# Model Invalidation for Switched Affine Systems with Applications to Fault and Anomaly Detection

**Farshad Harirchi and Necmiye Ozay**

*Electrical Engineering and Computer Science Department*
*University of Michigan, Ann Arbor, MI 48109 USA*
*(e-mail: harirchi,necmiye@umich.edu).*

**Abstract:** In this paper, the model (in)validation problem is addressed for the class of switched state space models. We pose the model invalidation problem as a mixed-integer linear program and solve it using the state-of-the-art MILP solvers. Model invalidation is mainly utilized to build trust in the models obtained from system identification. However, we turn our attention to solve another important class of problems using model invalidation approach proposed in this paper. It is shown that the model invalidation approach can be utilized to detect any general fault in cyber-physical systems. Moreover, it is illustrated that knowing the fault model can reduce the complexity of fault detection approach proposed here, if the fault and system model satisfy certain conditions.

*Keywords:* model invalidation, switched systems, fault detection, radiant systems.

## 1. INTRODUCTION

Cyber-physical systems are combinations of physical processes and embedded computers that collect data from these processes through sensors and control these processes in a closed loop manner. With the increase in data acquisition and storage capacity and the decrease in sensor costs, it is possible to collect large amounts of data during the operation of complex cyber-physical systems. For instance, "a four-engine jumbo jet can create 640 terabytes of data in just one crossing of the Atlantic Ocean" Rajah (2014). As discussed in Sznaier et al. (2014), this exponential growth in the data collection capabilities is a major challenge for systems and control community. Sensor-/information-rich networked cyber-physical systems, from air traffic or energy networks to smart buildings, are getting tightly integrated into our daily lives. As such, their safety-criticality increases. For such systems, it is crucial to detect faults or anomalies in real-time to support the decision-making process and to prevent potential large-scale failures. It is also important to obtain accurate models for these systems that can be used both for control design and later for monitoring the system.

Switched affine state-space models provide a convenient means to model many cyber-physical systems. In this paper, we consider two problems related to switched affine models: (i) model invalidation; (ii) fault and anomaly detection. In model invalidation problem, one starts with a family of models (i.e., a priori or admissible model set) and experimental input/output data collected from a system (i.e., a finite execution trace) and tries to determine whether the experimental data can be generated by one of the models in the initial model family. It was originally proposed as a way to build trust in models obtained through a system identification step or discard/improve them before using these models in robust control design Smith and Doyle (1992). There are more recent results applying model invalidation ideas for nonlinear systems Prajna (2006), and switched auto-regressive models Cheng et al. (2012); Ozay et al. (2010, 2014). Moreover, it is demonstrated via examples in Ozay et al. (2014) that model invalidation algorithms can be used for anomaly detection, where anomaly is roughly treated as anything that cannot be explained by the a priori model set. In this paper, we formalize the connection between anomaly/fault detection and model invalidation. Moreover, we show how model invalidation algorithms can be used in a receding horizon manner when fault models exist.

Fault detection techniques are developed in different communities Miljkovic (2011). A category of fault detection approaches is signal and data processing based, which utilizes techniques from pattern recognition Diallo et al. (2005) and spectrum analysis Isermann (2005, 2006) or simple algorithms of trend or limit checking of the signal Verron et al. (2010). Fault detection in control community has been investigated from the process model perspective. The methods developed in this community are based on residual generation and evaluation (see e.g., Chow and Willsky (1984); De Persis and Isidori (2001); Frank and Ding (1997)). Another set of approaches is based on designing the state or output observers and using the estimation error, or innovation, as the residual for the detection of the fault Frank (1990). Parameter estimation techniques have been also utilized as residual generators where the difference between physical parameters of the system and the estimated parameters from data are used as residuals Venkatasubramanian et al. (2003). Finally, utilizing neural networks for black box modeling of static or dynamical

⋆ This work is supported in part by DARPA grant N66001-14-1-4045.

systems and comparing the output of that model with the experimental data is another approach for fault detection Isermann (2006). In this work, we propose a new approach to fault detection from a controls perspective, which is based on model invalidation techniques for switched state-space models.

In Ozay et al. (2014), a model invalidation approach for switched auto-regressive models is proposed based on polynomial optimization and relaxation technique. Unavailability of switching sequence for measurement and noise in the input/output measurements render the model invalidation for switched systems challenging Ozay et al. (2014). The model invalidation problem setup we consider in the present paper is closely related to that in Ozay et al. (2014), however there are two important differences. First, we consider switched state-space models as opposed to switched autoregressive models. Arguably, when modeling a system using first principles, state-space representation is quite natural. Therefore, state-space models are more commonly used for modeling cyber-physical systems and developing model invalidation techniques for this class of systems is important. Second, we proceed by recasting the model invalidation problem as the feasibility check of a mixed integer linear programming (MILP) problem as opposed to the convex relaxation approach. Although from a complexity point of view convex relaxations are appealing, we experimentally observed that state-of-the-art MILP solvers (e.g., CPLEX (2009)) perform favorably on average without the need for a relaxation.

### *1.1 Contributions and Structure*

The contributions of this paper are: (i) to propose a model invalidation approach for the class of switched affine (SWA) models, (ii) to utilize model invalidation as a tool for anomaly and fault detection in cyber-physical systems, and (iii) to apply proposed method on the fault detection of radiant systems in smart buildings.

The structure of the paper is as follows: The proposed approach for model invalidation of switched affine models is discussed in Section 2. In Section 3, model-based fault and anomaly detection as well as the relation to model invalidation is described. Finally, academic and practical examples are provided in Section 4. The paper, then is concluded with discussion and future directions.

### *1.2 Notation*

Let $\mathbf{x} \in \mathbb{R}^n$ denote a vector and $\mathbf{x}^i$ indicate its $i$th element. Also, let $\mathbf{M} \in \mathbb{R}^{n\times m}$ represent a matrix and $\mathbf{M}^{i,j}$ indicate the element on $i$th row and $j$th column of the matrix $\mathbf{M}$. $I_n$ denotes the identity matrix of size $n$. The infinity norm of a vector $\mathbf{x}$ is denoted by $\|\mathbf{x}\|_\infty \doteq \max_i \mathbf{x}^i$. The set of positive integers up to $n$ is denoted by $\mathbb{Z}_n^+$, and the set of non-negative integers up to $n$ is denoted by $\mathbb{Z}_n^0$.

## 2. SWA MODEL (IN)VALIDATION

In this section, we state the invalidation problem, and provide a tractable approach to solve it.

### *2.1 Problem Definition*

We consider switched affine (SWA) systems of the form:

$$G = (\mathcal{X}, \mathcal{E}, \mathcal{U}, \{G_i\}_{i=1}^s) \tag{1}$$

where $\mathcal{X} \subset \mathbb{R}^n$ is the set of states, $\mathcal{E} \subset \mathbb{R}^{n_y}$ is the set of measurement noise values, $\mathcal{U} \subset \mathbb{R}^{n_u}$ is the set of inputs and $\{G_i\}_{i=1}^s$ is a collection of $s$ modes where for all $i \in \mathbb{Z}_s^+$ , the $i^{th}$ mode is an affine model $G_i = (\mathbf{A}_i, \mathbf{B}_i, \mathbf{f}_i, \mathbf{C}_i, \mathbf{D}_i)$.The evolution of $G$ is governed by:

$$\begin{aligned} \mathbf{x}(k+1) &= \mathbf{A}_{\sigma(k)}\mathbf{x}(k) + \mathbf{B}_{\sigma(k)}\mathbf{u}(k) + \mathbf{f}_{\sigma(k)}, \\ \mathbf{y}(k) &= \mathbf{C}_{\sigma(k)}\mathbf{x}(k) + \mathbf{D}_{\sigma(k)}\mathbf{u}(k) + \boldsymbol{\eta}(k), \end{aligned} \tag{2}$$

where $\mathbf{x}(k) \in \mathcal{X}$ is the state, $\mathbf{u}(k) \in \mathcal{U}$ is the control input, $\mathbf{y}(k) \in \mathbb{R}^{n_y}$ is the output, and $\boldsymbol{\eta}(k) \in \mathcal{E}$ is the measurement noise at time $k$. Here, $\sigma(k) \in \mathbb{Z}_s^+$ indicates the active mode at time $k$, that is, if $\sigma(k) = i$ the state evolves with respect to the dynamics of $G_i$. Throughout the paper we take $\mathcal{X}$, $\mathcal{E}$ and $\mathcal{U}$ to be infinity norm balls. That is, we let $\mathcal{X} = \{\mathbf{x} \mid \|\mathbf{x}\|_\infty \le M\}$, $\mathcal{E} = \{\boldsymbol{\eta} \mid \|\boldsymbol{\eta}\|_\infty \le \epsilon\}$ and $\mathcal{U} = \{\mathbf{u} \mid \|\mathbf{u}\|_\infty \le U\}$, where $M, \epsilon$ and $U$ are given constants. Usually physical constraints on the system impose the bounds $M$ and $U$. If no such bound is known, they can be taken to be infinite. On the other hand, the bound $\epsilon$ on the measurement noise value is based on the accuracy of the sensors and always assumed to be finite.

*Remark 1.* We do not consider process noise in the SWA system defined above, but the results in this paper can be extended to the systems with process noise, simply by adding variables to the problem.

In order to state the model invalidation problem, we first define the behavior of an SWA system.

*Definition 1.* The *$N$-truncated behavior* associated with an SWA system $G$ is the set of all length-$N+1$ input-output trajectories compatible with $G$, given by the set

$$\begin{aligned} \mathcal{B}_{swa}^N(G) := \big\{ \{\mathbf{u}(k), \mathbf{y}(k)\}_{k=0}^N, \mid \mathbf{u}(k) \in \mathcal{U} \text{ and } \exists \mathbf{x}(k) \in \mathcal{X}, \\ \sigma(k) \in \mathbb{Z}_s^+, \boldsymbol{\eta}(k) \in \mathcal{E} \text{ for } k = 0, \dots, N \text{ s.t. (2) holds} \big\}. \end{aligned}$$

With slight abuse of terminology, we will call $\mathcal{B}_{swa}^N(G)$ just the *behavior* of the system $G$.

Now we can state the model invalidation problem for SWA systems. Roughly speaking, given an input-output data sequence and a switched affine model, model invalidation problem is to determine whether or not the data is compatible with the model. This can be formally stated in terms of behaviors as follows:

*Problem 1.* Given $\big\{\mathbf{u}(k), \mathbf{y}(k)\big\}_{k=0}^N$, an input-output sequence, and a switched affine model $G$, determine whether or not the input-output sequence is contained in the behavior of $G$. That is, whether or not the following is true

$$\big\{\mathbf{u}(k), \mathbf{y}(k)\big\}_{k=0}^N \in \mathcal{B}_{swa}^N(G). \tag{3}$$

Next we define a series of feasibility problems that are equivalent to Problem 1. To this effect, start by noting that if at time $k$, mode $i$ is active, then the following system of equations is true:

$$\begin{cases} \mathbf{x}^1(k+1) - \mathbf{A}_i^{1,1}\mathbf{x}^1(k) - \ldots - \mathbf{A}_i^{1,n}\mathbf{x}^n(k) - \\ \quad \mathbf{B}_i^{1,1}\mathbf{u}^1(k) - \ldots \mathbf{B}_i^{1,n_u}\mathbf{u}^{n_u}(k) - \mathbf{f}_i^1 = 0 \\ \vdots \\ \mathbf{x}^n(k+1) - \mathbf{A}_i^{n,1}\mathbf{x}^1(k) - \ldots - \mathbf{A}_i^{n,n}\mathbf{x}^n(k) - \\ \quad \mathbf{B}_i^{n,1}\mathbf{u}^1(k) - \ldots \mathbf{B}_i^{n,n_u}\mathbf{u}^{n_u}(k) - \mathbf{f}_i^n = 0 \\ \mathbf{y}^1(k) - \boldsymbol{\eta}^1(k) - \mathbf{C}_i^{1,1}\mathbf{x}^1(k) - \ldots - \\ \quad \mathbf{C}_i^{1,n}\mathbf{x}^n(k) - \mathbf{D}_i^{1,1}\mathbf{u}^1(k) - \ldots - \mathbf{D}_i^{1,n_u}\mathbf{u}^{n_u}(k) = 0 \\ \vdots \\ \mathbf{y}^{n_y}(k) - \boldsymbol{\eta}^{n_y}(k) - \mathbf{C}_i^{n_y,1}\mathbf{x}^1(k) - \ldots - \\ \quad \mathbf{C}_i^{n_y,n}\mathbf{x}^n(k) - \mathbf{D}_i^{n_y,1}\mathbf{u}^1(k) - \ldots - \mathbf{D}_i^{n_y,n_u}\mathbf{u}^{n_u}(k) = 0 \end{cases} \tag{4}$$

The system of equations above has $n_y + n$ equations linear in the variables $\mathbf{x}(k:k+1)$ and $\boldsymbol{\eta}(k)$. Let us write the state equations in (4) in short as $\mathbf{h}_i^{(j)}\mathbf{x}(k:k+1) - l_{i,k} = 0$, and the output equations as $\mathbf{g}_i^{(j)}[\mathbf{x}(k+1);\boldsymbol{\eta}(k)] - q_{i,k} = 0$, where $l_{i,k}$ and $q_{i,k}$ represent the constant time-dependent terms that are obtained from input-output sequence and $\mathbf{h}_i^{(j)}, \mathbf{g}_i^{(j)}$ denote the coefficients of variables in the $j$th state and output equation.

*Proposition 1.* Given a SWA system $G$, and an input-output sequence $\{\mathbf{u}(k), \mathbf{y}(k)\}_{k=0}^{T}$, consider the following feasibility problem:

$$\begin{aligned} &\text{Find } \mathbf{x}(k), \boldsymbol{\eta}(k), a_{i,k} \text{ for } k \in \mathbb{Z}_T^0, i \in \mathbb{Z}_s^+ \\ &\text{s.t. } a_{i,k}(\mathbf{h}_i^{(j)}\mathbf{x}(k:k+1) - l_{i,k}) = 0 \ \forall j \in \mathbb{Z}_n^+, \ k \in \mathbb{Z}_T^0 \\ &\qquad a_{i,k}(\mathbf{g}_i^{(j')}[\mathbf{x}(k+1);\boldsymbol{\eta}(k)] - q_{i,k}) = 0 \ \forall j' \in \mathbb{Z}_{n_y}^+, \\ &\qquad\qquad k \in \mathbb{Z}_T^0 \\ &\qquad \sum_{i=1}^{s} a_{i,k} = 1, \text{ and } a_{i,k} \in \{0,1\}, \ \forall i \in \mathbb{Z}_s^+, \ k \in \mathbb{Z}_T^0 \\ &\qquad \|\boldsymbol{\eta}(k)\|_\infty \le \epsilon, \text{ and } \|\mathbf{x}(k)\|_\infty \le M, \ \forall k \in \mathbb{Z}_T^0 \\ &\qquad \|\mathbf{u}(k)\|_\infty \le U, \ \forall k \in \mathbb{Z}_T^0. \end{aligned} \tag{5}$$

The problem (5) is feasible if and only if (3) is satisfied.

Feasibility problem (5) is equivalent to the model invalidation problem, however, the formulation contains bilinear terms, and is not amenable to tractable solution procedures. Proceeding as in Cheng et al. (2012), we define the following auxiliary variables for all $i \in \mathbb{Z}_s^+$ and $k \in \mathbb{Z}_T^0$ to eliminate the bilinear terms in (5):

$$\begin{aligned} \boldsymbol{\eta}_i(k) &= a_{i,k}\boldsymbol{\eta}(k), \\ \mathbf{x}_i(k) &= a_{i,k}\mathbf{x}(k). \end{aligned} \tag{6}$$

Given a SWA system $G$, and an input-output sequence $\{\mathbf{u}(k), \mathbf{y}(k)\}_{k=k_1}^{k_2}$, let $\text{Feas}_G(\{\mathbf{u}(k), \mathbf{y}(k)\}_{k=k_1}^{k_2})$ denote the following feasibility problem:

$$\begin{aligned} &\text{Find } \mathbf{x}_i(k), \boldsymbol{\eta}_i(k), \mathbf{x}(k), \boldsymbol{\eta}(k), a_{i,k} \text{ for } k \in [k_1, k_2], i \in \mathbb{Z}_s^+ \\ &\text{s.t. } \mathbf{h}_i^{(j)}\mathbf{x}_i(k:k+1) - a_{i,k}l_{i,k} = 0 \ \forall j \in \mathbb{Z}_n^+, \ \forall k \in [k_1, k_2] \\ &\qquad \mathbf{g}_i^{(j')}[\mathbf{x}_i(k+1);\boldsymbol{\eta}_i(k)] - a_{i,k}q_{i,k}) = 0 \ \forall j' \in \mathbb{Z}_{n_y}^+, \\ &\qquad\qquad k \in [k_1, k_2] \\ &\qquad \sum_{i=1}^{s} a_{i,k} = 1, a_{i,k} \in \{0,1\}, \ \forall i \in \mathbb{Z}_s^+, \ k \in [k_1, k_2] \\ &\qquad \|\boldsymbol{\eta}_i(k)\|_\infty \le a_{i,k}\epsilon, \|\mathbf{x}(k)\|_\infty \le a_{i,k}M, \forall i \in \mathbb{Z}_s^+, \\ &\qquad\qquad \forall k \in [k_1, k_2] \\ &\qquad \|\mathbf{u}(k)\|_\infty \le U, \ \forall k \in [k_1, k_2] \\ &\qquad \sum_{i=1}^{s} \mathbf{x}_i(k) = \mathbf{x}(k), \forall i \in \mathbb{Z}_s^+, \ \forall k \in [k_1, k_2]. \end{aligned} \tag{7}$$

The feasibility problem, $\text{Feas}_G(\{\mathbf{u}(k), \mathbf{y}(k)\}_{k=k_1}^{k_2})$, is a mixed-integer linear program. Although its worst-case complexity is exponential, it can be solved relatively efficiently in practice using state-of-the-art solvers such as CPLEX CPLEX (2009).

*Proposition 2.* The MILP problem $\text{Feas}_G(\{\mathbf{u}(k), \mathbf{y}(k)\}_{k=0}^{T})$ is feasible if and only if problem (5) is feasible.

The two propositions above indicate that the model invalidation problem can be solved by checking the feasibility of $\text{Feas}_G(\{\mathbf{u}(k), \mathbf{y}(k)\}_{k=0}^{T})$.

## 3. FAULT AND ANOMALY DETECTION VIA MODEL INVALIDATION

In this section, we present two applications of the model invalidation framework in anomaly and fault detection. Let us first define what we mean by anomaly and fault.

*Definition 2.* An input/output sequence $\{\mathbf{u}(k), \mathbf{y}(k)\}_{k=0}^{N}$ is called *abnormal* for a switched system $G$ if and only if $\{\mathbf{u}(k), \mathbf{y}(k)\}_{k=0}^{N} \notin \mathcal{B}_{swa}^{N}(G)$.

With this definition, it is clear that detecting an abnormal behavior is equivalent to model being invalid. Therefore, the approach developed in the previous section can be readily applied for detecting anomalies in a system. On the other hand, we associate faults with fault models.

*Definition 3.* A *fault model* for a switched system $G = (\mathcal{X}, \mathcal{E}, \mathcal{U}, \{G_i\}_{i=1}^{s})$ is another switched system $G^f = (\mathcal{X}^f, \mathcal{E}^f, \mathcal{U}^f, \{G_i^f\}_{i=1}^{s^f})$ with the same number of states, inputs and outputs.

Note that with the definitions above, anomaly detection does not require explicit models for failure modes. Given that a cyber-physical system can fail (or be attacked) in infinitely many different ways, not needing to model these failure modes is advantageous. On the other hand, if one has certain models for the failures as in Def. 3 and is interested in detecting them, it might be possible to develop more efficient fault monitoring mechanisms. In what follows, we focus on persistent faults in the sense that once a fault occurs, the system starts evolving according to $G^f$.

*Definition 4.* A fault model $G^f$ for a switched system $G$ is called *$T$-step detectable* if $\mathcal{B}_{swa}^{N}(G) \cap \mathcal{B}_{swa}^{N}(G^f) = \emptyset$ for all $N \ge T$, where $T$ is a positive integer.

It is clear from the definition that if a fault model is $T$-step detectable, it is also $T'$-step detectable for all $T' \geq T$.

*Proposition 3.* Given a $T$-step detectable fault model $G^f$ for a switched system $G$, it is possible to detect the existence of a persistent fault with the given fault model by checking, at each time $k$, if $\text{Feas}_G(\{\mathbf{u}(j), \mathbf{y}(j)\}_{j=k-T}^{k})$ is feasible or not.

**Proof.** Let a fault occurs at time $i^*$, that is the input/output sequence $\{\mathbf{u}(j), \mathbf{y}(j)\}_{j \geq i^*}$ is generated by the fault model $G^f$. Because $G^f$ is $T$-step detectable, by Def. 4, there exists $k^* \leq i^* + T$ such that $\{\mathbf{u}(j), \mathbf{y}(j)\}_{j=i^*}^{k^*} \notin \mathcal{B}_{swa}^{k^*-i^*}(G)$. By Proposition 2, this is equivalent to the existence of $k^* \leq i^* + T$ such that $\text{Feas}_G(\{\mathbf{u}(j), \mathbf{y}(j)\}_{j=i^*}^{k^*})$ is infeasible. Since $[i^*, k^*] \subseteq [k^* - T, k^*]$, the infeasibility of $\text{Feas}_G(\{\mathbf{u}(j), \mathbf{y}(j)\}_{j=i^*}^{k^*})$ implies the infeasibility of $\text{Feas}_G(\{\mathbf{u}(j), \mathbf{y}(j)\}_{j=k^*-T}^{k^*})$.

For general anomaly detection, in theory we need to solve optimization problems with a growing size since all the data is coupled through the dynamics. However, Proposition 3 tells us that when the goal is to detect a specific fault, the size of the optimization problem solved at each time step can be fixed. Now the question is for a given fault model how to compute a $T$ (if it exists) such that the fault model is $T$-step detectable. We address this question next.

Our first step is to encode the behavioral definition of $T$-step detectability given in Def. 4. We eliminate the dependence on the mode signal $\sigma$ using propositional formulas. If at time $k$, mode $i$ is active, we know that the set of equations in (4) holds, that is the states $\mathbf{x}(k : k+1)$ and noise $\boldsymbol{\eta}(k)$ satisfy:

$$\varphi_{k,i}(\mathbf{x}(k:k+1), \boldsymbol{\eta}(k)) \doteq \left[\bigwedge_{l=1}^{n} [\mathbf{h}_i^{(j)} \mathbf{x}(k:k+1) - l_{i,k} = 0]\right] \wedge \left[\bigwedge_{l=1}^{n_y} [\mathbf{g}_i^{(j)}[\mathbf{x}(k+1); \boldsymbol{\eta}(k)] - q_{i,k} = 0]\right] \tag{8}$$

where $\wedge$ indicates the logical AND operation. At each time $k$, at least one mode is active, that is,

$$\bigvee_{i=1}^{s} \varphi_{k,i}(\mathbf{x}(k:k+1), \boldsymbol{\eta}(k)) \tag{9}$$

is satisfied by the states and noise values, where $\vee$ indicates the logical OR operation. Now, we can define a consistency set that is independent of the mode signal.

*Definition 5.* Let $\{\mathbf{u}(k), \mathbf{y}(k)\}_{k=k_1}^{k_2}$ be an input-output sequence over a time window $[k_1, k_2]$. The *consistency set*, $\mathcal{T}_G(\{\mathbf{u}(k), \mathbf{y}(k)\}_{k_1}^{k_2})$, is defined as follows:

$$\mathcal{T}_G(\{\mathbf{u}(k), \mathbf{y}(k)\}_{k=k_1}^{k_2}) = \Big\{ (\mathbf{x}(k_1:k_2), \boldsymbol{\eta}(k_1:k_2)) \mid \Big(\bigvee_{i=1}^{s} \varphi_{k,i}(\mathbf{x}(k:k+1), \boldsymbol{\eta}(k)) \text{ is TRUE}\Big), \|\mathbf{x}(k)\|_\infty \leq M, \|\boldsymbol{\eta}(k))\|_\infty \leq \epsilon, \|\mathbf{u}(k)\|_\infty \leq U, \forall k \in [k_1, k_2] \Big\}. \tag{10}$$

The model invalidation problem can be equivalently stated in terms of the consistency set as follows.

*Lemma 1.* The model invalidation problem is equivalent to checking the emptiness of the consistency set. That is, $\mathcal{T}_G(\{\mathbf{u}(k), \mathbf{y}(k)\}_{k=0}^{N}) \neq \emptyset$ if and only if $\{\mathbf{u}(k), \mathbf{y}(k)\}_{k=0}^{N} \in \mathcal{B}_{swa}^{N}(G)$.

**Proof.** Follows directly from the definitions.

Next, a characterization of $T$-step detectability in terms of consistency sets is given.

*Lemma 2.* A fault model $G^f$ for a switched system $G$ is $T$-step detectable if and only if for all $\{\mathbf{u}(k), \mathbf{y}(k)\}_{k=0}^{T}$ such that there exists $\{\mathbf{x}(k), \boldsymbol{\eta}(k)\}_{k=0}^{T} \in \mathcal{T}_{G^f}(\{\mathbf{u}(k), \mathbf{y}(k)\}_{i=0}^{T})$, we have $\mathcal{T}_G(\{\mathbf{u}(k), \mathbf{y}(k)\}_{k=0}^{T}) = \emptyset$.

**Proof.** A fault model $G^f$ for a switched system $G$ is $T$-step detectable if and only if for all $T+1$-length sequences $\{\mathbf{u}(k), \mathbf{y}(k)\}_{k=0}^{T}$, $\{\mathbf{u}(k), \mathbf{y}(k)\}_{k=0}^{T} \in \mathcal{B}_{swa}^{T}(G^f)$ implies $\{\mathbf{u}(k), \mathbf{y}(k)\}_{k=0}^{T} \notin \mathcal{B}_{swa}^{T}(G)$. By Lemma 1, this is equivalent to for all $T+1$-length sequences $\{\mathbf{u}(k), \mathbf{y}(k)\}_{k=0}^{T}$, $\mathcal{T}_{G^f}(\{\mathbf{u}(k), \mathbf{y}(k)\}_{k=0}^{T})$ being non-empty implies that the consistency set, $\mathcal{T}_G(\{\mathbf{u}(k), \mathbf{y}(k)\}_{k=0}^{T})$, is empty. Hence, the result follows.

$T$-step detectability is closely related to observability of a system with inputs, outputs, a function of outputs and states of $G$ and $G^f$, and a concatenation of the states of $G$ and $G^f$. For instance, there is no finite $T$ if $0 \in \mathcal{X}$ and the dynamics are linear (as opposed to affine), since zero input/output is a valid behavior of both $G$ and $G^f$ in this case. Instead of deriving an observability based result, which is challenging in the noisy switched setting, in what follows, we give a necessary and sufficient condition under which a fault $G^f$ is $T-$detectable for system $G$ that can be checked by checking the satisfiability of a logic formula.

*Theorem 1.* The fault model $G^f$ is $T$-step detectable for the system $G$ if and only if there does not exist $\mathbf{x}(0:T)$, $\mathbf{x}^f(0:T)$, $\boldsymbol{\eta}(0:T)$, $\boldsymbol{\eta}^f(0:T)$, $\mathbf{u}(0:T)$ such that

$$\begin{aligned}
&\bigwedge_{k=0}^{T} \Bigg( \bigvee_{i=1}^{s} \Big[ \big(\mathbf{x}(k+1) = \mathbf{A}_i \mathbf{x}(k) + \mathbf{B}_i \mathbf{u}(k) + \mathbf{f}_i\big) \wedge \\
&\Big( \bigvee_{j=1}^{s^f} \big[ (\mathbf{x}^f(k+1) = \mathbf{A}_j^f \mathbf{x}^f(k) + \mathbf{B}_j^f \mathbf{u}(k) + \mathbf{f}_j^f) \wedge (\mathbf{C}_i \mathbf{x}(k) \\
&+ \mathbf{D}_i \mathbf{u}(k) + \boldsymbol{\eta}(k) = \mathbf{C}_j^f \mathbf{x}^f(k) + \mathbf{D}_j^f \mathbf{u}(k) + \boldsymbol{\eta}^f(k)) \big] \Big) \Big] \\
&\wedge \|\mathbf{x}(k)\|_\infty \leq M \ \wedge \|\boldsymbol{\eta}(k))\|_\infty \leq \epsilon \ \wedge \|\boldsymbol{\eta}^f(k))\|_\infty \leq \epsilon \\
&\wedge \|\mathbf{u}(k)\|_\infty \leq \min(U, U^f) \wedge \|\mathbf{x}^f(k))\|_\infty \leq M^f \Bigg)
\end{aligned} \tag{11}$$

where the superscript $f$ indicates matrices and variables corresponding to the fault model.

**Proof.** Follows from Lemma 2.

By looking at (11), one can see that it consists of logic statements over linear inequalities in reals. Satisfiability of such a statement can be verified by off-the-shelf satisfiability modulo theory (SMT) solvers De Moura and Bjørner (2011). In practice, given a model $G$ and a fault model $G^f$, one can start with $T = 1$ and incrementally increase $T$ until $T$-detectability is verified (assuming a finite $T$ exists).

We demonstrate the application of this result in Section 4 with an example.

## 4. ILLUSTRATIVE EXAMPLES

In this section, several examples are presented in order to illustrate the effectiveness of the proposed approach. All the examples are implemented on a 3.5 GHz machine with 32 GB of memory running Ubuntu. The code used to generate the following results are available in MI4Hybrid [1], an open source Matlab toolbox developed at the University of Michigan, with CPLEX as the default MILP solver.

### *4.1 Numerical Examples*

Consider a switched system, $G = \left(\mathcal{X}, \mathcal{E}, \mathbb{R}, \{G_i\}_{i=1}^{3}\right)$ where $\mathcal{X} = \{\mathbf{x} \mid \|\mathbf{x}\|_\infty \leq 150\}$, $\mathcal{E} = \{\eta \mid \|\eta\|_\infty \leq 0.5\}$ and $G_i = (A_i, B_i, 0, C_i, 0)$, with

$$A_1 = \begin{pmatrix} 1 & 0.095 \\ -25 & -2 \end{pmatrix}, \quad B_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, C_1 = (1\ 0),$$
$$A_2 = \begin{pmatrix} 1 & 0.1 \\ -22.5 & -2 \end{pmatrix}, \quad B_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, C_2 = (1\ 0),$$
$$A_3 = \begin{pmatrix} 1 & 0.17 \\ -20 & -2 \end{pmatrix}, \quad B_3 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, C_3 = (1\ 0). \quad (12)$$

Initially, we illustrate the model invalidation algorithm with three simple case studies. The a priori model for the following experiments is $G_{ap} = \left(\mathcal{X}, \mathcal{E}, \mathbb{R}, \{G_i\}_{i=1}^{2}\right)$. The input to the system is from zero-mean normal distribution with standard deviation 5, and the switching sequence and the noise are also generated uniformly randomly. The input-output data sequence is generated in the following three ways.

- *Case 1*: The input-output sequence is generated by the system $G_{ap}$. The average signal to noise ratio for this data is 8.05 $dB$. We apply the model invalidation approach to this data with the a priori model $G_{ap}$ for 20 times and it is not invalidated in any of the trials.
- *Case 2*: This data is generated by the system $G_{ap}$ with $\mathcal{E}$ modified as $\mathcal{E} = \{\eta \mid \|\eta\|_\infty \leq 0.6\}$, and the signal to noise ratio is 9.21 $dB$. The model invalidation approach is applied to this data for 20 trials and all of them are invalidated.
- *Case 3*: The input-output sequence is generated by system $G_{ap}$ with the first mode modified as $G_1 = (1.15 * A_1, B_1, 0, C_1, 0)$. The signal to noise ratio is 11.87 $dB$. All the 20 trials for invalidation of this data sequence are successful.

Secondly, we explore the scalability of the proposed approach when the number of data points (time horizon) and the number of subsystems increase. In this regard, consider system $G$ as the a priori model for the experiments. We generate valid data using $G$ and invalid data using a model $\tilde{G} = \left(\mathcal{X}, \tilde{\mathcal{E}}, \mathbb{R}, \{G_i\}_{i=1}^{2}\right)$, where $\tilde{\mathcal{E}} = \{\eta \mid \|\eta\|_\infty \leq 0.7\}$. The input to both systems is random from normal distribution with standard deviation 5. The experiment is repeated twenty times for each time horizon and the results are given in Fig. 1 (top). The next experiment illustrates the changes in overall run-time as the number of modes increases. The invalidation approach is applied to systems with one to six stable modes, where each mode has two states. The time-horizon is fixed at 200 samples for the twenty trials for each number of modes and the results are shown in Fig. 1 (bottom). These experiments show that the MILP approach for invalidation is fairly scalable.

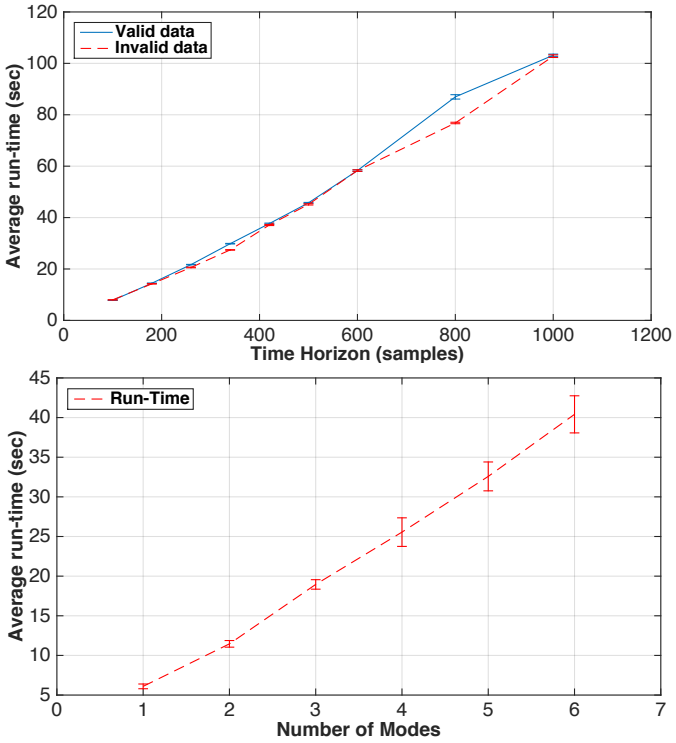


Fig. 1. Top: Run-time results for valid and invalid data with different time horizons, Bottom: Run-time results for invalid data with different number of modes.

### *4.2 T-Step Detectability*

Consider a switched system, $G = (\mathcal{X}, \mathcal{E}, \mathcal{U}, \{G_i\}_{i=1}^{2})$, with the following two modes:

$$A_1 = \begin{pmatrix} 0.8 & 1 \\ 0 & 0.1 \end{pmatrix}, B_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$
$$C_1 = I_2,\ D_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix},\ f_1 = \begin{pmatrix} 3 \\ 3 \end{pmatrix},$$
$$A_2 = \begin{pmatrix} 0.1 & 0.3 \\ 1 & 0 \end{pmatrix}, B_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$
$$C_2 = I_2,\ D_2 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, f_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad (13)$$

and a fault model $G^f = (\mathcal{X}, \mathcal{E}, \mathcal{U}, \{G_i^f\}_{i=1}^{1})$ with a single mode:

$$A_1^f = \begin{pmatrix} 0.1 & -1 \\ 0.3 & 0.3 \end{pmatrix}, B_1^f = \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$
$$C_1^f = I_2,\ D_1^f = 0,\ f_1^f = \begin{pmatrix} 10 \\ 10 \end{pmatrix}, \quad (14)$$

where $\mathcal{X} = \{x \mid x \geq 0\}$, $\mathcal{U} = \mathbb{R}$ and $\mathcal{E} = \{\eta \mid |\eta| \leq 0.1\}$. An application of Theorem 1 shows that the fault $G^f$ is 5-step detectable for the system $G$ under the mentioned constraints for noise and state variables. The Matlab code used to generate SMT files encoding the formula (11) is

[1] https://github.com/data-dynamics/MI4Hybrid

incorporated in MI4Hybrid toolbox and satisfiability is checked with CVC4 SMT solver Barrett et al. (2011).

### *4.3 Fault Detection in Hydronic Radiant Systems*

In this section, we consider hydronic radiant systems application in the smart buildings. Specifically, we are interested in a system with one pump and two zones as shown in Fig. 2.

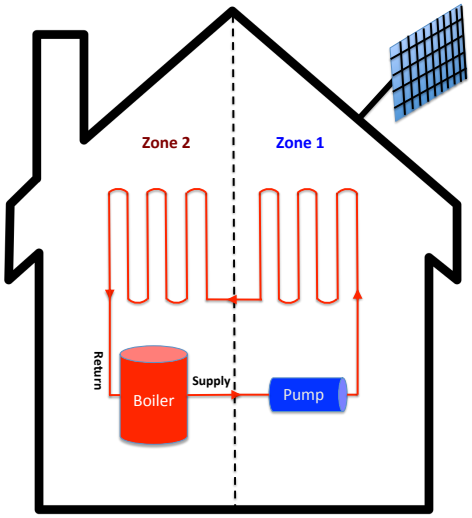


Fig. 2. Diagram of hydronic radiant system for two zones.

Such a system is represented with the following variables:

- 2 inputs: supply water temperature, $T_w$, and the pump speed.
- 3 state variables: temperature of each zone and temperature of the core water.
- 1 output: only the core water temperature is measured.

This system can be represented with a third order state space model as in Sun et al. (2014). It has been shown that making the following two assumptions reduces the amount of energy used by the radiant system and the maintenance costs Nghiem et al. (2013).

*A 1.* The supply water temperature is fixed.

*A 2.* The pump is either switched off or operates with constant speed that is known.

Considering the above assumptions, the system has a switched affine model represented by two modes in the state space form. Since, the two inputs are considered fixed, the system has no inputs, but a constant term will be added to the states for each mode that is represented by $F_i$ for mode $i$. One mode represents the system when the pump is off, and the other is for the case that the pump is running. When the pump is running, the system of differential equations for the radiant system is as follows:

$$C_r\dot{T}_c = \sum_{i=1}^{2} K_{r,i}(T_i - T_c) + K_w(T_w - T_c),$$
$$C_i\dot{T}_i = K_{r,i}(T_c - T_i) + \sum_{j\neq i} K_i j(T_j - T_i) \qquad (15)$$

and when it is off the equations are the same, except that $K_w = 0$. The values of the parameters in the model are provided in Table 1.

Table 1. Parameter values for the radiant system

| | |
|---|---|
| $T_w = 18°C$ | $K_{12} = 5(W/km^2)$ |
| $K_1 = 0.48(W/km^2)$ | $K_2 = 0.45(W/km^2)$ |
| $K_w = 20(W/km^2)$ | $C_r = 4000kJ/K$ |
| $K_{r,1} = 8(W/km^2)$ | $K_{r,2} = 7.7(W/km^2)$ |
| $C_1 = 1900kJ/K$ | $C_2 = 2100kJ/K$ |

The state space equations are discretized with sampling time of 5 minutes, and two modes with the following matrices are obtained:

$$A_1 = \begin{pmatrix} 0.54 & 0.21 & 0.23 \\ 0.44 & 0.27 & 0.24 \\ 0.43 & 0.21 & 0.31 \end{pmatrix}, \quad B_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix},$$
$$C_1 = (1\ 0\ 0), \ D_1 = 0, \quad F_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix},$$
$$A_2 = \begin{pmatrix} 0.16 & 0.11 & 0.12 \\ 0.22 & 0.23 & 0.19 \\ 0.22 & 0.17 & 0.27 \end{pmatrix}, \quad B_2 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix},$$
$$C_2 = (1\ 0\ 0), \ D_2 = 0, \quad F_2 = \begin{pmatrix} 10.8414 \\ 5.5494 \\ 5.2614 \end{pmatrix}. \qquad (16)$$

We assume failure of the furnace to heat the supply water as the fault for the system. This fault is injected in the last 30 minutes of the data as a ramp decrease in the temperature of supply water from 18° to 17°. Such a failure affects the system parameters associated with $F_2$ such that for the last 6 samples $F_2$ is the columns of the following matrix, which resembles ramp decrease in supply water temperature.

$$\begin{pmatrix} 10.6607 & 10.4800 & 10.2391 & 10.2391 & 10.2391 & 10.2391 \\ 5.4569 & 5.3644 & 5.2411 & 5.2411 & 5.2411 & 5.2411 \\ 5.1737 & 5.0860 & 4.9691 & 4.9691 & 4.9691 & 4.9691 \end{pmatrix}$$

Note that there is no large change in the value of the parameters, but such a small change can be detected by the proposed approach.

The output of the system when there is no failure and when the furnace failure occurs are illustrated in Fig. 3. We also add measurement noise, with uniform random distribution in $[-0.3, 0.3]$ to the output to emulate sensor inaccuracy.

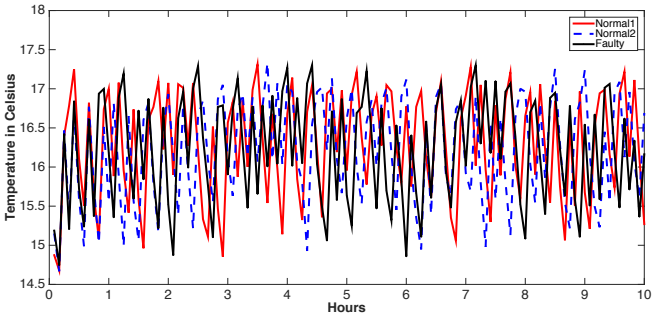


Fig. 3. Noisy output from the healthy and faulty systems.

As one can see, the outputs of the healthy and the faulty systems look very similar. Our approach is able

to invalidate the data illustrated with black solid line in Fig. 3, but it did not invalidate the other two output data generated by the healthy system. This experiment demonstrates the fact that the invalidation approach can detect relatively "small" faults in a reasonable time after occurrence.

## 5. CONCLUSIONS AND DISCUSSION

In this paper, we proposed an approach for checking the validity of a switched affine model based on input-output data. The first contribution of the paper is extending the recently proposed model invalidation techniques from switched regressor models to switched state-space models. Arguably, state-space models are more widely used in modeling cyber-physical systems, therefore are more appropriate in this context. The second contribution is to show the connection between model invalidation and fault/anomaly detection. In particular, the concept of $T$-step detectability of a fault was introduced and characterized that enables running model invalidation algorithms in a receding horizon manner while maintaining fault detection guarantees even when one discards most of the previously measured data. Finally, these techniques are demonstrated in an illustrative CPS example, namely for fault detection in the hydronic radiant systems in smart buildings.

In the future, we will extend this framework to account for parametric uncertainty in the models. Another interesting direction is to incorporate other behavioral constraints (such us those expressed in terms of temporal logics) into the proposed framework.

## REFERENCES


Barrett, C., Conway, C.L., Deters, M., Hadarean, L., Jovanović, D., King, T., Reynolds, A., and Tinelli, C. (2011). Cvc4. In *Proceedings of the 23rd International Conference on Computer Aided Verification*, CAV'11, 171–177. Springer-Verlag, Berlin, Heidelberg.

Cheng, Y., Wang, Y., Sznaier, M., Ozay, N., and Lagoa, C. (2012). A convex optimization approach to model (in)validation of switched arx systems with unknown switches. In *IEEE 51st Annual Conference on Decision and Control (CDC)*, 6284–6290.

Chow, E. and Willsky, A. (1984). Analytical redundancy and the design of robust failure detection systems. *IEEE Transactions on Automatic Control*, 29(7), 603–614.

CPLEX, I.I. (2009). V12. 1: User's manual for cplex. *International Business Machines Corporation*, 46(53), 157.

De Moura, L. and Bjørner, N. (2011). Satisfiability modulo theories: Introduction and applications. *Commun. ACM*, 54(9), 69–77.

De Persis, C. and Isidori, A. (2001). A geometric approach to nonlinear fault detection and isolation. *Automatic Control, IEEE Transactions on*, 46(6), 853–865.

Diallo, D., Benbouzid, M.E.H., Hamad, D., and Pierre, X. (2005). Fault detection and diagnosis in an induction machine drive: A pattern recognition approach based on concordia stator mean current vector. *Energy Conversion, IEEE Transactions on*, 20(3), 512–519.

Frank, P.M. (1990). Fault diagnosis in dynamic systems using analytical and knowledge-based redundancy—a survey and some new results. *Automatica*, 26(3), 459–474.

Frank, P. and Ding, X. (1997). Survey of robust residual generation and evaluation methods in observer-based fault detection systems. *Journal of Process Control*, 7(6), 403 – 424.

Isermann, R. (2005). Model-based fault-detection and diagnosis status and applications. *Annual Reviews in Control*, 29(1), 71 – 85.

Isermann, R. (2006). *Fault-Diagnosis Systems*. Springer, Berlin.

Miljkovic, D. (2011). Fault detection methods: A literature survey. In *MIPRO, 2011 Proceedings of the 34th International Convention*, 750–755.

Nghiem, T., Pappas, G., and Mangharam, R. (2013). Event-based green scheduling of radiant systems in buildings. In *American Control Conference (ACC)*, 455–460.

Ozay, N., Sznaier, M., and Lagoa, C. (2010). Model (in) validation of switched arx systems with unknown switches and its application to activity monitoring. In *49th IEEE Conference on Decision and Control (CDC)*, 7624–7630.

Ozay, N., Sznaier, M., and Lagoa, C. (2014). Convex certificates for model (in)validation of switched affine systems with unknown switches. *IEEE Transactions on Automatic Control*, 59(11), 2921–2932.

Prajna, S. (2006). Barrier certificates for nonlinear model validation. *Automatica*, 42(1), 117–126.

Rajah, V. (2014). Taming the data deluge. URL `http://www.oracle.com/us/corporate/profit/big-ideas/010312-vrajah-1917698.html`.

Smith, R.S. and Doyle, J.C. (1992). Model validation: A connection between robust control and identification. *IEEE Transactions on Automatic Control*, 37(7), 942–952.

Sun, F., Ozay, N., Wolff, E., Liu, J., and Murray, R. (2014). Efficient control synthesis for augmented finite transition systems with an application to switching protocols. In *American Control Conference (ACC)*, 3273–3280.

Sznaier, M., Camps, O., Ozay, N., and Lagoa, C. (2014). Surviving the upcoming data deluge: A systems and control perspective. In *51st IEEE Conference on Decision and Control (CDC)*.

Venkatasubramanian, V., Rengaswamy, R., Yin, K., and Kavuri, S.N. (2003). A review of process fault detection and diagnosis: Part i: Quantitative model-based methods. *Computers and Chemical Engineering*, 27(3), 293 – 311.

Verron, S., Tiplica, T., and Kobi, A. (2010). Fault detection of univariate non-gaussian data with bayesian network. In *IEEE International Conference on Industrial Technology (ICIT)*, 94–99.