{\gamma}F^\mathsf{T})(\frac{1}{\gamma}I_{n_\lambda})^{-1}(I_{n_\lambda} - \frac{1}{\gamma}F) \succ 0$. Since $\gamma > 0$, and we only need to show $\frac{1}{\gamma}F^\mathsf{T}F - (I_{n_\lambda} - \frac{1}{\gamma}F^\mathsf{T})(\frac{1}{\gamma}I_{n_\lambda})^{-1}(I_{n_\lambda} - \frac{1}{\gamma}F) = F^\mathsf{T} + F - \gamma I_{n_\lambda} \succ 0$. This is true because $\gamma < \sigma_{\min}(F^\mathsf{T} + F)$. This completes the proof. ■

In (7), we have introduced a proxy variable $\phi \geq \mathbf{0}$ to represent LCP constraint $F\boldsymbol{\lambda} + \boldsymbol{q}_t \geq \mathbf{0}$. Compared to other equivalences of LCP, such as (5), we emphasize the following benefits of (7) with the introduced proxy variable ϕ. First, (7) now only has box constraints which are *independent* from $\boldsymbol{\theta}$; this will facilitate the learning process because one does not need to explicitly track the active and inactive constraints and differentiate through constraints (which usually leads to numerical difficulty as shown in (Jin et al., 2021)). Second, compared to (5), (7) turns hard constraint $F\boldsymbol{\lambda} + \boldsymbol{q}_t \geq 0$ into a soft penalty; this may smooth the landscape of the proposed loss, facilitating the optimization over $\boldsymbol{\theta}$. With Lemma 2, we are now in a position to propose the following loss for learning LCS,

$$L_\epsilon(\boldsymbol{\theta}, \mathcal{D}) = \sum_{t=1}^{N} l_\epsilon(\boldsymbol{\theta}, \boldsymbol{x}_t^*, \boldsymbol{u}_t^*, \boldsymbol{x}_{t+1}^*) \quad \text{with} \tag{9a}$$

$$\begin{aligned} l_\epsilon(\boldsymbol{\theta}, \boldsymbol{x}_t^*, \boldsymbol{u}_t^*, \boldsymbol{x}_{t+1}^*) = \min_{\boldsymbol{\lambda}_t \geq \mathbf{0},\ \phi_t \geq \mathbf{0}} \ & \frac{1}{2}\|A\boldsymbol{x}_t^* + B\boldsymbol{u}_t^* + C\boldsymbol{\lambda}_t + \boldsymbol{d} - \boldsymbol{x}_{t+1}^*\|^2 + \\ & \frac{1}{\epsilon}\left(\boldsymbol{\lambda}_t^\mathsf{T}\phi_t + \frac{1}{2\gamma}\|D\boldsymbol{x}_t^* + E\boldsymbol{u}_t^* + F\boldsymbol{\lambda}_t + \boldsymbol{c} - \phi\|^2\right), \end{aligned} \tag{9b}$$

with $\epsilon > 0$. In $L_\epsilon(\boldsymbol{\theta}, \mathcal{D})$, the loss $l_\epsilon(\boldsymbol{\theta}, \boldsymbol{x}_t^*, \boldsymbol{u}_t^*, \boldsymbol{x}_{t+1}^*)$ on each data point $(\boldsymbol{x}_t^*, \boldsymbol{u}_t^*, \boldsymbol{x}_{t+1}^*)$ includes two parts: the violation of dynamics (1a) and the violation of the LCP, as stated in Lemma 2. We have introduced $\epsilon > 0$ to control the weight of penalties on the two violations. (9) is to minimize data's violation to both dynamics (1a) and complementarity constraints (1b), thus we name it *violation-based loss*. In what follows, we will show that the violation-based loss attains some good properties both for analysis and algorithmic implementation, in comparison with prediction-based loss (6).

4.2. Properties of Violation-based Loss

The first lemma shows that the violation-based loss (9) is a strongly-convex quadratic program w.r.t. $(\boldsymbol{\lambda}, \phi)$ and allows much easier computation of the gradient w.r.t. $\boldsymbol{\theta}$.

Lemma 3 *Given F satisfying Assumption 1 and any constant $0 < \gamma < \sigma_{\min}(F^\mathsf{T} + F)$,*

(a) (9b) is strongly-convex quadratic program with respect to $(\boldsymbol{\lambda}_t, \phi_t)$ for any $\epsilon > 0$.

(b) Let $(\boldsymbol{\lambda}_t^{\epsilon,\boldsymbol{\theta}}, \phi_t^{\epsilon,\boldsymbol{\theta}})$ be the solution to (9b). $L_\epsilon(\boldsymbol{\theta}, \mathcal{D})$ is differentiable with respect to $\boldsymbol{\theta}$ if the strict complementarity holds for both $\boldsymbol{\lambda}_t \geq \mathbf{0}$ and $\phi_t \geq \mathbf{0}$ at $(\boldsymbol{\lambda}_t^{\epsilon,\boldsymbol{\theta}}, \phi_t^{\epsilon,\boldsymbol{\theta}})$, $t = 1, 2, ..., N$. The gradient is given by

$$\begin{aligned} &\nabla_A L_\epsilon = \sum_{t=1}^{N} \boldsymbol{e}_t^{\text{dyn}} \boldsymbol{x}_t^{*\mathsf{T}},\ \nabla_B L_\epsilon = \sum_{t=1}^{N} \boldsymbol{e}_t^{\text{dyn}} \boldsymbol{u}_t^{*\mathsf{T}},\ \nabla_C L_\epsilon = \sum_{t=1}^{N} \boldsymbol{e}_t^{\text{dyn}} \boldsymbol{\lambda}_t^{*,\boldsymbol{\theta}\mathsf{T}},\ \nabla_{\boldsymbol{d}} L_\epsilon = \sum_{t=1}^{N} \boldsymbol{e}_t^{\text{dyn}}, \\ &\nabla_D L_\epsilon = \sum_{t=1}^{N} \boldsymbol{e}_t^{\text{lcp}} \boldsymbol{x}_t^{*\mathsf{T}},\ \nabla_E L_\epsilon = \sum_{t=1}^{N} \boldsymbol{e}_t^{\text{lcp}} \boldsymbol{u}_t^{*\mathsf{T}},\ \nabla_F L_\epsilon = \sum_{t=1}^{N} \boldsymbol{e}_t^{\text{lcp}} \boldsymbol{\lambda}_t^{\epsilon,\boldsymbol{\theta}\mathsf{T}},\quad \nabla_{\boldsymbol{c}} L_\epsilon = \sum_{t=1}^{N} \boldsymbol{e}_t^{\text{lcp}}, \end{aligned} \tag{10}$$

with $\boldsymbol{e}_t^{\text{dyn}} := A\boldsymbol{x}_t^* + B\boldsymbol{u}_t^* + C\boldsymbol{\lambda}_t^{\epsilon,\boldsymbol{\theta}} + \boldsymbol{d} - \boldsymbol{x}_{t+1}^*$ and $\boldsymbol{e}_t^{\text{lcp}} := \frac{1}{\epsilon\gamma}(D\boldsymbol{x}_t^* + E\boldsymbol{u}_t^* + F\boldsymbol{\lambda}_t^{\epsilon,\boldsymbol{\theta}} + \boldsymbol{c} - \phi_t^{\epsilon,\boldsymbol{\theta}})$.

Proof Claim (a) in Lemma 3 can be easily proved by verifying that the Hessian of the objective function in (9b) is positive definite.

In Claim (b), the differentiability of $L_\epsilon(\boldsymbol{\theta}, \mathcal{D})$ depends on the differentiability of $(\boldsymbol{\lambda}_t^{\epsilon,\boldsymbol{\theta}}, \phi_t^{\epsilon,\boldsymbol{\theta}})$ with respect to $\boldsymbol{\theta}$, $t = 1, 2, ..., N$. In fact, $(\boldsymbol{\lambda}_t^{\epsilon,\boldsymbol{\theta}}, \phi_t^{\epsilon,\boldsymbol{\theta}})$ is differentiable with respect to $\boldsymbol{\theta}$ if the strict

complementarity condition holds for constrained optimization in (9b). This is a direct result from the well-known sensitivity analysis theory (see Theorem 2.1 in (Fiacco, 1976)). The gradient of $L_\epsilon(\boldsymbol{\theta}, \mathcal{D})$ can be obtained directly applying the envelope theorem (Afriat, 1971). For example, the gradient of $l_\epsilon(\boldsymbol{\theta}, \boldsymbol{x}_t^*, \boldsymbol{u}_t^*, \boldsymbol{x}_{t+1}^*)$ with respect to matrix A is

$$\nabla_{\mathrm{vec}\,(A)} l_\epsilon = \left(\frac{dl_\epsilon}{d\,\mathrm{vec}\,(A)}\right)^\mathsf{T} = \left((\boldsymbol{e}_t^{\text{dyn}})^\mathsf{T}\left(\boldsymbol{x}_t^{*\mathsf{T}} \otimes I_{n_x}\right)\right)^\mathsf{T} = \left(\boldsymbol{x}_t^* \otimes I_{n_x}\right)\boldsymbol{e}_t^{\text{dyn}} = \mathrm{vec}(\boldsymbol{e}_t^{\text{dyn}}\boldsymbol{x}_t^{*\mathsf{T}}).$$

Writing the above into the matrix form leads to $\nabla_A L_\epsilon = \sum_{t=1}^N \boldsymbol{e}_t^{\text{dyn}}(\boldsymbol{x}_t^*)^\mathsf{T}$. Similar derivations also apply to $\nabla_B L_\epsilon$, $\nabla_B L_\epsilon$, $\nabla_C L_\epsilon$, $\nabla_D L_\epsilon$, $\nabla_E L_\epsilon$, and $\nabla_F L_\epsilon$. This completes the proof. ■

In addition to the strongly-convex quadratic problem with bound constraints in (9b), Lemma 3 state that $L_\epsilon(\boldsymbol{\theta}, \mathcal{D})$ allows for much simpler differentiation, as stated in claim (b). Note that differentiation of $L_\epsilon(\boldsymbol{\theta}, \mathcal{D})$ in (10) does not involve any matrix inverse. This is in stark contrast with the prediction-based loss (6), whose differentiation (de Avila Belbute-Peres et al., 2018) is based on the implicit function theorem (Rudin et al., 1976) and requires the inverse of Jacobian matrix of KKT equations (which is computationally expensive).

Another implication of Lemma 3 is that the Lipschitz constant of $L_\epsilon(\boldsymbol{\theta}, \mathcal{D})$ with respect to the LCP matrices $(D, E, F, \boldsymbol{c})$ can be controlled by the choice of ϵ. Specifically, the second line of (10) shows that one can always choose a large ϵ to produce a small Lipschitz constant of the loss landscape with respect to $(D, E, F, \boldsymbol{c})$. This property can facilitate the learning process by controlling the smoothness of the loss landscape, and also helpful in the generalization of learned results as analyzed in a concurrent work (Bianchini et al., 2022). However, we also need to note that the large choice of ϵ could lead to the bias learning results, as shown in the later examples.

We further have the following result, which states the second-order derivative of $L_\epsilon(\boldsymbol{\theta}, \mathcal{D})$.

Lemma 4 *Given F satisfying Assumption 1 and any constant $0 < \gamma < \sigma_{\min}(F^\mathsf{T} + F)$, suppose that the differentiability in Lemma 3 holds. Then,*

$$\nabla^2_{\boldsymbol{\theta}} L_\epsilon = \sum_{t=1}^N \left(\frac{\partial^2 L_\epsilon}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}} - \frac{\partial^2 L_\epsilon}{\partial\boldsymbol{\theta}\partial\boldsymbol{z}_t^\epsilon}\left(\mathrm{diag}\left(\frac{\partial L_\epsilon}{\partial \boldsymbol{z}_t^\epsilon}\right) + \mathrm{diag}(\boldsymbol{z}_t^\epsilon)\frac{\partial^2 L_\epsilon}{\partial \boldsymbol{z}_t^\epsilon \partial \boldsymbol{z}_t^\epsilon}\right)^{-1}\mathrm{diag}(\boldsymbol{z}_t)\frac{\partial^2 L_\epsilon}{\partial \boldsymbol{z}_t^\epsilon \partial\boldsymbol{\theta}}\right), \tag{11}$$

with $\boldsymbol{z}_t^\epsilon = (\boldsymbol{\lambda}_t^{\epsilon,\boldsymbol{\theta}}, \boldsymbol{\phi}_t^{\epsilon,\boldsymbol{\theta}})$ being the solution to (9b).

Proof To prove Lemma 4, we need first to show $\left(\mathrm{diag}\left(\frac{\partial L_\epsilon}{\partial \boldsymbol{z}_t^\epsilon}\right) + \mathrm{diag}(\boldsymbol{z}_t^\epsilon)\frac{\partial^2 L_\epsilon}{\partial \boldsymbol{z}_t^\epsilon \partial \boldsymbol{z}_t^\epsilon}\right)$ is invertible. We only provide the sketch of the proof due to page limits. First, the KKT conditions at $\boldsymbol{z}_t^\epsilon = (\boldsymbol{\lambda}_t^{\epsilon,\boldsymbol{\theta}}, \boldsymbol{\phi}_t^{\epsilon,\boldsymbol{\theta}})$ can be written as the following LCP

$$\mathbf{0} \le \left(\frac{\partial L_\epsilon}{\partial \boldsymbol{z}_t^\epsilon}\right)' \perp \boldsymbol{z}_t^\epsilon \ge \mathbf{0}. \tag{12}$$

The strict complementarity (differentiblity of L_ϵ) stated in claim (b) of Lemma 3 is equivalent to say the above LCP in (12) is strictly complementarity. By claim (a) in Lemma 3, we have known $\frac{\partial^2 L_\epsilon}{\partial \boldsymbol{z}_t^\epsilon \partial \boldsymbol{z}_t^\epsilon} \succ 0$, which is a P-matrix. Following the same proof in Appendix 6, one can show that $\left(\mathrm{diag}\left(\frac{\partial L_\epsilon}{\partial \boldsymbol{z}_t^\epsilon}\right) + \mathrm{diag}(\boldsymbol{z}_t^\epsilon)\frac{\partial^2 L_\epsilon}{\partial \boldsymbol{z}_t^\epsilon \partial \boldsymbol{z}_t^\epsilon}\right)$ is invertible.

Now we prove Lemma 4. By applying envelop theorem (Afriat, 1971) to $l_\epsilon(\boldsymbol{\theta}, \boldsymbol{x}_t^*, \boldsymbol{u}_t^*, \boldsymbol{x}_{t+1}^*)$ in (9b), one can write $\nabla_{\boldsymbol{\theta}} L_\epsilon = \left(\frac{\partial L_\epsilon}{\partial \boldsymbol{\theta}}\right)^\mathsf{T}$. When taking the second-order derivative, one has

$$\nabla^2_{\boldsymbol{\theta}} L_\epsilon = \sum_{t=1}^{N} \left(\frac{\partial^2 L_\epsilon}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}} + \frac{\partial^2 L_\epsilon}{\partial \boldsymbol{\theta} \partial \boldsymbol{z}_t^\epsilon} \frac{d \boldsymbol{z}_t^\epsilon}{d \boldsymbol{\theta}} \right). \tag{13}$$

Here, by differentiating through the LCP in (12), one can obtain

$$\frac{\partial \boldsymbol{z}_t^\epsilon}{\partial \boldsymbol{\theta}} = -\left(\operatorname{diag}\left(\frac{\partial L_\epsilon}{\partial \boldsymbol{z}_t^\epsilon}\right) + \operatorname{diag}(\boldsymbol{z}_t^\epsilon) \frac{\partial^2 L_\epsilon}{\partial \boldsymbol{z}_t^\epsilon \partial \boldsymbol{z}_t^\epsilon} \right)^{-1} \operatorname{diag}(\boldsymbol{z}_t^\epsilon) \frac{\partial^2 L_\epsilon}{\partial \boldsymbol{z}_t^\epsilon \partial \boldsymbol{\theta}}. \tag{14}$$

Plugging the above to (13) leads to (11). ■

Lemma 4 states that first, Hessian of the violation-based loss with respect to the system parameter $\boldsymbol{\theta}$ can also be analytically obtained. Such Hessian is important both for algorithmic implementation and theoretical analysis. Arithmetically, the above Hessian can be used to develop second-order methods for optimizing (4). Analytically, the Hessian can be used to analyze the convexity of the problem. Specifically, if (9b) is convex jointly with respect to $(\boldsymbol{\lambda}_t, \boldsymbol{\phi}_t)$ and $\boldsymbol{\theta}$, one can show that $L_\epsilon(\boldsymbol{\theta}, \mathcal{D})$ will be convex (also see Section 3.2.5 in (Boyd et al., 2004)). This holds for all other system matrices/vectors except matrices C and F, which imposes challenges for learning process.

Finally, we give the following result showing the violation-based loss $L_\epsilon(\boldsymbol{\theta}, \mathcal{D})$ in (9) can be controlled to approximate the prediction-based loss $L^{\text{pred}}(\boldsymbol{\theta}, \mathcal{D})$ (6) in terms of both the loss itself and its differentiability.

Lemma 5 *Given F satisfying Assumption 1 and any constant $0 < \gamma < \sigma_{\min}(F^\mathsf{T} + F)$, there exists $\Delta > 0$ such that for any $\epsilon \in (0, \Delta]$,*

(a) $L_\epsilon(\boldsymbol{\theta}, \mathcal{D})$ is differentiable (Lemma 3) if $L^{\text{pred}}(\boldsymbol{\theta}, \mathcal{D})$ is differentiable (Lemma 1).

(b) $L_\epsilon(\boldsymbol{\theta}, \mathcal{D}) \to L^{\text{pred}}(\boldsymbol{\theta}, \mathcal{D})$ as $\epsilon \to 0$.

Proof We here only provide the sketch for the proof of the above lemma due to page limits. In the proof of claim (a), first, we can show that the strict complementarity in Lemma (1) is equivalently to say the strict complementarity for (7) in Lemma 2, i.e., the following LCP

$$\mathbf{0} \leq \left(\frac{\partial f}{\partial \boldsymbol{z}_t}\right)' \perp \boldsymbol{z}_t \geq \mathbf{0} \quad \text{with} \quad \boldsymbol{z}_t := (\boldsymbol{\lambda}_t, \boldsymbol{\phi}_t) \quad \text{and} \quad f(\boldsymbol{\lambda}, \boldsymbol{\phi}) := \boldsymbol{\lambda}^\mathsf{T} \boldsymbol{\phi} + \frac{1}{\gamma} \|F\boldsymbol{\lambda} + \boldsymbol{q}_t - \boldsymbol{\phi}\|^2 \tag{15}$$

is strict complementarity. Further, one can show that (12) will converge to (15) as $\epsilon \to 0$. Since the strict complementarity preserves as ϵ falls in a small neighborhood around 0 (this is similar to the proof of Theorem 2.1 in (Fiacco, 1976)), one can say $L_\epsilon(\boldsymbol{\theta}, \mathcal{D})$ is differentiable with any small $\epsilon > 0$ in the neighborhood around 0. The proof of claim (b) can directly follow the standard proof in penalty-based optimization (Fiacco and McCormick, 1990). ■

The above lemma has shown that the proposed violation-based $L_\epsilon(\boldsymbol{\theta}, \mathcal{D})$ and prediction-based loss $L^{\text{pred}}(\boldsymbol{\theta}, \mathcal{D})$ are essentially related to each other in term of function itself and its differentiability with respect to $\boldsymbol{\theta}$. Specifically, the differentiability of prediction-based loss $L^{\text{pred}}(\boldsymbol{\theta}, \mathcal{D})$ in Lemma 1, i.e., strict complementarity for the LCP, always implies the differentiability of the violation-based

loss $L_\epsilon(\boldsymbol{\theta}, \mathcal{D})$ for any choice of small $\epsilon > 0$. Second, by controlling $\epsilon \to 0$, the violation-based formulation approximates to prediction-based one.

In light of all properties stated above, we now summarize the advantage of the proposed violation-based loss $L_\epsilon(\boldsymbol{\theta}, \mathcal{D})$ over the prediction-based loss $L^{\text{pred}}(\boldsymbol{\theta}, \mathcal{D})$. First, $L_\epsilon(\boldsymbol{\theta}, \mathcal{D})$ allows for more efficient computation of gradient, while differentiation of $L_{\text{pred}}(\boldsymbol{\theta}, \mathcal{D})$ always requires the matrix inverse. Second, $L_\epsilon(\boldsymbol{\theta}, \mathcal{D})$ permits analytical Hessian information, which is important both for the algorithmic implementation and theoretical analysis, such Hessian matrix is difficult to obtain for $L^{\text{pred}}(\boldsymbol{\theta}, \mathcal{D})$. Finally, the violation-based loss $L_\epsilon(\boldsymbol{\theta}, \mathcal{D})$ is flexible to approximate $L^{\text{pred}}(\boldsymbol{\theta}, \mathcal{D})$ in terms of both the function itself and its differentiability with respect to $\boldsymbol{\theta}$, by controlling ϵ.

5. Examples

In implementation, one way to enforce Assumption 1 is using re-parameterization tricks: by re-parameterizing $F = GG^\mathsf{T} + H - H^\mathsf{T}$ with any $G \in \mathbb{R}^{n_\lambda \times n_\lambda}$ and $H \in \mathbb{R}^{n_\lambda \times n_\lambda}$, one can easily see $F + F^\mathsf{T} \succeq 0$. Also note that F and C in (1) are permutation- and scaling- invariant with respect to $\mathcal{D}$: if $(\boldsymbol{x}_t^*, \boldsymbol{u}_t^*, \boldsymbol{x}_{t+1}^*)$ satisfies (1) with hidden $\boldsymbol{\lambda}_t^*$, it also satisfies the following LCS with $\tilde{\boldsymbol{\lambda}}_t^* = PS\boldsymbol{\lambda}_t$.

$$\begin{aligned} &\boldsymbol{x}_{t+1}^* = A\boldsymbol{x}_t^* + B\boldsymbol{u}_t^* + CS^{-1}P^\mathsf{T}\tilde{\boldsymbol{\lambda}}_t^* + \boldsymbol{d} \\ &\mathbf{0} \le \tilde{\boldsymbol{\lambda}}_t^* \perp PSD\boldsymbol{x}_t + PSE\boldsymbol{u}_t + PSFS^{-1}P^\mathsf{T}\tilde{\boldsymbol{\lambda}}_t^* + \boldsymbol{c} \ge \mathbf{0} \end{aligned} \tag{16}$$

for any permutation matrix $P \in \mathbb{R}^{n_\lambda \times n_\lambda}$ and any diagonal matrix $S \in \mathbb{R}^{n_\lambda \times n_\lambda}$ with positive entries. To mitigate this ambiguity, we add a regularizing cost $R(\boldsymbol{\theta}) = \omega \|C\|_F^2$, where $\|\cdot\|_F$ is the matrix Frobenius norm and ω is the weighting parameter. We set $\omega = 10^{-5}$ in our following experiments.

We randomly define a ground-truth LCS with $\boldsymbol{\theta}^*$, where all parameters are selected from a uniform distribution in range $[-1, 1]$. To generate training data $\mathcal{D} = \{(\boldsymbol{x}_t^*, \boldsymbol{u}_t^*, \boldsymbol{x}_{t+1}^*)\}_{t=1}^{N_{\text{train}}}$, we sample $\boldsymbol{x}_t^*$ and $\boldsymbol{u}_t^*$ from uniform distributions over $[-10, 10]$ and $[-5, 5]$, respectively, and then solve $\boldsymbol{x}_{t+1}^*$ based on $\boldsymbol{\theta}^*$. We also add zero-mean Gaussian noise with standard deviation $\sigma{=}10^{-2}$ to $\mathcal{D}$. We generate similar, but noiseless, testing data $\mathcal{T} = \{(\bar{\boldsymbol{x}}_t, \bar{\boldsymbol{u}}_t, \bar{\boldsymbol{x}}_{t+1})\}_{t=1}^{N_{\text{test}}}$. To evaluate the learned LCS on $\mathcal{T}$, we define the following *mean relative prediction error*,

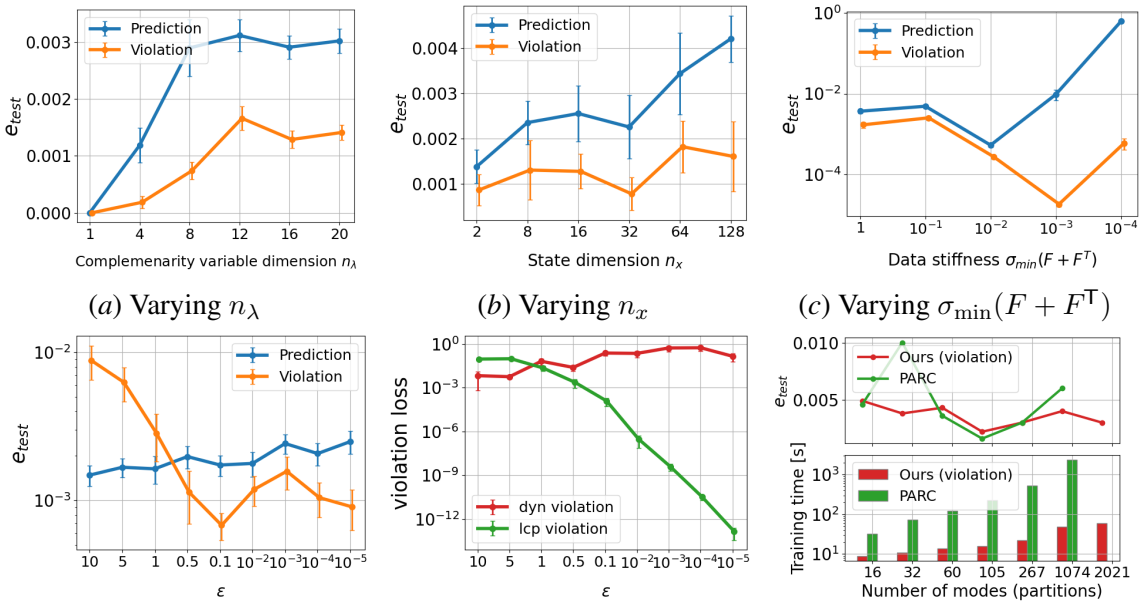
$$e_{\text{test}} = \frac{\sum_{t=1}^{N_{\text{test}}} \|\boldsymbol{x}_{t+1}^{\boldsymbol{\theta}} - \bar{\boldsymbol{x}}_{t+1}\|^2}{\sum_{t=1}^{N_{\text{test}}} \|\bar{\boldsymbol{x}}_{t+1}\|^2}, \tag{17}$$

with $\boldsymbol{x}_t^{\boldsymbol{\theta}}$ the state predicted by the learned LCS at $(\bar{\boldsymbol{x}}_t, \bar{\boldsymbol{u}}_t)$. The size of $\mathcal{T}$ is $N_{\text{test}}{=}1000$.

The following simulations evaluate different aspects of the proposed violation-based learning formulation (9), in comparison with the prediction-based learning formulation (6) and a state-of-the-art clustering-based PWA regression method (Bemporad, 2021). Each training case includes a total of 30 trials (otherwise stated), and each trial uses a random ground-truth LCS to generate $\mathcal{D}$ and $\mathcal{T}$ and randomly initializes $\boldsymbol{\theta}$. The training uses the Adam algorithm (Kingma and Ba, 2014) with mini-batch size 100 and learning rate 10^{-3} (other Adam parameters: $\beta_1 = 0.9$, $\beta_2 = 0.9$, $\epsilon = 10^{-6}$). Because the data is generated from random ground-truth LCS systems, and some systems may be easier to identify than others, we expect fairly high variance of the results. The implementation code of the following examples can be found at https://github.com/DAIRLab/Learning-LCS.

5.1. Results and Analysis

Comparisons between the proposed violation-based formulation (9) and the prediction-based one (6) are shown in Fig. 1(*a*)-1(*d*). Fig. 1(*e*) shows for the violation-based formulation, how different ϵ in (9b) impacts the final training LCP violation loss $\left(\boldsymbol{\lambda}_t^\mathsf{T}\boldsymbol{\phi}_t + \frac{1}{2\gamma}\|D\boldsymbol{x}_t^* + E\boldsymbol{u}_t^* + F\boldsymbol{\lambda}_t + \boldsymbol{c} - \boldsymbol{\phi}\|^2\right)$ and dynamics loss $\|A\boldsymbol{x}_t^* + B\boldsymbol{u}_t^* + C\boldsymbol{\lambda}_t + \boldsymbol{d} - \boldsymbol{x}_{t+1}^*\|^2$. Fig. 1(*f*) compares the proposed violation-based method with PARC (Bemporad, 2021), a state-of-the-art PWA regression method based on clustering. In summary of all evaluations, one can conclude that the proposed violation-based learning outperforms the prediction-based method, when handling high numbers of hybrid modes, e.g., 16k modes at $n_\lambda = 20$ in Fig. 1(*a*); dealing with high-dimension systems, e.g., $n_x = 128$ in Fig. 1(*b*); and identifying *high-stiffness systems* in Fig. 1(*c*). Although Fig. 1(*e*) suggests a smaller ϵ leads to smaller LCP violation in training, Fig. 1(*d*) shows that the prediction performance of the learned model is largely not influenced by any further smaller ϵ. Thus, it could be not difficult to find a reasonable small ϵ in the violation-based method for both benign training loss landscape and good prediction performance. Fig. 1(*f*) shows that both the clustering-based regression (Bemporad, 2021) and the proposed violation-based method achieve comparable prediction accuracy, but the proposed method demonstrates significant efficiency for handling large number of hybrid modes. Details for each evaluations are given below.



(*a*) Varying n_λ (*b*) Varying n_x (*c*) Varying $\sigma_{\min}(F + F^\mathsf{T})$

(*d*) Varying ϵ (*e*) Training violation losses in (9b) given different ϵ (*f*) Timeout (≥2h) for PARC at 2021 modes

Figure 1: Evaluations of the violation-based learning (9). Error bars indicate the standard errors.

In Fig. 1(*a*), we vary the number of complementarity constraints, i.e., the dimension of complementarity variable, n_λ, with fixed n_x=10 and n_u=4. $\mathcal{D}$ has the size of $N_{\text{train}} = 50$k, and $\sigma_{\min}(F+F^\mathsf{T}) = 1$. Note that the maximum achievable number of hybrid modes in $\mathcal{D}$ depends on both n_λ and n_x, and is 2^{n_λ} if $n_x = n_\lambda$. In our case, at $n_\lambda = 20$, $\mathcal{D}$ contains around 16k modes. In our violation-based loss formulation (9), we fix $\gamma = 10^{-2}$ and $\epsilon = 10^{-4}$.

In Fig. 1(*b*), we vary system state dimension n_x with fixed $n_\lambda = 10$ and $n_u = 4$. Here, each learning case includes 15 trials, and other settings follow the ones in Fig. 1(*a*). Fig. 1(*c*) varies the system stiffness indicated by $\sigma_{\min}(F+F^\mathsf{T})$, i.e., smaller $\sigma_{\min}(F+F^\mathsf{T})$ means a stiffer dynamics (Aydinoglu et al., 2021). Here, n_x=8, n_u=2, n_λ=10, and others follow the ones in Fig. 1(*a*).

Fig. 1(*e*) shows how varying ϵ will impact the final training loss of the dynamics prediction term and LPC violation term in (9b). Here, n_x=4, n_λ=4, n_u=2, N_{train}=5k, and other settings follow the ones in Fig. 1(*a*). In the same settings, Fig. 1(*d*) shows the prediction performance of the learned model, in comparison with the model learned by prediction-based method.

In Fig. 1(*f*), both the violation-based method and PARC (Bemporad, 2021) are given the same datasets with different number of hybrid modes (generated ground-truth LCS with different n_λ, fixed n_x=4, n_λ=4, n_u=2, other settings follow the ones in Fig. 1(*a*)). The upper panel compares the prediction error of the learned models using both methods, and bottom panel shows the training time (until convergence) for both methods.

6. Conclusion

We have proposed a violation-based loss formulation which enables to learn a LCS using gradient-based methods. The violation-based loss is a sum of dynamics prediction loss and a novel complementarity violation loss. We have shown several some properties of new formulation. The numerical results demonstrate a state-of-the-art ability to identify piecewise-affine dynamics, outperforming the state-of-the-art clustering-based regression and the methods which must differentiate through non-smooth linear complementarity problems.

Appendix: Proof of Lemma 1

We first prove if $\boldsymbol{\lambda}_t^* = \text{LCP}(F, D\boldsymbol{x}_t^* + E\boldsymbol{u}_t^* + \boldsymbol{c})$ is the strictly complementarity, then matrix $S_t := \text{diag}(D\boldsymbol{x}_t^* + E\boldsymbol{u}_t^* + F\boldsymbol{\lambda}_t^* + \boldsymbol{c}) + \text{diag}(\boldsymbol{\lambda}_t^*)F$ is invertible. We prove this by contradiction. Suppose S_t is singular, and there exists a non-zero $\boldsymbol{v} \in \mathbb{R}^{n_\lambda}$ s.t. $S_t^\mathsf{T}\boldsymbol{v} = \boldsymbol{0}$. Since F satisfying Assumption 1 is the P-matrix, so is F^T. Consider the two cases. If $\text{diag}(\boldsymbol{\lambda}_t^*)\boldsymbol{v} = \boldsymbol{0}$, thus $\text{diag}(D\boldsymbol{x}_t^* + E\boldsymbol{u}_t^* + F\boldsymbol{\lambda}_t^* + \boldsymbol{c})\boldsymbol{v} = \boldsymbol{0}$. There must exist $i \in \{1, ..., n_\lambda\}$ such that $\boldsymbol{\lambda}_t^*[i] = 0$ and $\boldsymbol{v}[i] \neq 0$. By the strict complementarity, $(D\boldsymbol{x}_t^* + E\boldsymbol{u}_t^* + F\boldsymbol{\lambda}_t^* + \boldsymbol{c})[i] \cdot \boldsymbol{v}[i] \neq 0$, which contradicts $\text{diag}(D\boldsymbol{x}_t^* + E\boldsymbol{u}_t^* + F\boldsymbol{\lambda}_t^* + \boldsymbol{c})\boldsymbol{v} = \boldsymbol{0}$. If $\text{diag}(\boldsymbol{\lambda}_t^*)\boldsymbol{v} \neq \boldsymbol{0}$, we have $F^\mathsf{T}\,\text{diag}(\boldsymbol{\lambda}_t^*)\boldsymbol{v} = -\,\text{diag}(D\boldsymbol{x}_t^* + E\boldsymbol{u}_t^* + F\boldsymbol{\lambda}_t^* + \boldsymbol{c})\boldsymbol{v}$. Then, for all $i \in \{1, 2, ..., n_\lambda\}$, $(\text{diag}(\boldsymbol{\lambda}_t^*)\boldsymbol{v})\,[i] \cdot \left(F^\mathsf{T}\,\text{diag}(\boldsymbol{\lambda}_t^*)\boldsymbol{v}\right)[i] = (\text{diag}(\boldsymbol{\lambda}_t^*)\boldsymbol{v})\,[i] \cdot (-\,\text{diag}(D\boldsymbol{x}_t^* + E\boldsymbol{u}_t^* + F\boldsymbol{\lambda}_t^* + \boldsymbol{c})\boldsymbol{v})\,[i] = 0$. In fact, since F^T is a P-matrix, the above result contradicts with the reverse-sign property of P-matrix (see Theorem 3.3.4 in (Cottle et al., 2009)). Combine the above two cases, we conclude that S_t is non-singular.

Next, we prove Lemma 1. Define $\boldsymbol{g}(\boldsymbol{\lambda}_t, D, E, F, \boldsymbol{c}) = \text{diag}(\boldsymbol{\lambda}_t)\,(D\boldsymbol{x}_t^* + E\boldsymbol{u}_t^* + F\boldsymbol{\lambda}_t + \boldsymbol{c}) = \boldsymbol{0}$. It is obvious that $\boldsymbol{\lambda}_t^* = \text{LCP}(F, D\boldsymbol{x}_t^* + E\boldsymbol{u}_t^* + \boldsymbol{c})$ satisfies the above equation. Next, we take the Jacobian Matrix of $\boldsymbol{g}(\boldsymbol{\lambda}_t, D, E, F, \boldsymbol{c})$ with respect to $\boldsymbol{\lambda}_t$ evaluated at $\boldsymbol{\lambda}^*$, leading to $\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{\lambda}_t}|_{\boldsymbol{\lambda}_t^*} = \text{diag}(D\boldsymbol{x}_t^* + E\boldsymbol{u}_t^* + F\boldsymbol{\lambda}_t^* + \boldsymbol{c}) + \text{diag}(\boldsymbol{\lambda}_t^*)F = S_t$. Since S_t is invertible due to the previous proof, by applying the implicit function theorem (Rudin et al., 1976), one can reach the differentiability in Lemma 1. This completes the proof. ■

Acknowledgments

Toyota Research Institute provided funds to support this work. This work was also supported by the National Science Foundation under Grant No. CMMI-1830218 and an NSF Graduate Research Fellowship under Grant No. DGE-1845298.

References

SN Afriat. Theory of maxima and the method of lagrange. *SIAM Journal on Applied Mathematics*, 20(3):343–357, 1971.

Mihai Anitescu and Gary D Hart. A constraint-stabilized time-stepping approach for rigid multibody dynamics with joints, contact and friction. *International Journal for Numerical Methods in Engineering*, 60(14):2335–2371, 2004.

Alp Aydinoglu and Michael Posa. Real-time multi-contact model predictive control via admm. In *International Conference on Robotics and Automation*, 2022.

Alp Aydinoglu, Philip Sieg, Victor M Preciado, and Michael Posa. Stabilization of complementarity systems via contact-aware controllers. *IEEE Transactions on Robotics*, 2021.

Laurent Bako, Khaled Boukharouba, Eric Duviella, and Stéphane Lecoeuche. A recursive identification algorithm for switched linear/affine models. *Nonlinear Analysis: Hybrid Systems*, 5(2): 242–253, 2011.

Peter W Battaglia, Razvan Pascanu, Matthew Lai, Danilo Rezende, and Koray Kavukcuoglu. Interaction networks for learning about objects, relations and physics. *Advances in neural information processing systems*, 2016.

Alberto Bemporad. Piecewise linear regression and classification. *arXiv preprint arXiv:2103.06189*, 2021.

Alberto Bemporad and Manfred Morari. Control of systems integrating logic, dynamics, and constraints. *Automatica*, 35(3):407–427, 1999.

Alberto Bemporad, Francesco Borrelli, and Manfred Morari. Piecewise linear optimal controllers for hybrid systems. In *American Control Conference*, volume 2, pages 1190–1194. IEEE, 2000.

Alberto Bemporad, Andrea Garulli, Simone Paoletti, and Antonio Vicino. A bounded-error approach to piecewise affine system identification. *IEEE Transactions on Automatic Control*, 50 (10):1567–1580, 2005.

Bibit Bianchini, Mathew Halm, Nikolai Matni, and Michael Posa. Generalization bounds for implicit learning of nearly discontinuous functions. In *Annual Learning for Dynamics & Control Conference*, 2022.

Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Leo Breiman. Hinging hyperplanes for regression, classification, and function approximation. *IEEE Transactions on Information Theory*, 39(3):999–1013, 1993.

Valentina Breschi, Dario Piga, and Alberto Bemporad. Piecewise affine regression via recursive multiple least squares and multicategory discrimination. *Automatica*, 73:155–162, 2016.

Bernard Brogliato. *Nonsmooth mechanics*. Springer, 1999.

M Kanat Camlibel, Jong-Shi Pang, and Jinglai Shen. Lyapunov stability of complementarity and extended systems. *SIAM Journal on Optimization*, 17(4):1056–1101, 2007.

Alejandro Castro, Frank Permenter, and Xuchen Han. An unconstrained convex formulation of compliant contact. *arXiv preprint arXiv:2110.10107*, 2021.

Richard W Cottle, Jong-Shi Pang, and Richard E Stone. *The linear complementarity problem*. SIAM, 2009.

Filipe de Avila Belbute-Peres, Kevin Smith, Kelsey Allen, Josh Tenenbaum, and J Zico Kolter. End-to-end differentiable physics for learning and control. *Advances in neural information processing systems*, 31:7178–7189, 2018.

Danny Driess, Jung-Su Ha, Marc Toussaint, and Russ Tedrake. Learning models as functionals of signed-distance fields for manipulation planning. In *Conference on Robot Learning*, 2021.

Yingwei Du, Fangzhou Liu, Jianbin Qiu, and Martin Buss. A semi-supervised learning approach for identification of piecewise affine systems. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 67(10):3521–3532, 2020.

Ehsan Elhamifar, Samuel A Burden, and S Shankar Sastry. Adaptive piecewise–affine inverse modeling of hybrid dynamical systems. *IFAC Proceedings Volumes*, 47(3):10844–10849, 2014.

Giancarlo Ferrari-Trecate, Marco Muselli, Diego Liberati, and Manfred Morari. A clustering technique for the identification of piecewise affine systems. *Automatica*, 39(2):205–217, 2003.

Anthony V Fiacco. Sensitivity analysis for nonlinear programming using penalty methods. *Mathematical programming*, 10(1):287–311, 1976.

Anthony V Fiacco and Garth P McCormick. *Nonlinear programming: sequential unconstrained minimization techniques*. SIAM, 1990.

Moritz Geilinger, David Hahn, Jonas Zehnder, Moritz Bächer, Bernhard Thomaszewski, and Stelian Coros. Add: analytically differentiable dynamics for multi-body systems with frictional contact. *ACM Transactions on Graphics*, 39(6):1–15, 2020.

András Hartmann, João M Lemos, Rafael S Costa, João Xavier, and Susana Vinga. Identification of switched arx models via convex optimization and expectation maximization. *Journal of Process Control*, 28:9–16, 2015.

Wilhemus PMH Heemels, Bart De Schutter, and Alberto Bemporad. Equivalence of hybrid dynamical models. *Automatica*, 37(7):1085–1091, 2001.

WPMH Heemels, Johannes M Schumacher, and S Weiland. Linear complementarity systems. *SIAM journal on applied mathematics*, 60(4):1234–1269, 2000.

WPMH Heemels, M Kanat Camlibel, and Johannes M Schumacher. On the dynamic analysis of piecewise-linear networks. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 49(3):315–327, 2002.

Wanxin Jin, Shaoshuai Mou, and George J Pappas. Safe pontryagin differentiable programming. *Advances in Neural Information Processing Systems*, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Fabien Lauer. On the complexity of piecewise affine system identification. *Automatica*, 62:148–153, 2015.

Yunzhu Li, Jiajun Wu, Russ Tedrake, Joshua B Tenenbaum, and Antonio Torralba. Learning particle dynamics for manipulating rigid bodies, deformable objects, and fluids. In *International Conference on Learning Representations*, 2019.

J-N Lin and Rolf Unbehauen. Canonical piecewise-linear approximations. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 39(8):697–699, 1992.

Hayato Nakada, Kiyotsugu Takaba, and Tohru Katayama. Identification of piecewise affine systems based on statistical clustering technique. *Automatica*, 41(5):905–913, 2005.

Simone Paoletti, Aleksandar Lj Juloski, Giancarlo Ferrari-Trecate, and René Vidal. Identification of hybrid systems a tutorial. In *European journal of control*, volume 13, pages 242–260. Elsevier, 2007.

Mihir Parmar, Mathew Halm, and Michael Posa. Fundamental challenges in deep learning for stiff contact dynamics. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2021.

Samuel Pfrommer, Mathew Halm, and Michael Posa. Contactnets: Learning discontinuous contact dynamics with smooth, implicit representations. In *Conference on Robot Learning*, 2020.

Jacob Roll, Alberto Bemporad, and Lennart Ljung. Identification of piecewise affine systems via mixed-integer programming. *Automatica*, 40(1):37–50, 2004.

Walter Rudin et al. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1976.

David Stewart and Jeffrey C Trinkle. An implicit time-stepping scheme for rigid body dynamics with coulomb friction. In *IEEE International Conference on Robotics and Automation*, volume 1, pages 162–169. IEEE, 2000.

Michael J Tsatsomeros. Generating and detecting matrices with positive principal minors. *Asian Information-Science-Life: An International Journal*, 1(2):115–132, 2002.