Purdue University
Purdue e-Pubs

Open Access Dissertations

Theses and Dissertations

12-2016

Observability and observer design for switched linear systems

Scott C. Johnson
Purdue University

Graduate School Form 30
(Revised 08/14)

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Scott Johnson

Entitled Observability and Observer Design for Switched Linear Systems

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

RAYMOND A. DE CARLO	MILOS ZEFRAN
GREGORY M. SHAVER	
JIANGHAI HU	
STEVEN D. PEKAREK	

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification/Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy on Integrity in Research" and the use of copyrighted material.

Approved by Major Professor(s): RAYMOND A. DE CARLO

Approved by: V. Balakrishnan 11/03/2016

Head of the Department Graduate Program Date

OBSERVABILITY AND OBSERVER DESIGN
FOR SWITCHED LINEAR SYSTEMS

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Scott C. Johnson

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

December 2016

Purdue University

West Lafayette, Indiana

This thesis is dedicated to my wife, April. Her patience and steadfast encouragement has made this journey possible.

ACKNOWLEDGMENTS

The author would like to thank Dr. Raymond DeCarlo for his commitment to research at the highest level attainable. The author would also like to thank Dr. Richard T. Meyer, Dr. Steven Pekarek, and Dr. Milǒs Žefran for helpful advice and encouragement along the way. The author has been graciously supported financially by the Graduate School, Electrical and Computer Engineering Department, and the Department of Energy in collaboration with Dr. Gregory Shaver and the Hoosier Heavy Hybrid Center of Excellence.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

ABSTRACT

Johnson, Scott C. PhD, Purdue University, December 2016. Observability and Observer Design for Switched Linear Systems. Major Professor: Raymond DeCarlo.

Hybrid vehicles, HVAC systems in new/old buildings, power networks, and the like require safe, robust control that includes switching the mode of operation to meet environmental and performance objectives. Such switched systems consist of a set of continuous-time dynamical behaviors whose sequence of operational modes is driven by an underlying decision process. This thesis investigates feasibility conditions and a methodology for state and mode reconstruction given input-output measurements (not including mode sequence). An application herein considers insulation failures in permanent magnet synchronous machines (PMSMs) used in heavy hybrid vehicles.

Leveraging the feasibility literature for switched linear time-invariant systems, this thesis introduces two additional feasibility results: 1) detecting switches from safe modes into failure modes and 2) state and mode estimation for switched linear time-varying systems. This thesis also addresses the robust observability problem of computing the smallest structured perturbations to system matrices that causes observer infeasibility (with respect to the Frobenius norm). This robustness framework is sufficiently general to solve related robustness problems including controllability, stabilizability, and detectability.

Having established feasibility, real-time observer reconstruction of the state and mode sequence becomes possible. We propose the embedded moving horizon observer (EMHO), which re-poses the reconstruction as an optimization using an embedded state model which relaxes the range of the mode sequence estimates into a continuous space. Optimal state and mode estimates minimize an L2-norm between the measured output and estimated output of the associated embedded state model. Necessary conditions for observer convergence are developed. The EMHO is adapted to solve the surface PMSM fault detection problem.

1. INTRODUCTION AND PROBLEM STATEMENT

This thesis investigates observability and observer design for switched state models, possibly time-varying. Switched systems consist of a set of continuous-time dynamical behaviors (vector fields) in the state and input of the form:

$$\dot{x} = f_{v(t)}(t, x, u) \tag{1.1a}$$

$$y = g_{v(t)}(t, x, u), \tag{1.1b}$$

where (i) $v(t) \in S_M \triangleq \{0, 1, \ldots, M-1\}$, that is $v(t)$ takes values in a finite set meaning only finite set of possible dynamical behaviors of the system, (ii) for the linear case (1.1) has the form

$$\dot{x}(t) = A_{v(t)}(t)x(t) + B_{v(t)}(t)u(t), \;\; x(t_0) = x_0 \tag{1.2a}$$

$$y(t) = C_{v(t)}(t)x(t) \tag{1.2b}$$

where at time t, $x(t) \in \mathbb{R}^n$ and $u(t) \in \mathbb{R}^m$ are the current state and known control input, respectively; $y(t)$ is the measured output; and for each $i \in S_M$ the system matrices $A_i(t)$, $B_i(t)$, and $C_i(t)$ are piecewise analytic with dimension $\mathbb{R}^{n\times n}$, $\mathbb{R}^{n\times m}$, and $\mathbb{R}^{p\times n}$, respectively. Here piecewise analytic functions are used for convenience but one only needs functions with the number of continuous derivatives needed for subsequent theorems.

The function $v(t)$ evolves by some underlying decision process or environmental triggers which determine the switched dynamics of (1.1) and (1.2). The active vector field is termed the mode of operation. When the mode of operation is driven by environmental factors or otherwise uncontrolled, the mode sequence is referred to as autonomous.

This report investigates conditions of feasibility, robustness, and methods for reconstruction of both the continuous state of the dynamical system and the mode of operation from input-output measurements. Chapter 2 discusses the relevant literature for feasibility including extensions for switched linear time-varying (SLTV) state models. Chapter 3 develops a robustness metric for reconstructing the state and mode and an algorithm for computing this robustness metric. Specifically, Chapter 3 considers a larger family of robustness problems which includes the state and mode reconstruction problem for SLTI systems as a special case. Chapter 4 combines a literature review and a novel observer algorithm for reconstructing the state and mode of operation from the input-output measurements. The effectiveness of the observer is demonstrated in the context of fault detection in Chapter 5. We first motivate the switched system observer problem.

1.1 Motivation

Autonomous mode switching can model faults such as wheel-slippage in a wheeled mobile robot (wmr) [1, 2], for which the slipping dynamics are modeled as another mode of operation. An example is when a wmr encounters a patch of ice. How can one detect when the wmr enters the slipping dynamics?

Autonomous modes can also be used to model cyber-physical attacks on a power network [3]. When an external agent attacks the power network, energy is diverted, generators are overloaded, etc. which causes a change in the overall power network dynamics. How does one observe this change in the mode of operation?

Another example is insulation failure in Permanent Magnet Synchronous Machines (PMSM) which are commonly used in heavy hybrid vehicles such as the 644k hybrid wheel loader built by Deere and Co. Here insulation failures along the phase windings can cause shorts which cause a discrete change in the dynamics of the PMSM. Detecting these shorts is critical to machine integrity and thus robust and safe operation.

The automotive industry has many examples of controlled mode switching. For example, the energy saving capabilities of hybrid vehicles are linked to the power train configurations (or modes): combustion engine propulsion with and without charging, electric drive propulsion, regenerative braking, etc. In the hybrid vehicle, the mode sequence may be controlled by an underlying decision process designed to balance energy efficiency and system performance. Other examples of controlled mode switching include the power train configuration of hybrid fuel cell vehicles [4] and the PWM signal in a boost converter [5]. Given input-output measurements can one observe the state of the vehicle as well as its mode of operation when the mode is unavailable?

The work in this thesis is motivated by these autonomous and controlled switched observer and detection problems. Algorithm feasibility and design represent the first stage of observer development. The second stage is to implement a real-time observer which reconstructs both the state and the mode sequence for a class of switched systems. The real-time observer is necessary for practical implementation on systems which require state and mode estimates for real-time control.

1.2 Definitions, Assumptions, and Problem Statement

In this thesis we consider switched linear time-varying (SLTV) systems in (1.2). The following assumptions are also necessary.

Assumption 1.1. *The state $x(t)$ does not exhibit state jumps.*

Assumption 1.2. *The switching sequence $v(t)$ has a minimum dwell time T_{min}, that is $v(t)$ is piecewise constant and for two subsequent switching times t_1 and t_2 satisfies $t_2 - t_1 \geq T_{min}$.*

Assumption 1.3. *The input $u(t) : \mathbb{R} \to \mathbb{R}^m$ is piecewise continuous.*

Before one can construct an algorithm for reconstructing the state and mode of operation in an interval $[t_0, t_f]$, one needs to set forth conditions for feasibility of this reconstruction. This section introduces the framework for the feasibility problem and the proposed observer algorithm. We begin with the feasibility framework.

For the observability or feasibility problem, we note that conditions on $A_i(t)$, $B_i(t)$, $C_i(t)$, and $u(t)$ are sufficient for the existence and uniqueness of the solution to (1.2a) given a piecewise constant $v(t)$ and initial condition x_0. Thus the output is uniquely described. One need only reconstruct the initial condition and mode sequence. This motivates the definition of initial state and mode sequence (SMS) observability adapted from [6]. As with classical observability, the definition of SMS observability begins with the notion of SMS indistinguishability, that is unobservability.

Definition 1.1. *For the system in* (1.2)*, two initial state and mode sequences (SMS),* $\{x_0, v(t)\}$ *and* $\{\bar{x}_0, \bar{v}(t)\}$*, are indistinguishable on the interval* $[t_0, t_0+T]$ *if the output responses are equivalently equal, i.e.,* $y(t) \equiv \bar{y}(t)$*, and either (i)* $u \not\equiv 0$ *or (ii)* x_0 *and* $\bar{x}_0$ *are not both zero.* $I(x_0, v(t))$ *denotes the set of SMS that are indistinguishable from* $\{x_0, v(t)\}$*.*

Definition 1.2. *We say that system* (1.2) *is SMS observable with input* $u(t)$ *over* $[t_0, t_0+T]$ *if no two SMS* $\{x_0, v(t)\}$ *and* $\{\bar{x}_0, \bar{v}(t)\}$*, are indistinguishable over* $[t_0, t_0+T]$*, i.e.* $I(x_0, v(t)) = \{x_0, v(t)\}$ *for all* x_0 *and* $v(t)$*.*

For a linear system without input $u(t) \equiv 0$, an initial condition of $x_0 = 0$ results in a state trajectory of $x(t) \equiv 0$ regardless of the mode sequence. Subsequently, the output is identically zero $y(t) \equiv 0$ for all mode sequences. This implies the mode sequence cannot be reconstructed.

The addition of the continuous input complicates the observability problem. For an unknown mode sequence, the effect of the input on the output trajectory depends on the mode sequence. In special cases, the input can cause two modes of operation to be indistinguishable. In other cases, an input $u(t)$ can cause two SMS $\{x_0, v(t)\}$ and $\{\bar{x}_0, \bar{v}(t)\}$ to be distinguishable. It is shown in Chapter 2, that under certain conditions the set of inputs causing distinguishability is generic [7,8]. This result will hold for all initial conditions so excluding $x_0 = 0$ is unnecessary when the input is present.

If the observer problem is feasible, the goal of the observer is to reconstruct (in real-time) the state $x(t)$ and mode sequence $v(t)$ using knowledge of the output $y(t)$

and input $u(t)$ over an interval $[t_0, t_f]$. Here real-time denotes solvability which is instantaneous or in the dynamic observer case the estimates at each step are delayed but converge asymptotically. The proposed observer design is a modified version of the moving horizon estimator or moving horizon observer (MHO).

1.3 Embedded MHO Problem Statement

The basic structure of the MHO is shown in Figure 1.3. The MHO considers a finite horizon $[t_1 - T, t_1]$ where $t_1 \in [t_0, t_f]$. The MHO objective is to choose an optimal state and mode estimate $\hat{x}(t)$ and $\hat{v}(t)$ minimizing the error between the measured output $y^M(t)$ and the estimated output $\hat{y}(t)$, for example minimizing the L_2 norm, $\int_{t_1-t}^{t_1} \|y^M(t) - \hat{y}(t)\|^2 dt$. Since a fixed initial condition and mode sequence uniquely describes a state trajectory which satisfies (1.2), the MHO problem can be reduced to picking an estimate $\hat{x}(t_1 - h)$ and the mode sequence $\hat{v}(t)$ over $[t_1 - T, t_1]$ for $0 \leq h \leq T$. Here h allows one to pick the state estimate to be at the beginning, end, or in the interior of the interval $[t_1 - T, t_1]$.

Switched System Observer Problem (SSOP): Reconstruct the state $x(t)$ and the mode sequence $v(t)$ over $[t_0, t_f]$ in real-time given the measured output $y^M(t)$ and the known control input $u^M(t)$ so that some output error metric is minimized.

Clearly, a brute force method to solving the SSOP is to create a bank of state observers, one observer for each mode, then choose the active mode and state based on which observer is tracking the measured output the best. This method is explored in [9–11] using various types of observers for the state estimation in each mode. The basic structure of these observers is shown in Figure 1.3. The bank of observers approach requires estimation of n states in each of the M modes. A mode change from mode i to j is identified after the mode j observer outperforms all other observers with respect to output tracking over some small interval of time. For the bank of observers approach in a MHO context, the result is an optimization problem in $n \times M$

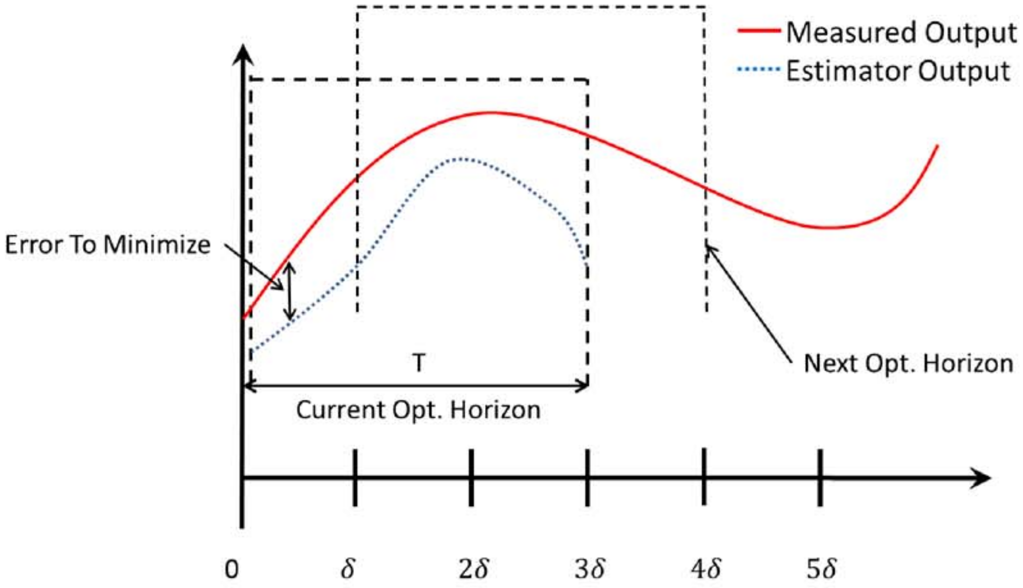

Fig. 1.1. The moving horizon observer scheme with horizon T with uniform time δ between subsequent horizons. The estimator output $\hat{y}(t)$ depends on the state estimate $\hat{x}(t-h)$ and mode estimate $\hat{v}(t)$ over the current optimization horizon.

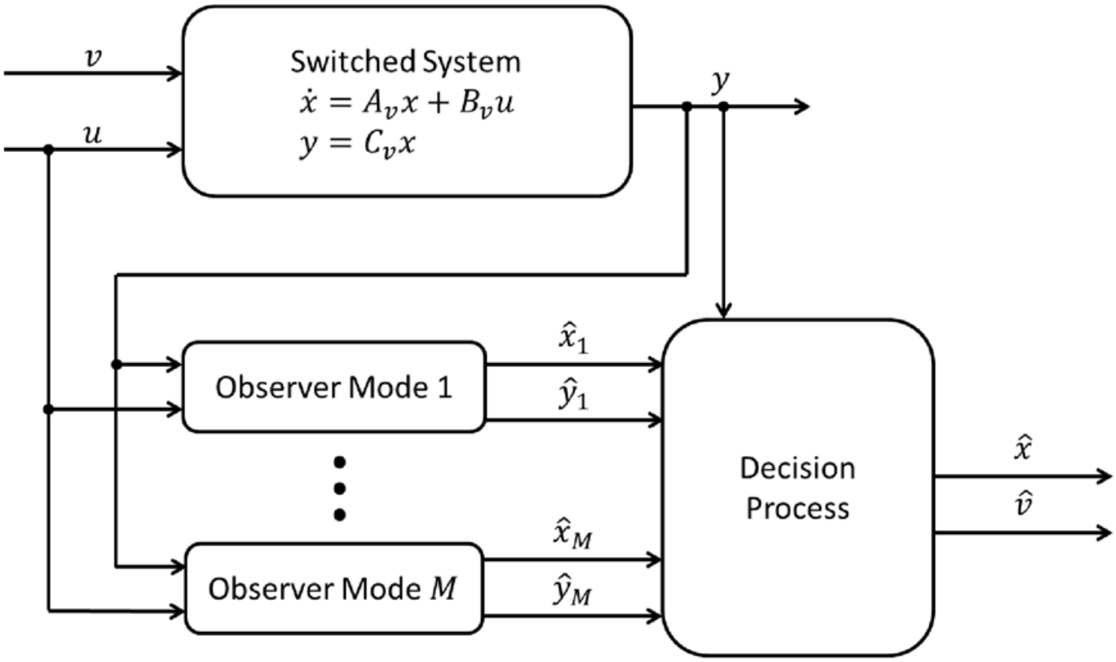

Fig. 1.2. Bank of observers each producing a state estimate $\hat{x}_i$ with the overall state and mode estimate based on a decision process using estimator outputs $\hat{y}_i$ and measured outputs y.

variables. The computational complexity of the bank of observers approach increases significantly as the number of modes becomes large.

To avoid computational complexity associated with searches over all discrete modes, we relax (embed) the range of the mode sequence estimates as in [12]. Specifically for the two mode case, this relaxation entails expanding the range of $\hat{v}(t)$ from $\{0,1\}$ to $[0,1]$. The embedded system dynamics for a two mode SLTI system has the form

$$\begin{aligned}\dot{\hat{x}}(t) &= ((1-\hat{v}(t))A_0 + \hat{v}(t)A_1)\hat{x}(t) \\ &\quad + ((1-\hat{v}(t))B_0 + \hat{v}(t)B_1)u^M(t) \qquad (1.3a)\\ \hat{y}(t) &= ((1-\hat{v}(t))C_0 + \hat{v}(t)C_1)\hat{x}(t). \qquad (1.3b)\end{aligned}$$

This embedding allows for the use of classical continuous solvers in contrast to searching all discrete modes which results in a mixed-integer programming problem. The embedding approach using an MHO will result in a classical nonlinear optimization problem in $n + M$ variables, as compared to $n \times M$ variables when using a bank of observers. In [12] it is proven that the switched system trajectories are dense in the trajectories of the embedded system. This implies that if sufficiently fast switching is allowed, any embedded system trajectory can be approximated arbitrarily close by a switched system trajectory. Conversely, a projection of the embedded mode reconstruction on the the set $\{0, 1\}$ yields mode estimates for underlying switched system mode sequence. These properties motivate the application of the embedded system formulation as a basis for the moving horizon observer. For simplicity we consider the two-mode problem. Extensions to $M > 2$ modes will follow a few simple modifications. The two-mode embedded MHO problem for switched linear systems (possibly time-varying) is formalized below.

Embedded Moving Horizon Observer (EMHO) Problem: For each finite horizon $[t_1 - T, t_1]$ and $0 \leq h \leq T$ the EHMO problem is given by

$$\min_{\substack{\hat{x}(t_1 - h) \\ \hat{v}:[t_1 - T, t_1] \to [0,1]}} \int_{t_1 - T}^{t_1} \left\| y^M(t) - \hat{y}(t) \right\|^2 dt$$

subject to:

$$\begin{aligned} \dot{\hat{x}}(t) &= ((1 - \hat{v}(t))A_0 + \hat{v}(t)A_1)\hat{x}(t) \\ &\qquad + ((1 - \hat{v}(t))B_0 + \hat{v}(t)B_1)u^M(t) \\ \hat{y}(t) &= ((1 - \hat{v}(t))C_0 + \hat{v}(t)C_1)\hat{x}(t) \end{aligned}$$

where $u^M(t)$ is the measured input. The next horizon with final time t_1' is assumed to shift in time by an amount δ, i.e. $t_1' = t_1 + \delta$.

In addition to the EMHO, we will explore a modified EMHO scheme which adds a mild penalty for deviating from previous state estimates (if available). The modified EMHO scheme is given below.

Modified Embedded Moving Horizon Observer (MEMHO) Problem: For each finite horizon $[t_1 - T, t_1]$ and $0 \leq h \leq T$ the MEHMO problem is given by

$$\min_{\substack{\hat{x}(t_1-h) \\ \hat{v}:[t_1-T,t_1]\mapsto[0,1]}} \int_{t_1-T}^{t_1} \left\| y^M(t) - \hat{y}(t) \right\|^2 dt + \Gamma(\hat{x}(t_1 - h))$$

subject to:

$$\begin{aligned} \text{(i)} \quad \dot{\hat{x}}(t) &= ((1-\hat{v}(t))A_0 + \hat{v}(t)A_1)\hat{x}(t) \\ &\quad + ((1-\hat{v}(t))B_0 + \hat{v}(t)B_1)u^M(t) \\ \text{(ii)} \quad \hat{y}(t) &= ((1-\hat{v}(t))C_0 + \hat{v}(t)C_1)\hat{x}(t) \end{aligned}$$

where

$$\Gamma(\hat{x}(t_1 - h)) = \int_{t_1-T}^{t_1-h} \gamma(t) \left\| \hat{x}(t) - \hat{x}_{prev}(t) \right\|^2 dt$$

$$\gamma : \mathbb{R} \mapsto \mathbb{R}^+ \text{ measurable penalty function}$$

and $\hat{x}_{prev}$ is the previous state estimate. If at any time t, $\hat{x}_{prev}(t)$ is unavailable, it is replaced with $\hat{x}(t)$ effectively removing it from the penalty term. The next horizon with final time t_1' is assumed to shift in time by δ, i.e. $t_1' = t_1 + \delta$.

The EMHO and MEMHO have the practical advantage of improving the computation complexity, but at what cost? Searching in the larger space of trajectories, $\mathcal{X} \triangleq \mathbb{R}^n \times [0, 1]$, risks converging to an optimal mode estimate in the interior $(0, 1)$ which does not correspond to an original switched system trajectory.

However, given conditions on the SLTI or SLTV system which guarantee SMS observability, it is proven in Chapter 4 that the set of optimal solutions with a mode estimate in $(0, 1)$ is contained in a set $L \subset \mathcal{X}$ which has codimension at least 2 in $\mathcal{X}$. This implies that the set of problem points is a small subset of the search space (if such points exist at all), but we need a stronger result to guarantee the EMHO or MEMHO can converge. We need to show is that we can navigate through the search space $\mathcal{X}$ while avoiding the set contained in L. This is achieved by proving $\mathcal{X} \setminus L$ is path connected (see Chapter 4 for details). Loosely speaking, this characterization of the embedded search space implies that almost all search paths in $\mathcal{X}$ will not pass through this set L and the optimal solution for the original switched system can be reached.

In addition to characterizing the search space for the EMHO and MEMHO, Chapter 4 reviews the observer literature for the switched system observer problem. Chapter 5 demonstrates the usefulness of the EMHO in the context of fault detection within a surface permanent magnet synchronous machine (SPMSM).

2. OBSERVABILITY OF SWITCHED SYSTEMS

This chapter explores relevant SMS observability results for switched linear time-invariant (SLTI) and switched linear time-varying (SLTV) systems. Section 2.1 compiles relevant LTI system properties needed for the switched system results to follow. Section 2.3 introduces an extension to SMS observability for SLTI systems known as set-transition observability. Section 2.4 extends the SMS observability conditions to SLTV systems. The SLTI system has the form

$$\dot{x}(t) = A_{v(t)}x(t) + B_{v(t)}u(t) \tag{2.1a}$$

$$y(t) = C_{v(t)}x(t), \tag{2.1b}$$

which is special case of SLTV system in (1.2) where $v(\cdot) \in S_M = \{0, 1, \ldots, M\}$. As in (1.2), we will assume there are no state jumps and that the mode sequence $v(t)$ has a minimum dwell time T_{min}, as per Assumptions 1.1 and 1.2. We begin with linear time-invariant (LTI) system observability results which are the basis for SMS observability of SLTI and SLTV systems.

2.1 LTI System Background

The background material in this section is comprised of two topics: observability and disturbance decoupling for LTI systems.

2.1.1 Review of LTI System Observability Results

The LTI system model is given by

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{2.2a}$$

$$y(t) = Cx(t), \tag{2.2b}$$

where A, B, C are real matrices of dimension $n \times n$, $n \times m$, and $p \times n$, respectively. The input $u(t)$ is assumed piecewise continuous (as a sufficient condition for existence and uniqueness of the state solution). The state trajectory $x(t)$ with dynamics in (2.2a) and initial condition $x(t_0) = x_0$ has solution structure

$$x(t) = e^{A(t-t_0)}x_0 + \int_{t_0}^{t} e^{A(t-\tau)}Bu(\tau)d\tau. \tag{2.3}$$

Thus the output is

$$y(t) = Ce^{A(t-t_0)}x_0 + C\int_{t_0}^{t} e^{A(t-\tau)}Bu(\tau)d\tau. \tag{2.4}$$

From (2.3), it is clear that the state $x(t)$ is uniquely defined given an initial condition x_0 and input $u(t)$. Since the input $u(t)$ is assumed known, the reconstruction of the entire state trajectory is equivalent to reconstructing the state $x(t_1)$ for any time $t_1 \leq t$. Specifically, one often computes the initial state x_0.

The last term in (2.4) depends only on the input $u(t)$ so this term can be computed and its effect subtracted from the measured output $y(t)$. As such, system observability reduces to the null space of $Ce^{A(t-t_0)}$ containing only $x_0 = 0$. This is summarized in the formal definition below.

Definition 2.1. *For the system in* (2.2)*, the state* $x_0 \in \mathbb{R}^n \backslash 0$ *is unobservable if the zero-input system response is identically zero, i.e.* $0 \equiv Ce^{A(t-t_0)}x_0$*. The system in* (2.2) *is said to be observable if no state is unobservable.*

The set of all unobservable states for a pair (C, A) is the unobservable subspace. The following theorem characterizes the unobservable subspace.

Theorem 2.1. *The state* x_0 *is unobservable for a given pair* (C, A) *if and only if* $Rx_0 = 0$ *where*

$$R = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} \tag{2.5}$$

The proof of theorem (2.1) follows using a Taylor series expansion of Ce^{At} and application of the Cayley-Hamilton Theorem. The following theorem summarizes a number of equivalences for observability for LTI system.

Theorem 2.2. *[13][1] For the LTI system in* (2.2)*, the following are equivalent:*

i. The pair (C,A) is observable.

ii.

$$\operatorname{rank}\begin{bmatrix} C \\ \lambda_i I - A \end{bmatrix} = n$$

for each eigenvalue λ_i of A.

iii. rank $(R) = n$*, where R is defined in* (2.5)*.*

iv. rank $(Ce^{At}) = n$*, i.e., there are n linearly independent columns each of which is a vector-valued function of time defined over $[t_0, \infty)$.*

v. The observability Gramian

$$W_O(t_1 - T, t_1) = \int_{t_1 - T}^{t_1} e^{A^\top q} C^\top C e^{Aq} dq \tag{2.6}$$

is nonsingular for all $T > 0$. In which case the current state $x(t_1)$ is given by

$$x(t_1) = e^{At_1} W_O(t_1 - T, t_1)^{-1} \int_{t_1 - T}^{t_1} e^{A^\top q} C^\top y^M(q) dq$$

where

$$y^M(t) = y(t) - C \int_{t_1 - T}^{t_1} e^{A(t_1 - q)} Bu(q) dq.$$

2.1.2 Disturbance Decoupling Problem For LTI Systems

The geometric approach [14–17] provides another lens for analyzing LTI systems. This review of the disturbance decoupling problem (DDP) uses basic geometric control concepts [14]. This is included because [8] uses these concepts for developing observability conditions for switched LTI systems.

[1] This theorem is an equivalent form of that found in [13].

To discuss the DDP, consider a linear system with disturbance $d \in \mathbb{R}^l$ represented by

$$\dot{x}(t) = Ax(t) + Bu(t) + Sd(t) \tag{2.7a}$$

$$y(t) = Cx(t) \tag{2.7b}$$

where A, B, C, and S are real matrices of dimension $n \times n$, $n \times m$, $p \times n$, and $n \times l$, respectively. In this report $\mathcal{B}$ stands for $\operatorname{Im} B$, $\mathcal{S}$ for $\operatorname{Im} S$ and $\mathcal{K}$ for $\ker C$. The term $d(t)$ represents a disturbance which is not directly measurable. Informally, the DDP is to find a state feedback $F \in \mathbb{R}^{m\times n}$ such that $u(t) = Fx$ and $d(\cdot)$ has no effect on the output $y(\cdot)$, i.e.

$$C \int_0^t e^{(A+BF)(t-\tau)} d(\tau) d\tau \equiv 0 \tag{2.8}$$

The formal statement of the DDP requires a few definitions.

Definition 2.2. *A linear subspace $\mathcal{L}$ is called A–invariant if $A\mathcal{L} \subset \mathcal{L}$, i.e. $\mathcal{L}$ is A–invariant if $w \in \mathcal{L}$ implies that $Aw \in \mathcal{L}$.*

Definition 2.3. *The controllable subspace of a pair (A, B) is*

$$\langle A \mid \mathcal{B} \rangle = \sum_{i=1}^{n} A^{i-1}\mathcal{B}. \tag{2.9}$$

Recall that this subspace is A–invariant by Cayley-Hamilton.

The definition of A–invariant can be used to describe the unobservable subspace stated in Theorem 2.1. The unobservable subspace, $\mathcal{N}$, for the pair (A, C) is the largest A–invariant subspace in $\mathcal{K}$, i.e. $\mathcal{N} = \cap_{i=1}^{n} \ker(CA^{i-1})$ [14, pg. 59]. Note that $\mathcal{N}$ is exactly the null space of the matrix R in (2.5). We can now formally state the DDP.

Disturbance Decoupling Problem: (DDP). Given A, B, $\operatorname{Im} S \triangleq \mathcal{S}$ and $\ker C \triangleq \mathcal{K}$ from (2.7), find (if possible) a feedback matrix $F \in \mathbb{R}^{m\times n}$, such that

$$\langle A + BF \mid \mathcal{S} \rangle \subset \mathcal{K} \tag{2.10}$$

The controllable subspace $\langle A+BF \mid \mathcal{S}\rangle$ of the pair $(A+BF, S)$ describes the entire effect of the disturbance $d(t)$ on the state space. So the condition $\langle A+BF \mid \mathcal{S}\rangle \subset \mathcal{K}$ implies that the disturbance is decoupled from the output, i.e. for all disturbances $d(\cdot)$

$$C\int_{t_0}^{t} e^{(A+BF)(t-\tau)} S d(\tau) d\tau \equiv 0. \tag{2.11}$$

When is the DDP solvable? To answer this question, we need to define the concept of (A, B)–invariant subspaces.

Definition 2.4. *[14] For a pair $(A, B) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m}$, a subspace $\mathcal{V} \subset \mathbb{R}^n$ is (A, B)–invariant if there exists a map $F \in \mathbb{R}^{m\times n}$ such that*

$$(A+BF)\mathcal{V} \subset \mathcal{V} \tag{2.12}$$

or equivalently $A\mathcal{V} \subset \mathcal{V} + \mathcal{B}$. The class of (A, B)–invariant subspaces contained in a subspace $\mathcal{X} \subset \mathbb{R}^n$ is denoted $\Im(A, B; \mathcal{X})$. A matrix F satisfying (2.12) *for a subspace $\mathcal{V}$ is called a friend of $\mathcal{V}$; the set of friends of $\mathcal{V}$ is denoted $\boldsymbol{F}(\mathcal{V})$.*

The class of subspaces $\Im(A, B; \mathcal{X})$ has the critical property that it is closed under the operation of subspace addition. This implies that $\Im(A, B; \mathcal{X})$ admits a supremal element, denoted by $\mathcal{V}^* = \sup \Im(A, B; \mathcal{X})$ (see [14, Lemma 4.3,4.4]). For the DDP, we consider replacing $\mathcal{X}$ with $\mathcal{K}$. This space $\mathcal{V}^*$ now represents the largest invariant subspace created by feedback matrix F which is in $\mathcal{K} = \ker C$. So if the disturbance $d(\cdot)$ (which enters through the matrix S) can be forced to lie within $\mathcal{K}$, we can solve the DDP. This insight is summarized in the following important theorem.

Theorem 2.3. *The DDP is solvable if and only if*

$$\mathcal{S} \subset \mathcal{V}^* \tag{2.13}$$

where $\mathcal{V}^* = \sup \Im(A, B; \mathcal{K})$.

The preceding theorem characterizes the DDP solution, but does not supply an algorithm for calculating $\mathcal{V}^*$. The following theorem specifies the algorithm which requires the definition of the inverse map A^{-1}.

Definition 2.5. *Let $\mathcal{S}$ be a subspace of $\mathbb{R}^n$. Then*

$$A^{-1}\mathcal{S} \triangleq \{x \in \mathbb{R}^n : Ax \in \mathcal{S}\}.$$

Theorem 2.4. *Let $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, and $\mathcal{K}$ be a subspace of $\mathbb{R}^n$. Define a sequence $\mathcal{V}^\mu$ given by*

$$\begin{aligned}\mathcal{V}^0 &= \mathcal{K}\\ \mathcal{V}^\mu &= \mathcal{K} \cap A^{-1}(\mathcal{B} + \mathcal{V}^{\mu-1})\end{aligned}$$

Then $\mathcal{V}^\mu \subset \mathcal{V}^{\mu-1}$, and for some $k \leq \dim(\mathcal{K})$

$$\mathcal{V}^k = \sup \Im(A, B; \mathcal{K}).$$

2.2 SMS Observability of SLTI Systems

2.2.1 Without Input

We can now address the SMS observability problem for SLTI systems, that is observability of both the state and the mode sequence. We begin with the case when $u(\cdot) \equiv 0$. Recall from Chapter 1, the definition of SMS observability included here again for reference.

Definition 2.6. *For the system in* (1.2)*, two initial state and mode sequences (SMS), $\{x_0, v(t)\}$ and $\{\bar{x}_0, \bar{v}(t)\}$, are indistinguishable on the interval $[t_0, t_0 + T]$ if the output responses $y(t) \equiv \bar{y}(t)$ and either (i) $u \not\equiv 0$ or (ii) x_0 and $\bar{x}_0$ are not both zero. $I(x_0, v(t))$ denotes the set of SMS that are indistinguishable from $\{x_0, v(t)\}$.*

Definition 2.7. *We say that system* (1.2) *is SMS observable with input $u(t)$ over $[t_0, t_0{+}T]$ if no two SMS $\{x_0, v(t)\}$ and $\{\bar{x}_0, \bar{v}(t)\}$, are indistinguishable over $[t_0, t_0{+}T]$, i.e. $I(x_0, v(t)) = \{x_0, v(t)\}$ for all x_0 and $v(t)$.*

The SLTI system without input is given by

$$\dot{x}(t) = A_{v(t)}x(t) \tag{2.14a}$$

$$y(t) = C_{v(t)}x(t) \tag{2.14b}$$

where A_i, C_i are real matrices with dimension $n \times n$ and $p \times n$, resp., and $v(t) \in S_M$. We will divide the observability problem into two subproblems: (i) identification of the initial state $x(t_0)$ and initial mode $v(t_0)$ and (ii) identification of switching times. We begin with the former problem.

Identification of the Initial State and Mode

Let $\mathcal{O}_{2n}(i)$ for $i \in S_M$ be an extended observability matrix of mode i:

$$\mathcal{O}_{2n}(i) = \begin{bmatrix} C_i \\ C_i A_i \\ \vdots \\ C_i A_i^{2n-1} \end{bmatrix}. \tag{2.15}$$

A sufficient condition for identification of the initial state $x(t_0)$ and initial mode $v(t_0)$ is given by Lemma 2.5.

Lemma 2.5. *[6] For the SLTI system* (2.1), *the initial state $x(t_0)$ and initial mode $v(t_0)$ is observable if and only if for each mode $i, j \in S_M$ with $i \neq j$*

$$\operatorname{rank}\left(\begin{bmatrix} \mathcal{O}_{2n}(i) & \mathcal{O}_{2n}(j) \end{bmatrix}\right) = 2n. \tag{2.16}$$

The proof of Lemma 2.5 is the objective for this subsection. Let the first switching time be given by t_1. Consider two different initial conditions (x_0, v) and $(\bar{x}_0, \bar{v})$ which are indistinguishable over $[t_0, t_1)$. This will imply that (2.16) is not satisfied. Since no switching occurs in $[t_0, t_1)$, the outputs of the two initial conditions are from (2.4):

$$C_v e^{A_v(t-t_0)} x_0 = y(t) = \bar{y}(t) = C_{\bar{v}} e^{A_{\bar{v}}(t-t_0)} \bar{x}_0.$$

By simple algebraic manipulation this implies

$$y(t) - \bar{y}(t) = \begin{bmatrix} C_v & -C_{\bar{v}} \end{bmatrix} \begin{bmatrix} e^{A_v(t-t_0)} & 0 \\ 0 & e^{A_{\bar{v}}(t-t_0)} \end{bmatrix} \begin{bmatrix} x_0 \\ \bar{x}_0 \end{bmatrix} = 0. \tag{2.17}$$

Note that (2.17) can hold over an interval $[t_0, t_1)$ if and only if each derivative is zero, i.e. $(\frac{d}{dt})^k[y(t) - \bar{y}(t)] = 0$ for each time $t \in [t_0, t_1)$ and each $k = 0, 1, 2, \ldots$. Thus for $t = t_0$, (2.17) holds if and only if for each $k = 0, 1, 2, \ldots$

$$\begin{bmatrix} C_v & -C_{\bar{v}} \end{bmatrix} \begin{bmatrix} A_v^k & 0 \\ 0 & A_{\bar{v}}^k \end{bmatrix} \begin{bmatrix} x_0 \\ \bar{x}_0 \end{bmatrix} = 0. \tag{2.18}$$

The key observation in [6] is that (2.17) is exactly the output of the following LTI system

$$\dot{\tilde{x}}(t) = \begin{bmatrix} A_v & 0 \\ 0 & A_{\bar{v}} \end{bmatrix} \tilde{x}(t) \triangleq A\tilde{x}(t) \tag{2.19a}$$

$$\tilde{y}(t) = \begin{bmatrix} C_v & -C_{\bar{v}} \end{bmatrix} \tilde{x}(t) \triangleq C\tilde{x}(t). \tag{2.19b}$$

The initial state $[x_0^\top, \bar{x}_0^\top]^\top$ is unobservable for the extended system in (2.19) if the pair $\{x_0, v\}$ and $\{\bar{x}_0, \bar{v}\}$ are indistinguishable, i.e. (2.17) holds. Specifically, the pair (A, C) in (2.19) is observable if and only if the observability matrix in (2.5) is full rank, i.e.

$$\text{rank}\left(\begin{bmatrix} C \\ CA \\ \vdots \\ CA^{2n-1} \end{bmatrix}\right) = \text{rank}\left(\begin{bmatrix} C_v & -C_{\bar{v}} \\ C_v A_v & -C_{\bar{v}} A_{\bar{v}} \\ \vdots & \\ C_v A_v^{2n-1} & -C_{\bar{v}} A_{\bar{v}}^{2n-1} \end{bmatrix}\right) = 2n \tag{2.20}$$

Recall that in the definition of SMS observability without an input, the point $x_0 = 0 = \bar{x}_0$ was excluded. With this in mind, we can see (2.20) guarantees that the initial state $x(t_0)$ and initial mode $v(t_0)$ are observable which is the result introduced in Lemma 2.5.

Identification of Switching Times

Assuming each successive switching time, say t_k, is identifiable and that there is a minimum dwell time with $t_{k+1} - t_k \geq T_{min}$, then Lemma 2.5 can be re-applied over

$[t_k, t_{k+1})$. The key result for this process is Theorem 2.6 below which summarizes necessary and sufficient conditions for identifying all switching times.

Theorem 2.6. *[6] For the SLTI system in* (2.1)*, all switching times are observable if and only if for all $i \neq j \in S_M$,*

$$\operatorname{rank}\left(\left[\mathcal{O}_{2n}(i) - \mathcal{O}_{2n}(j)\right]\right) = n, \tag{2.21}$$

and the switching times can be identified as the times t_k such that the output $y(t)$ is not smooth.

To explore the proof of Theorem 2.6, consider the first switching time t_1, which is unknown. Since the output of the LTI subsystem in each mode is smooth, a switching time from mode i to mode j at t_1 is undetectable from the output $y(t)$ if and only if for each $k = 0, 1, \ldots$

$$y^{(k)}(t_1^-) = y^{(k)}(t_1^+).$$

Combining the above equality for $k = 0, 1, \ldots, 2n-1$ yields

$$\mathcal{Y}_{2n}(t_1^-) = \mathcal{Y}_{2n}(t_1^+), \tag{2.22}$$

where

$$\mathcal{Y}_{2n}(t_1^-) \triangleq \begin{bmatrix} y(t_1^-) \\ \dot{y}(t_1^-) \\ \vdots \\ y^{(2n-1)}(t_1^-) \end{bmatrix} = \mathcal{O}_{2n}(i)x(t_1), \tag{2.23}$$

and the last equality in (2.23) follows by direct calculation. Thus (2.22) implies that the switch from mode i to j at t_1 is unobservable if and only if

$$\left(\mathcal{O}_{2n}(i) - \mathcal{O}_{2n}(j)\right) x(t_1) = 0. \tag{2.24}$$

Thus $x(t_1)$ must be in the null space of $\mathcal{O}_{2n}(i) - \mathcal{O}_{2n}(j)$ for the switching time to be unobservable leading to the necessary and sufficient condition in Theorem 2.6.

Identification of the State and Mode Sequence

Combining Lemma 2.5 and Theorem 2.6 provides the complete result for SMS observability for SLTI systems without input. As one can prove, (2.21) is a necessary condition for (2.16) so (2.16) is the only condition one needs to verify as per the following theorem.

Theorem 2.7. *[6] The SLTI system in* (2.1) *is SMS observable if and only if for all* $i, j \in S_M$

$$\operatorname{rank}[\mathcal{O}_{2n}(i), \mathcal{O}_{2n}(j)] = 2n. \tag{2.25}$$

The mode sequence can be reconstructed as $v(t') = \{k : \operatorname{rank}[\mathcal{O}_{2n}(k), \mathcal{Y}_{2n}(t')] = n\}$*. As such, the initial state is reconstructed as* $x_0 = \mathcal{O}_{2n}(v(t_0))^{-L}\mathcal{Y}_{2n}(t_0)$*, where "*$-L$*" denotes a left-inverse.*

2.2.2 With Input

When the input $u(t)$ is included, the general SLTI system is given in (2.1). As discussed previously, the primary issue with the addition of the input is that although the input $u(t)$ is known, the effect of the input on the output depends on the unknown active mode. As observed in [18] and [7], there is a large class of SLTI systems where particular inputs and initial conditions may cause indistinguishability, but where most admissible inputs and initial conditions are distinguishable. It is shown in both [18] and [7], that given certain conditions, the set of inputs which promote distinguishability for all initial conditions is generic. By generic we mean that the complement of this set has Lebesgue measure zero (See [19] for additional background on measure theory not included in this preliminary report). The generic distinguishability property will apply to *all initial conditions*; so the special case of $x_0 = 0$ is not excluded.

As seen in the case without input, derivatives of the output are useful in deriving conditions for mode distinguishability. The derivatives of the output remain important for deriving conditions for mode distinguishability in the presence of the input. Specifically, if $u(\cdot)$ is analytic (i.e. C^∞), the output of the SLTI system is piecewise analytic (piecewise because of potential mode switching). Hence, differentiation of the output will be seen to lead to necessary and sufficient conditions for the existence of an input causing mode distinguishability.

To develop such conditions, consider two SMS $(x_0, v(t) \equiv i)$ and $(x_0', v'(t) \equiv j)$ with outputs $y(t)$ and $y'(t)$ and corresponding state trajectories $x(t)$ and $x'(t)$ both satisfying (2.1). Taking time derivatives of the output difference $y(t) - y'(t)$ and borrowing notation from (2.15) and (2.23), we obtain

$$\mathcal{Y}_{2n}(t) - \mathcal{Y}'_{2n}(t) = \begin{bmatrix} \mathcal{O}_{2n}(i) & \mathcal{O}_{2n}(j) \end{bmatrix} \begin{bmatrix} x(t) \\ -x'(t) \end{bmatrix} + (\Gamma_{2n-1}(i) - \Gamma_{2n-1}(j))\,\mathcal{U}_{2n}(t),$$

where $\mathcal{U}_{2n}(t)$ denotes the input and its first $2n-1$ derivatives at t, i.e.,

$$\mathcal{U}_{2n}(t) = \begin{bmatrix} u(t) \\ \dot{u}(t) \\ \vdots \\ u^{(2n-1)}(t), \end{bmatrix} \tag{2.26}$$

and $\Gamma_{2n-1}(i)$ is the extended Toeplitz matrix for mode i given by

$$\Gamma_k(i) = \begin{bmatrix} 0 & \cdots & 0 & 0 \\ C_iB_i & \cdots & 0 & 0 \\ C_iA_iB_i & \cdots & \vdots & \vdots \\ \vdots & \cdots & 0 & 0 \\ C_iA_i^{k-1}B_i & \cdots & C_iB_i & 0 \end{bmatrix} \tag{2.27}$$

In [18], it was noted that for SLTI systems the input has an effect on the output difference with $q-1$ derivatives, $\mathcal{Y}_q(t) - \mathcal{Y}'_q(t)$, only if $\Gamma_{k_0}(i) - \Gamma_{k_0}(j) \neq 0$ for some $k_0 \in \mathbb{N}$. As it turns out by the Cayley-Hamilton theorem, $k_0 = 2n$ is necessary and sufficient for the existence of an analytic input u causing mode distinguishability as per the following proposition.

Proposition 2.8. *[18] For the SLTI system in* (2.1)*, there exists an analytic input* $u(\cdot)$ *such that modes* i *and* j *with* $i \neq j$ *are distinguishable for all initial conditions if and only if*

$$\Gamma_{2n}(i) - \Gamma_{2n}(j) \neq 0. \tag{2.28}$$

The proof of Proposition 2.8 is not within the scope of this review, but interested readers are referred to [18]. The existence of an input causing mode distinguishability implies almost every input causes mode distinguishability. The proof of this statement is proven in Section 2.4. Once the mode is determined, the problem reduces to the classical LTI state observability problem. The result is summarized in the following theorem combining results from [18] with the current notation.

Theorem 2.9. *The SLTI system in* (2.1) *is SMS observable for almost all analytic inputs* $u(\cdot)$ *if for each* $i, j \in S_M$ *with* $i \neq j$*,*

1. *the pair* (A_i, C_i) *is observable and*
2. $\Gamma_{2n}(i) - \Gamma_{2n}(j) \neq 0$.

In [7] and [8] the analytic requirement on the input is relaxed. Specifically in [8], the input $u : [0, \infty) \mapsto \mathbb{R}^m$ is considered to be in $\mathcal{U}_f = L_P(\mathbb{R}^m)$ which is the class of all piecewise continuous inputs such that

$$\int_0^\infty \sum_{i=1}^m |u_i(t)|^p dt < \infty. \tag{2.29}$$

As with [18], [8] begins by exploring when there exists an input causing distinguishability between two modes i and j. Before conditions guaranteeing such an input can be developed, we first consider the set of initial conditions for which there exists an input causing indistinguishability. Let $W_{i,j}$ be the set of initial conditions where there exists an input causing indistinguishability, i.e.

$$W_{i,j} = \left\{ \begin{bmatrix} x_0 \\ x_0' \end{bmatrix} \in \mathbb{R}^{2n} : \exists u(\cdot), \text{ s.t. } y_i(t; x_0, u) = y_j(t; x_0', u), t \geq 0 \right\} \tag{2.30}$$

where $y_i(t; x_0, u)$ denotes the output of (2.1) with initial condition x_0 and mode sequence $v(t) \equiv i$ with input $u(t)$. One can verify that $W_{i,j}$ is a subspace of $\mathbb{R}^{2n}$. The key insight in [8] is to realize $W_{i,j}$ is exactly the largest $(A_{i,j}, B_{i,j})$—invariant subspace in $\ker C_{i,j}$ where

$$\begin{aligned} A_{i,j} &= \begin{bmatrix} A_i & 0 \\ 0 & A_j \end{bmatrix}, \quad B_{i,j} = \begin{bmatrix} B_i \\ B_j \end{bmatrix}, \\ C_{i,j} &= \begin{bmatrix} C_i & -C_j \end{bmatrix} \end{aligned} \tag{2.31}$$

where these matrices represent an extended state model as first introduced in (2.19) in connection to the mode-distinguishability problem. This characterization of $W_{i,j}$ is found in the following lemma.

Lemma 2.10. *[7] For two modes i and j of the SLTI system in* (2.1)*, the indistinguishability subspace $W_{i,j}$ is equal to the supremal $(A_{i,j}, B_{i,j})$—invariant subspace contained in $\mathcal{K}_{i,j} = \ker C_{i,j}$, denoted as* $\sup \Im(A_{i,j}, B_{i,j}; \mathcal{K}_{i,j})$.

The result in Lemma 2.10 can be understood by considering the input $u(\cdot)$ as a disturbance acting on the extended system. If the disturbance "$u(\cdot)$" is not decoupled, i.e., has a measurable effect on the output of the extended system then the two modes i and j are distinguishable. To this end, distinguishing modes i and j can be resolved using the main results of the DDP in Theorem 2.3. This connection is summarized in the following theorem.

Theorem 2.11. *For two modes i and j of the SLTI system in* (2.1)*, there exists a time t and input $u(t)$ such that*

$$\forall x_0, \forall x_0', \quad y_i(t; x_0, u) \neq y_j(t; x_0', u) \tag{2.32}$$

if and only if $\mathcal{B}_{i,j} \not\subset W_{i,j}$.

Proof. See [7] for a complete proof. Included here is a sketch of the proof for conceptual understanding. From Lemma 2.10, $W_{i,j} = V^* = \sup \Im(A_{i,j}, B_{i,j}; \mathcal{K}_{i,j})$. Hence

$\mathcal{B}_{i,j} \not\subset V^*$ is equivalent to $\mathcal{B}_{i,j} \not\subset W_{i,j}$. The condition $\mathcal{B}_{i,j} \not\subset V^*$ has a practical meaning. To see this, recall that the controllable subspace of the pair $(A_{i,j}, B_{i,j})$ is the largest $A_{i,j}$-invariant subspace containing $\mathcal{B}_{i,j}$. In addition, the controllable subspace of $(A_{i,j} + B_{i,j}F_{i,j}, B_{i,j})$ is the same as the pair $(A_{i,j}, B_{i,j})$.

If $\mathcal{B}_{i,j} \subset V^*$, then the controllable subspace of the pair $(A_{i,j}, B_{i,j})$ is contained in V^*. Since $V^* \subset \mathcal{K}_{i,j}$, this implies that the controllable subspace is contained in $\mathcal{K}_{i,j}$, i.e. the input has no effect on the output of the extended system. Thus the SMS $\{x_0 = 0, i\}$ and $\{\bar{x}_0 = 0, j\}$ are indistinguishable for all inputs (by definition of SMS observability when the input is nonzero).

If $\mathcal{B}_{i,j} \not\subset V^*$, then a portion of the controllable subspace of the extended system is visible in the output of the extended system. Now we consider two classes of initial state pairs: those in the unobservable subspace of the extended system and those that are not. The pairs $[x_0^\top, \bar{x}_0^\top]^\top$ outside the unobservable subspace are distinguishable for all inputs except those driving the extended state into the unobservable subspace (which is a set of measure zero).

The pairs $[x_0^\top, \bar{x}_0^\top]^\top$ inside the unobservable subspace of the extended system need to be moved out of the unobservable subspace. Distinguishability of these states is achieved by inputs $u(\cdot)$ which (i) excite the portion of the range of $B_{i,j}$ not contained in V^* and (ii) effect the output with a function which is functionally independent of the columns of $C_{i,j} \exp(A_{i,j}(t - t_0))$. Almost all inputs in $\mathcal{U}_f$ have these properties. The desired input $u(\cdot)$ is one which satisfies (i), (ii), and the conditions for state pairs outside the unobservable subspace of the extended system. □

So if for each pair of distinct modes i and j, $\mathcal{B}_{i,j} \not\subset W_{i,j}$ then the mode sequence for the SLTI system in (2.1) is discernible for almost all inputs. Reconstructing the state then reduces to the classical observability problem for LTI systems, i.e. each LTI subsystem must be observable. This is summarized in the following theorem which repackages a few results from [7].

Theorem 2.12. *The SLTI system in* (2.1) *is SMS observable for generic inputs (in $L_p(\mathbb{R}^m)$) if for each pair of modes i and j in S_M with $i \neq j$, the pair (A_i, C_i) is observable and $\mathcal{B}_{i,j} \not\subset W_{i,j}$.*

2.3 Set-Transition Observability

This section addresses the Set-Transition (ST) observability problem for SLTI systems without a continuous input. For the ST observability problem, we consider the set of modes S_M partitioned into non-empty sets of safe and the failure modes denoted SM and FM, respectively. The partitioning is known *a priori*; however, the mode sequence $v(\cdot)$ is unavailable for direct measurement, although the initial mode $v(t_0)$ is assumed to be in SM (representing the common practice of an operator verifying initial safe operation). The ST observability problem is detecting when the system moves into a failure mode, i.e. the mode sequence changes from SM to FM. These results are published in [20]. We begin with the definition of ST observability.

Definition 2.8. *Consider the SLTI $\Sigma = \{A_i, C_i \, for \, i \in S_M\}$ in* (2.14) *with S_M partitioned into two nonempty sets SM and FM, denoted $S_M = SM \sqcup FM$ for the disjoint union. The mode sequence $v(t)$ has a minimum dwell time and $v(t_0) \in SM$. The system Σ is ST observable over $[t_0, t_f]$ if there does not exist two SMS, $\{x_0, v(t)\}$ and $\{\bar{x}_0, \bar{v}(t)\}$, indistinguishable over $[t_0.t_f]$ such that $v(t) \in SM$ for all t in $[t_0, t_f]$, $\bar{v}(t_0) \in SM$, and $\bar{v}(t_1) \in FM$ for some $t_1 \in (t_0, t_f)$.*

Remark 2.1. *Recall that the definition of indistinguishable SMS excludes the case when x_0 and $\bar{x}_0$ are both zero. This excludes the trivial case when the state is identically zero.*

The conditions of Theorem 2.7 are sufficient for ST observability because Theorem 2.7 guarantees every pair of distinct SMS's are distinguishable given the output. Hence pairs across the SM and FM boundary are distinguishable. However, ST observability only requires trajectories that evolve safely are distinguishable from those that transition into a failure mode. This allows for a relaxation of Theorem 2.7, as set forth in Theorem 2.13 below.

Theorem 2.13. *Let $\Sigma = \{A_i, C_i, i \in S_M\}$ as in (2.14) where $S_M = SM \sqcup FM$. If for all $k_s \in SM$ and $k_f \in FM$*

$$rank \begin{bmatrix} \mathcal{O}_{2n}(k_s) & \mathcal{O}_{2n}(k_f) \end{bmatrix} = 2n \tag{2.33}$$

then Σ is ST observable over $[t_0, t_f)$.

The proof of Theorem 2.13 follows directly from Theorem 2.7, but can be found in [20]. The condition in (2.33) is sufficient to guarantee the output $y(t)$ is not smooth at the set switching times, i.e. the switching times are detectable, and that each mode can be distinguished. However, this condition is sufficient but not necessary for ST observability. A specific example proving that (2.33) is not necessary is when there is only one safe mode. In this case, since the system starts in the safe mode, any mode switch is a transition into a failure mode. So in this case, the necessary and sufficient condition is that for the safe mode k_s and any failure mode k_f

$$\operatorname{rank}\left(\mathcal{O}_{2n}(k_s) - \mathcal{O}_{2n}(k_f)\right) = n, \tag{2.34}$$

which guarantees the switching times from k_s to k_f are observable from Theorem 2.6. When there is only one safe mode, (2.34) is both necessary and sufficient.

When there are multiple safe modes, the condition in (2.34) for each safe mode k_s and each failure mode k_f is necessary but not sufficient for ST observability. The issue is that one needs to distinguish safe-to-safe mode switches from safe-to-fail mode switches. The condition in (2.34) guarantees the output is not smooth at safe-to-fail mode switches, but safe-to-safe mode switches can also be non-smooth. The necessary and sufficient conditions are provided after the following technical lemma.

Lemma 2.14. *Let $\Sigma = \{A_i, C_i, i \in S_m\}$ be an SLTI system with $S_M = SM \sqcup FM$. Let $\{x_0, v(t)\}$ and $\{\bar{x}_0, \bar{v}(t)\}$ be two SMS with corresponding extended outputs $\mathcal{Y}_\infty(t)$ and $\bar{\mathcal{Y}}_\infty(t)$, respectively. At time t', $\mathcal{Y}_\infty(t') = \bar{\mathcal{Y}}_\infty(t')$ if and only if*

$$\begin{bmatrix} \mathcal{O}_{2n}(v(t')) & \mathcal{O}_{2n}(\bar{v}(t')) \end{bmatrix} \begin{bmatrix} x(t') \\ -\bar{x}(t') \end{bmatrix} = 0 \tag{2.35}$$

where $x(t)$ and $\bar{x}(t)$ are the state trajectories corresponding to $\{x_0, v(\cdot)\}$ and $\{\bar{x}_0, \bar{v}(\cdot)\}$.

Proof. See Section 2.5. □

Theorem 2.15. *Let $\Sigma = \{A_i, C_i, i \in S_M\}$ be an SLTI system in (2.14) where $S_M = SM \sqcup FM$ and $|SM| \geq 2$. Σ is ST observable over $[t_0, t_f)$ if and only if each pair (C_i, A_i) is observable for $i \in SM$ and for all $k_{s1}, k_{s2}, k_{s3} \in SM$ such that $k_{s1} \neq k_{s2}$ and for all $k_f \in FM$*

$$rank \begin{bmatrix} \mathcal{O}_{2n}(k_{s1}) & \mathcal{O}_{2n}(k_{s3}) \\ \mathcal{O}_{2n}(k_{s2}) & \mathcal{O}_{2n}(k_f) \end{bmatrix} = 2n. \tag{2.36}$$

Proof. See Section 2.5. □

2.3.1 Examples

This subsection illustrates Theorem 2.15 through two examples. The first example considers a SLTI system which is not ST observable and does not satisfy the rank condition in Theorem 2.15. This example will demonstrate how some sequences may be indistinguishable. The second example constructs another partition of S_M which is ST observable.

Example 1. Consider the SLTI system

$$\dot{x}(t) = A_{v(t)}x(t), \quad x(t_0) = x_0 \tag{2.37a}$$

$$y(t) = C_{v(t)}x(t) \tag{2.37b}$$

with three modes $v(t) \in S_M = \{0, 1, 2\}$ where

$$C_0 = \begin{bmatrix} 1 & 1 \end{bmatrix}, \quad C_1 = \begin{bmatrix} 0 & 1 \end{bmatrix}, \quad C_2 = \begin{bmatrix} 0 & 1 \end{bmatrix}$$

$$A_0 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad A_1 = \begin{bmatrix} 4 & 0 \\ 1 & 3 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$$

and the modes partitioned as $SM = \{0, 1\}$ and $FM = \{2\}$. Assume that the mode sequence $v(t)$ is known to have a minimum dwell time $T_{min} = 0.5$ and no state jumps. Is this system ST observable? To determine if this system is set observable using Theorem 2.15 the extended observability matrices are calculated as follows:

$$\mathcal{O}_{2n}(0) = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix}, \mathcal{O}_{2n}(1) = \begin{bmatrix} 0 & 1 \\ 1 & 3 \\ 7 & 9 \\ 37 & 27 \end{bmatrix}, \mathcal{O}_{2n}(2) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 2 & 0 \\ 4 & 0 \end{bmatrix}$$

Note that each LTI subsystem is observable in the classical sense, since the observability matrices are rank n. Calculating the joint observability matrices between members of the two sets results in

$$\begin{aligned} \text{rank} \begin{bmatrix} \mathcal{O}_{2n}(0) & \mathcal{O}_{2n}(2) \end{bmatrix} &= \text{rank} \left[\begin{array}{cc|cc} 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 2 & 0 \\ 1 & 0 & 4 & 0 \end{array}\right] \\ &= 3 < 2n \end{aligned}$$

$$\begin{aligned} \text{rank} \begin{bmatrix} \mathcal{O}_{2n}(1) & \mathcal{O}_{2n}(2) \end{bmatrix} &= \text{rank} \left[\begin{array}{cc|cc} 0 & 1 & 0 & 1 \\ 1 & 3 & 1 & 0 \\ 7 & 9 & 2 & 0 \\ 37 & 27 & 4 & 0 \end{array}\right] \\ &= 4 = 2n. \end{aligned}$$

To construct a pair of indistinguishable SMS, consider the two mode sequences $v(t)$ and $\bar{v}(t)$ defined over $[0, 2]$ as

$$v(t) = \begin{cases} 1, & \text{if } 0 \le t < 1 \\ 0, & \text{if } 1 \le t \le 2 \end{cases}, \quad \bar{v}(t) = \begin{cases} 1, & \text{if } 0 \le t < 1 \\ 2, & \text{if } 1 \le t \le 2 \end{cases}$$

Consider initial states $x_0 = \bar{x}_0 = [0; e^{-3}]^T$. We will show that the SMS $\{x_0, v(t)\}$ and $\{\bar{x}_0, \bar{v}(t)\}$ are indistinguishable. Since $v(t)$ stays within SM and $\bar{v}(t)$ moves from SM to FM, if $\{x_0, v(t)\}$ and $\{\bar{x}_0, \bar{v}(t)\}$ are indistinguishable then (2.37) is not ST observable. The outputs are calculated in each of the time intervals. For $t \in [0, 1)$,

$$y(t) = C_1 e^{A_1} x_0 = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} e^{4t} & 0 \\ e^{4t} - e^{3t} & e^{3t} \end{bmatrix} \begin{bmatrix} 0 \\ e^{-3} \end{bmatrix} = e^{3(t-1)}$$

$$\bar{y}(t) = C_1 e^{A_1} \bar{x}_0 = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} e^{4t} & 0 \\ e^{4t} - e^{3t} & e^{3t} \end{bmatrix} \begin{bmatrix} 0 \\ e^{-3} \end{bmatrix} = e^{3(t-1)}$$

Similarly for $t \in [1, 2]$

$$y(t) = C_0 e^{A_0} x_1 = \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} e^{t} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = 1,$$

$$\bar{y}(t) = C_2 e^{A_2} \bar{x}_1 = \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} e^{2t} & 0 \\ \frac{1}{2}(e^{2t} - 1) & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = 1$$

As discussed previously, this implies that (2.37) is not ST observable. So observability of each LTI subsystem is insufficient to guarantee ST observability of a SLTI system. One might suppose from this example that the rank condition in (2.33) is not only sufficient, but also necessary. However, this is not the case in general. Specifically, the rank condition is not necessary because the output prior to a set change provides additional information concerning the initial state. Because of this unutilized information, (2.33) is not a necessary condition for ST observability.

The next example will show that changing the partition of S_M by moving one mode from SM to FM will cause this system to become ST observable.

Example 2. Consider the same SLTI system in (2.37), with the new partition of S_M given by $SM = \{1\}$ and $FM = \{0, 2\}$. In this case the condition in Theorem 2.15 requires only two rank conditions as follows:

$$\begin{aligned} \text{rank}\begin{bmatrix} \mathcal{O}_{2n}(1) & \mathcal{O}_{2n}(2) \end{bmatrix} &= \text{rank}\left[\begin{array}{cc|cc} 0 & 1 & 0 & 1 \\ 1 & 3 & 1 & 0 \\ 7 & 9 & 2 & 0 \\ 37 & 27 & 4 & 0 \end{array}\right] \\ &= 4 = 2n \end{aligned}$$

$$\begin{aligned} \text{rank}\begin{bmatrix} \mathcal{O}_{2n}(1) & \mathcal{O}_{2n}(0) \end{bmatrix} &= \text{rank}\left[\begin{array}{cc|cc} 0 & 1 & 1 & 1 \\ 1 & 3 & 1 & 0 \\ 7 & 9 & 1 & 0 \\ 37 & 27 & 1 & 0 \end{array}\right] \\ &= 4 = 2n \end{aligned}$$

Thus Theorem 2.15 guarantees that this system is ST observable with this partition of S_M.

2.4 Observability of Switched Linear Time-Varying Systems

This section presents new contributions to SMS observability of Switched Linear Time-Varying (SLTV) systems. As in the SLTI case, the goal is reconstruction of the initial state x_0 (or the final state) and the entire mode sequence $v(\cdot)$. In this section, we set up feasibility conditions for this reconstruction for SLTV systems. When not mentioned, it is assumed throughout this section that the input and output are measured but mode sequence measurements are unavailable. Knowledge of the state at any time t for a fixed input and mode sequence uniquely describes the state and output trajectories. For convenience, we will develop the feasibility conditions using the final time t_1 of the interval $[t_1 - T, t_1]$. We will also limit our study to SLTV systems with 2 modes. For additional modes, one can apply the developed 2 mode conditions for each pair of modes. This section is divided into three subsections: (i) SMS observability without input (ii) SMS observability with inputs, and (iii) extensions to nonlinear SMS observability without input.

2.4.1 Without a Continuous Input

For the SLTV system in (1.2) with two modes $v(t) \in \{0, 1\}$ and without continuous input, we recall the extended system

$$\dot{\tilde{x}}(t) = \begin{bmatrix} A_0(t) & 0 \\ 0 & A_1(t) \end{bmatrix} \tilde{x}(t) \triangleq A(t)\tilde{x}(t) \tag{2.38a}$$

$$\tilde{y}(t) = \begin{bmatrix} C_0(t) & -C_1(t) \end{bmatrix} \tilde{x}(t) \triangleq C(t)\tilde{x}(t) \tag{2.38b}$$

For notation, we define $\Phi(\cdot, \cdot)$ to be the state transition matrix for (2.38a). The extended observability Gramian of (2.38) over an interval $[t_1 - T, t_1]$ is $W_O(t_1 - T, t_1)$ given by

$$W_O(t_1 - T, t_1) = \int_{t_1 - T}^{t_1} \Phi^\top(\tau, t_1) C^\top(\tau) C(\tau) \Phi(\tau, t_1) d\tau \tag{2.39}$$

The extended observability Gramian is critical in developing feasibility conditions for the SMS observability problem. We begin by addressing the problem of switching time identification.

Switching Time Identification

One key insight into developing conditions for SLTV observability is the identification of switching times. This section presents sufficient conditions for switching time identification. Algorithms for identifying these switching times are not the focus of this section. If the input and state model matrices are smooth, the conditions in this section are sufficient for switching times to be identified as those times where the output is not smooth, i.e. the output or its derivative of some order is discontinuous. However, such output behavior can be induced when the input or state model matrices are not smooth. Hence, any method must be able to distinguish the effects of mode switching from those of model or input induced discontinuities. Although these methods are important, this section focuses on conditions which guarantee feasibility of switching time identification.

Lemma 2.16. *Consider a SLTV system, Σ, in (1.2) satisfying Assumption 1.2 with two modes, $v(t) \in \{0, 1\}$, and no continuous input. Then Σ is switching time observable in the interval $[t_1 - T, t_1]$ if the final state is nonzero and over any subinterval $[t_1', t_2'] \subset [t_1 - T, t_1]$ the extended observability Gramian in (2.39), $W_O(t_1', t_2')$, is positive definite.*

Proof. For notation, let $x(\tau; w, t_1, v_w)$ denote the solution to (1.2a) evaluated at time τ passing through the final state w at time t_1 with $u(\cdot) = 0$ and mode sequence v_w. For contradiction assume Σ is not switching time observable in $[t_1 - T, t_1]$. This implies there exists two indistinguishable final state and mode sequences $(w \neq 0, v_w(t))$ and $(z \neq 0, v_z(t))$ and a nontrivial subinterval $[t_1', t_2'] \subset [t_1 - T, t_1]$ in which v_w and v_z are constant and $v_w(t) \neq v_z(t)$; such a subinterval $[t_1', t_2']$ exists due to the minimum dwell time in Assumption 1.2. Without loss of generality let $v_w(t) = 0$ and $v_z(t) = 1$ for $t \in [t_1', t_2']$. Since the output in (1.2) is piecewise continuous, $(w, v_w(t))$ and $(z, v_z(t))$ are indistinguishable if and only if the L_2 norm of the output difference is zero. Thus

$$
\begin{aligned}
0 &= \int_{t_1'}^{t_2'} \left\| C_0(\tau)x(\tau; w, t_1, v_w) - C_1(\tau)x(\tau; z, t_1, v_z) \right\|^2 d\tau \\
&= \int_{t_1'}^{t_2'} \left\| \begin{bmatrix} C_0(\tau) & -C_1(\tau) \end{bmatrix} \begin{bmatrix} \Phi_0(\tau, t_1') & 0 \\ 0 & \Phi_1(\tau, t_1') \end{bmatrix} \begin{bmatrix} x(t_1'; w, t_1, v_w) \\ x(t_1'; z, t_1, v_z) \end{bmatrix} \right\|^2 d\tau \\
&= \begin{bmatrix} x(t_1'; w, t_1, v_w) \\ x(t_1'; z, t_1, v_z) \end{bmatrix}^T W_O(t_1', t_2') \begin{bmatrix} x(t_1'; w, t_1, v_w) \\ x(t_1'; z, t_1, v_z) \end{bmatrix} \qquad (2.40) \\
&\geq \lambda_{min}(W_O(t_1', t_2')) \left\| \begin{bmatrix} x(t_1'; w, t_1, v_w) \\ x(t_1'; z, t_1, v_z) \end{bmatrix} \right\|^2 \qquad (2.41)
\end{aligned}
$$

Note that if the states $x(t_1'; w, t_1, v_w)$ and $x(t_1'; z, t_1, v_z)$ are zero then w and z are zero, which contradicts the assumption of a nonzero final state. Finally, since $W_O(t_1', t_2')$ is positive definite, $\lambda_{min}(W_O(t_1', t_2')) > 0$. Hence, the right side of(2.41) is nonzero, i.e. $(w, v_w(t))$ and $(z, v_z(t))$ are distinguishable, which is a contradiction. □

In the SLTI case, the extended system in (2.38) is also time-invariant and the observability Gramian, $W_O(t_1', t_2')$, is positive definite if and only if

$$2n = \operatorname{rank}\left(\begin{bmatrix} C \\ CA \\ \vdots \\ CA^{2n-1} \end{bmatrix}\right) = \operatorname{rank}\left(\begin{bmatrix} \mathcal{O}_{2n}(0) & \mathcal{O}_{2n}(1) \end{bmatrix}\right). \tag{2.42}$$

This is the exact result in Theorem 2.7. As it turns out, the conditions in Lemma 2.16 will be sufficient for complete SMS observability. To obtain this result we consider feasibility for reconstructing the state and mode over an interval without switching.

State and Mode Observability Without Switching

If the switching times are observable, we can analyze intervals in which no switching occurs. For an interval $[t_1 - T, t_1]$ without switching, observability of the state and mode sequence reduces to the extended system observability Gramian being positive definite over this interval as per the following theorem.

Theorem 2.17. *Consider a SLTV system, Σ, in (1.2) with two modes, i.e. $v(t) \in \{0, 1\}$ and no continuous input. Then Σ is SMS observable over an interval without switching, $[t_1 - T, t_1]$, if and only if the extended system observability Gramian in (2.38), $W_O(t_1 - T, t_1)$, is positive definite.*

Proof. **Sufficiency:** First we assume the extended observability Gramian $W_O(t_1 - T, t_1)$ is positive definite, i.e. $\lambda_{min}(W_O(t_1 - T, t_1)) > 0$, and show that all SMS with nonzero final states are distinguishable. By assumption, switching does not occur in the interval $[t_1 - T, t_1]$, i.e. the unknown mode is constant over this interval. So we need only consider constant mode sequences over $[t_1 - T, t_1]$. Let $(w, v_w), (z, v_z) \in \mathbb{R}^n \times \{0, 1\}$ denote two indistinguishable nonzero final state and constant mode pairs for the interval $[t_1 - T, t_1]$. There are now two cases: (i) $v_w \neq v_z$ and (ii) $v_w = v_z$. In each case we prove that if the observability Gramian is positive definite then the pair of SMS (w, v_w) and (z, v_z) will be distinguishable. We begin with case (i).

Case (i): if $v_w \neq v_z$, without loss of generality let $v_w = 0$ and $v_z = 1$. The L_2 norm of the difference between the outputs of the two SMS is

$$\begin{aligned}
&\int_{t_1-T}^{t_1} \|C_0(\tau)x(\tau; w, t_1) - C_1(\tau)x(\tau; z, t_1)\|^2 \, d\tau \\
&\quad = \int_{t_1-T}^{t_1} \|C_0(\tau)\Phi_0(\tau, t_1)w - C_1(\tau)\Phi_1(\tau, t_1)z\|^2 \, d\tau \\
&\quad = \int_{t_1-T}^{t_1} \left\| \begin{bmatrix} C_0(\tau) & -C_1(\tau) \end{bmatrix} \begin{bmatrix} \Phi_0(\tau, t_1) & 0 \\ 0 & \Phi_1(\tau, t_1) \end{bmatrix} \begin{bmatrix} w \\ z \end{bmatrix} \right\|^2 d\tau \\
&\quad = \begin{bmatrix} w \\ z \end{bmatrix}^\top \int_{t_1-T}^{t_1} \Phi^\top(\tau, t_1)C^\top(\tau)C(\tau)\Phi(\tau, t_1)d\tau \begin{bmatrix} w \\ z \end{bmatrix} \\
&\quad = \begin{bmatrix} w \\ z \end{bmatrix}^\top W_O(t_1 - T, t_1) \begin{bmatrix} w \\ z \end{bmatrix} && (2.43) \\
&\quad \geq \lambda_{min}(W_O(t_1 - T, t_1)) \left\| \begin{bmatrix} w \\ z \end{bmatrix} \right\|^2 && (2.44)
\end{aligned}$$

Thus the right side of (2.44) is zero only when $w = z = 0$, a case excluded from the definition of SMS observability without input.

Case (ii): if $v_w = v_z$, then without loss of generality let $v_w = 0$. After algebraic manipulation, the L_2 norm of the output difference is

$$\begin{aligned}
&\int_{t_1-T}^{t_1} \|C_0(\tau)x(\tau; w, t_1) - C_0(\tau)x(\tau; z, t_1)\|^2 d\tau \\
&\quad = \begin{bmatrix} w - z \\ 0 \end{bmatrix}^\top W_O(t_1 - T, t_1) \begin{bmatrix} w - z \\ 0 \end{bmatrix} && (2.45) \\
&\quad \geq \lambda_{min}(W_O(t_1 - T, t_1)) \left\| \begin{bmatrix} w - z \\ 0 \end{bmatrix} \right\|^2 && (2.46)
\end{aligned}$$

This the right side of (2.46) is zero only when $w = z$, i.e. when the two SMS are equal.

Thus from the conclusions of the above two cases, (w, v_w) and (z, v_z) are indistinguishable only if $(w, v_w) = (z, v_z)$ or $w = z = 0$, i.e. Σ is SMS observable if $W_O(t_1 - T, t_1)$ is positive definite.

Necessity: if the extended observability Gramian, $W_O(t_1 - T, t_1)$, is not positive definite, then there exists a nonzero vector $h \in \mathbb{R}^{2n}$ such that

$$h^\top W_O(t_1 - T, t_1)h = 0. \tag{2.47}$$

As such, there exists $w, z \in \mathbb{R}^n$, not both zero, such that one of the following must hold: (i) $h = \begin{bmatrix} w \\ z \end{bmatrix}$, (ii) $h = \begin{bmatrix} w - z \\ 0 \end{bmatrix}$, or (iii) $h = \begin{bmatrix} 0 \\ w - z \end{bmatrix}$. In case (i), SMS $\{w, v_w(\cdot) \equiv 0\}$ and $\{z, v_z(\cdot) \equiv 1\}$ are indistinguishable by (2.43). In case (ii), (2.45) implies that SMS $\{w, v_w \equiv 0\}$ and $\{z, v_z \equiv 0\}$ are indistinguishable and $w \neq z$ since $h \neq 0$. Case (iii) follows from case (ii) via relabeling. Thus if the extended observability Gramian, $W_O(t_1 - T, t_1)$, is not positive definite then Σ is not SMS observable. □

State and Mode Observability With Switching

Combining the two preceding subsections leads to the main result for SMS observability without input.

Theorem 2.18. *Consider a SLTV system, Σ, in (1.2) satisfying Assumption 1.2 with two modes, i.e. $v(t) \in \{0, 1\}$, and no continuous input. Then Σ is SMS observable over an interval $[t_1 - T, t_1]$ if the final state $x(t_1)$ is nonzero and over each subinterval $[t_1', t_2'] \subset [t_1 - T, t_1]$ the extended observability Gramian is positive definite, i.e. $W_O(t_1', t_2') > 0$.*

Proof. By Lemma 2.16, any switching time in $[t_1 - T, t_1]$ is observable, i.e. all indistinguishable SMS have the same switching times. By Assumption 1.2, there are only a finite number of switching times in $[t_1 - T, t_1]$. So $[t_1 - T, t_1]$ can be partitioned into a finite number of subintervals $[t_k', t_{k+1}')$ in which there is no switching. For each subinterval $[t'_k, t'_{k+1}) \subset [t_1 - T, t_1]$ Theorem 2.17 implies Σ is SMS observable since $W_O(t'_1, t'_1) > 0$. Thus Σ is SMS observable over $[t_1 - T, t_1]$. □

2.4.2 Observability with Input

The addition of the continuous input causes observability of the state and mode sequence to become more complex, in general. The issue is that although the input is known, the active mode is unknown. Thus the effect of the input on the output is uncertain. The input also affects switching time identification. In the case without input, Lemma 2.16 provides conditions for switching time identification. However, with an input, the effects of inputs and mode switches need to be distinguished in the measured output. For simplicity, in this section we assume that switching times in $[t_1 - T, t_1]$ are contained in an ordered and finite set $\mathcal{A}$ as per the following assumption.

Assumption 2.1. *All switching times in* $[t_1 - T, t_1]$ *are contained in an ordered and finite set* $\mathcal{A} = \{s_\alpha \in [t_1 - T, t_1] | \alpha = 0, 1, 2, \dots K\}$ *where* $t_1 - T \triangleq s_0 < s_1 < \cdots < s_K \triangleq t_1$.

Note that Assumption 2.1 does not imply that each time s_i is a switching time. Although switching times are in $\mathcal{A}$, the input can still cause mode indistinguishability. To explore how the input can cause mode indistinguishability, consider two distinct final state and mode sequences $(w, v_w(t))$ and $(z, v_z(t))$ for the SLTV system (1.2) which are indistinguishable on an interval $[t_1 - T, t_1]$. The solution to (1.2a) for final state and mode sequence $(w, v_w(t))$ is

$$x(t; w, t_1, v_w) = \Phi_{v_w}(t, t_1)w + \int_{t_1}^{t} \Phi_{v_w}(t, q)B_{v_w(q)}(q)u(q)dq, \tag{2.48}$$

where $\Phi_{v_w}(\cdot,\cdot)$ denotes the state transition matrix for (1.2a) with fixed mode sequence $v_w(t)$. Since $(w, v_w(t))$ and $(z, v_z(t))$ are indistinguishable, the L_2 norm of the output difference for the two SMS is zero, i.e.

$$
\begin{aligned}
0 &= \int_{t_1-T}^{t_1} \left\| C_{v_w(\tau)}(\tau)x(\tau; w, t_1, v_w) - C_{v_z(\tau)}(\tau)x(\tau; z, t_1, v_z) \right\|^2 d\tau \\
&= \int_{t_1-T}^{t_1} \Bigg\| \begin{bmatrix} C_{v_w(\tau)}(\tau) & -C_{v_z(\tau)}(\tau) \end{bmatrix} \begin{bmatrix} \Phi_{v_w}(\tau, t_1) & 0 \\ 0 & \Phi_{v_z}(\tau, t_1) \end{bmatrix} \begin{bmatrix} w \\ z \end{bmatrix} \qquad (2.49a) \\
&\quad + \int_{t_1}^{\tau} \begin{bmatrix} C_{v_w(\tau)}(\tau) & -C_{v_z(\tau)}(\tau) \end{bmatrix} \begin{bmatrix} \Phi_{v_w}(\tau, q) & 0 \\ 0 & \Phi_{v_z}(\tau, q) \end{bmatrix} \begin{bmatrix} B_{v_w(q)}(q) \\ B_{v_z(q)}(q) \end{bmatrix} u(q)dq \Bigg\|^2 d\tau \qquad (2.49b)
\end{aligned}
$$

If the right side of (2.49) is nonzero for all pairs of distinct SMS, then the SLTV system is SMS observable. The first term in (2.49a) is nonzero if the final state is nonzero ($w \neq 0 \neq z$) and the observability Gramian in (2.39) is full rank, which follows from the preceding section. Even if the first term in (2.49a) is zero, the input can cause distinguishability through the second term in (2.49b). Although rare, the input can also cause indistinguishablility if the second term in (2.49b) negates the first term in (2.49a). In the results that follow, the developed sufficient conditions guarantee that almost all inputs cause the right side of (2.49) to be nonzero over $[t_1 - T, t_1]$, regardless of the final state.

Following Theorem 2.20, it is proven that the existence of a mode distinguishing input (for all final states) is sufficient for almost all inputs to be mode distinguishing. To develop conditions for the existence of a mode distinguishing input, we introduce the output-controllability Gramian after some notation. For the two modes 0 and 1, let the tuple $(A(t), B(t), C(t))$ denote the extended system matrices

$$
\begin{aligned}
A(t) &= \begin{bmatrix} A_0(t) & 0 \\ 0 & A_1(t) \end{bmatrix}, \quad B(t) = \begin{bmatrix} B_0(t) \\ B_1(t) \end{bmatrix}, \qquad (2.50) \\
C(t) &= \begin{bmatrix} C_0(t) & -C_1(t) \end{bmatrix}.
\end{aligned}
$$

Introduced in [21], the output-controllability Gramian (OCG) for the extended system is

$$P(t_1 - T, t_1) \triangleq \int_{t_1 - T}^{t_1} C(t_1)\Phi(t_1, \tau)B(\tau)(C(t_1)\Phi(t_1, \tau)B(\tau))^\top d\tau \tag{2.51}$$

For LTV systems, [21] proves that the OCG having full row rank is necessary and sufficient for output controllability, i.e. for the existence on an input driving the output to a specified value. For SLTV systems, we prove, in the following theorem, that a nonzero extended OCG and positive definite extended observability Gramian over each subinterval is sufficient for driving the extended system to a nonzero output, i.e. sufficient for mode distinguishability. The following technical lemma develops the key building block for proving the existence of an input causing mode distinguishability.

Lemma 2.19. *Consider the two mode SLTV system Σ in (1.2) with a minimum dwell time, Assumption 1.2. Consider two final state and mode sequences $\{w, v_w\}$ and $\{z, v_z\}$ such that over the subinterval $[s_1, s_2) \subset (t_1 - T, t_1)$, v_w and v_z are constant and $v_w(t) \neq v_z(t)$. If over every nonempty subinterval, $[t_1', t_2'] \subset [t_1 - T, t_1]$*

(i) $W_O(t_1', t_2') > 0$ and

(ii) $P(t_1', t_2') \neq 0$,

then there exists an input $u(\cdot)$ distinguishing $\{w, v_w\}$ and $\{z, v_z\}$.

Proof. Without loss of generality, let $v_w(t) = 0$ and $v_z(t) = 1$ for all $t \in [s_1, s_2)$. Since $P(s_1, s_2) \neq 0$, there exists $u_m \in \mathbb{R}^m$ such that $P(s_1, s_2)u_m \neq 0$. We claim the input

$$u(t) = \begin{cases} (C(s_2)\Phi(s_2, t)B(t))^\top u_m, & t \in [s_1, s_2) \\ 0, & \text{otherwise} \end{cases} \tag{2.52}$$

distinguishes $\{w, v_w\}$ and $\{z, v_z\}$. Since over the interval $[t_1 - T, s_1)$, $u(t) = 0$ and $W_O(t_1 - T, s_1) > 0$, Theorem 2.17 implies $\{w, v_w\}$ and $\{z, v_z\}$ are indistinguishable only if the corresponding state trajectories x_w and x_z are zero at s_1. We now consider the case when $x_w(s_1) = x_z(s_1) = 0$. Let $y_w(t)$ and $y_z(t)$ denote the outputs of $\{w, v_w\}$ and $\{z, v_z\}$, respectively. Then

$$\begin{aligned} y_w(s_2) - y_z(s_2) &= \int_{s_1}^{s_2} C(s_2)\Phi(s_2,\tau)B(\tau)u(\tau)d\tau \\ &= P(s_1, s_2)u_m \neq 0. \end{aligned}$$

implying $\{w, v_w\}$ and $\{z, v_z\}$ are distinguishable. □

When all switching times are in $\mathcal{A}$, Lemma 2.19 allows one to construct a mode disinguishing input for all final states, as per the following theorem.

Theorem 2.20. *Consider the two mode SLTV system* Σ *in* (1.2) *with a minimum dwell time and all switching times in* $\mathcal{A}$, *Assumptions 1.2 and 2.1, respectively. If over every nonempty subinterval,* $[t_1', t_2'] \subset [t_1 - T, t_1]$

(i) $W_O(t_1', t_2') > 0$ *and*

(ii) $P(t_1', t_2') \neq 0$,

then there exists a mode distinguishing input $u(t)$ *for all final states over interval* $[t_1 - T, t_1]$.

Proof. The proof will proceed by considering separately each subinterval $[s_i, s_{i+1})$ with $s_i, s_{i+1} \in \mathcal{A}$. For the subinterval $[s_0, s_1)$, fix a time $t_m \in (s_0, s_1)$ and let $u(t) = 0$ on the interval $[s_0, t_m)$. Since there are no switching times in $[t_m, s_1)$, Lemma 2.19 implies the existence of a mode distinguishing input over $[s_0, s_1)$ for all states at s_1. Repeating this construction over each interval $[s_i, s_{i+1})$ results in a mode distinguishing input $u(t)$ over $[t_1 - T, t_1]$ for all final states, i.e. all final states at t_1. □

We now prove that the existence of an input distinguishing two subsystems for all final states implies mode distinguishability for generic inputs. We assume that the switching times are observable or contained in the known set $\mathcal{A}$. Because potential switching times are known, we need only prove generic mode distinguishability over an interval $[s_1, s_2)$ without switching. For the interval $[s_1, s_2)$, let the set of inputs not causing mode distinguishability for all final states be denoted $\mathcal{U}_i$, a subset of the entire input space $\mathcal{U}_f = L_P(\mathbb{R}^m)$. Recall that proving $\mathcal{U}_i$ is a proper subspace of $\mathcal{U}_f$ implies that $\mathcal{U}_i$ has measure zero in $\mathcal{U}_f = L_p(\mathbb{R}^m)$.

First we prove $\mathcal{U}_i$ is a subspace. For each $u \in \mathcal{U}_i$, i.e. each input not mode distinguishing, there exists an extended final state x_f causing the output of the extended system to be zero over $[s_1, s_2)$, an interval without switching. Consider $u, u' \in \mathcal{U}_i$ with the extended final states $\tilde{x}_f, \tilde{x}'_f \in \mathbb{R}^{2n}$, respectively, which cause the output of the extended system to be identically zero over $[s_1, s_2)$. Then $\tilde{y}(t; \tilde{x}_f, u) = \tilde{y}(t; \tilde{x}'_f, u') = 0$ for all $t \in [s_2, s_2)$. Since the extended system is linear, superposition implies that for all $\alpha, \beta \in \mathbb{R}$

$$\begin{aligned} \tilde{y}(t; \alpha\tilde{x}_f + \beta\tilde{x}'_f, \alpha u + \beta u') &= \alpha\tilde{y}(t; \tilde{x}_f, u) + \beta\tilde{y}(t; \tilde{x}'_f, u') \\ &= 0. \end{aligned}$$

Thus $\alpha u + \beta u' \in \mathcal{U}_i$ implying $\mathcal{U}_i$ is a subspace of $\mathcal{U}_f$.

Given the conditions of Theorem 2.20 are satisfied, there exists a mode distinguishing input, i.e. an input $u \notin \mathcal{U}_i$. Thus $\mathcal{U}_i$ is a proper subspace and has Lebesgue measure zero in $\mathcal{U}_f$, i.e. mode distinguishability holds for generic (almost all) inputs.

With the mode observable, only state reconstruction remains, i.e. the classical LTV observability problem. Since the positive definite extended observability Gramian implies observability of each LTV subsystem, the conditions in Theorem 2.20 are sufficient for SMS observability of SLTV systems as per the following theorem.

Theorem 2.21. *Consider the two mode SLTV system Σ in* (1.2) *with a minimum dwell time and all switching times in $\mathcal{A}$, Assumptions 1.2 and 2.1. If over every nonempty subinterval, $[t'_1, t'_2] \subset [t_1 - T, t_1]$,*

(i) $W_O(t'_1, t'_2) < 0$

(ii) $P(t'_1, t'_2) \neq 0$,

then Σ is SMS observable for almost all inputs.

Proof. By Assumptions 1.2 and 2.1, there are a finite number of intervals $[s_i, s_{i+1}) \subset [t_1 - T, t_1]$ without mode switches which partition $[t_1 - T, t_1]$. Over each interval $[s_i, s_{i+1})$ without switching, Theorem 2.20 implies there exists and input causing mode distinguishability. This implies that the mode sequence is observable for almost all inputs. All that remains is to determine the final state. Since $W_O(t_1', t_2') > 0$, each mode (active mode) has a positive definite observability Gramian over $[t_1', t_2']$, i.e. the state is observable over $[t_1', t_2']$. Thus the state $x(t)$ is observable in each subinterval $[s_i, s_{i+1})$ implying Σ is SMS observable for generic inputs. □

2.4.3 Extensions to Nonlinear Switched Systems

In this section we extend results in Section 2.4.1 to switched nonlinear systems without input and make a connection to the strong observability condition in [22]. Specifically, we consider two-mode switched nonlinear systems of the form

$$\dot{x} = f_{v(t)}(t, x) \tag{2.53a}$$

$$y = g_{v(t)}(t, x) \tag{2.53b}$$

which satisfy the following assumptions:

(i) $v(\cdot)$ has a minimum dwell time,

(ii) there exists a unique solution to (2.53a) for each initial (final) condition $x_0 \in \mathbb{R}^n$ and any admissible mode sequence $v(t)$,

(iii) $f_i(\cdot, 0) = 0$ and $g_i(\cdot, 0) = 0$ for all modes $i \in \{0, 1\}$.

As with the case for SLTV systems without input, we begin by considering when switching times are observable. In Lemma 2.16, switching times for SLTV systems without input are observable if the final state is nonzero and the observability Gramian is positive definite over each nontrivial subinterval. The observability Gramian is not defined for nonlinear systems; so a suitable nonlinear analog is desired. The following lemma extends Lemma 2.16 after some notation. Let the output (2.53b) due to the final state x at t_1 and mode sequence $v \equiv i$ be denoted $y_i(\tau; x, t_1)$.

Lemma 2.22. *Let Σ be a nonlinear switched system in* (2.53)*. Consider an interval* $[t_1 - T, t_1]$ *such that the final state is nonzero. If for any nontrivial subinterval* $[t_1', t_2']$ *and for all nonzero* $x, \bar{x} \in \mathbb{R}^n$ *there exists* $\gamma_m = \gamma_m(t_1', t_2') > 0$ *such that*

$$\int_{t_1'}^{t_2'} \|y_0(\tau; x, t_2') - y_1(\tau; \bar{x}, t_2)\|^2 d\tau \geq \gamma_m \left\| \begin{bmatrix} x \\ \bar{x} \end{bmatrix} \right\|^2, \tag{2.54}$$

then Σ is switching time observable over $[t_1 - T, t_1]$.

Note that (2.54) has the same effect as a positive definite extended observability Gramian. Specifically, if the system in (2.53) were linear, then

$$\begin{aligned}\int_{t_1'}^{t_2'} \|y_0(\tau; x, t_2') - y_1(\tau; \bar{x}, t_2)\|^2 d\tau &= \begin{bmatrix} x \\ \bar{x} \end{bmatrix}^\top W_O(t_1', t_2') \begin{bmatrix} x \\ \bar{x} \end{bmatrix} \\ &\geq \lambda_{min}(W_O(t_1', t_2')) \left\| \begin{bmatrix} x \\ \bar{x} \end{bmatrix} \right\|^2 .\end{aligned}$$

With this observation, the proof of Lemma 2.22 follows Lemma 2.16.

With switching times observable, the main extension of Theorem 2.18 can be introduced as per the following theorem.

Theorem 2.23. *Let Σ be a nonlinear switched system in* (2.53)*. Consider an interval* $[t_1 - T, t_1]$ *such that the final state is nonzero. If for any nontrivial subinterval* $[t_1', t_2']$ *and for all final states* $x, \bar{x} \in \mathbb{R}^n$ *there exists* $\gamma_m = \gamma_m(t_1', t_2') > 0$ *and* $\gamma_x = \gamma_x(t_1', t_2') > 0$ *such that*

$$\int_{t_1'}^{t_2'} \|y_0(\tau; x, t_2') - y_1(\tau; \bar{x}, t_2)\|^2 d\tau \geq \gamma_m \left\| \begin{bmatrix} x \\ \bar{x} \end{bmatrix} \right\|^2, \tag{2.55}$$

and for each mode $i \in \{0, 1\}$

$$\int_{t_1'}^{t_2'} \|y_i(\tau; x, t_2') - y_i(\tau; \bar{x}, t_2')\|^2 d\tau \geq \gamma_x \|x - \bar{x}\|^2 \tag{2.56}$$

then Σ is SMS observable over $[t_1 - T, t_1]$.

Satisfying (2.55) implies mode distinguishability over intervals without switching. The minimum dwell time assumption guarantees that the interval $[t_1 - T, t_1]$ can be partitioned into a finite set of intervals $[t_1', t_2)$ without switching. Thus, the mode sequence is observable. What remains is to guarantee feasibility of state reconstruction.

The condition in (2.56) is the strong state observability condition in [22]. The strong observability condition guarantees the state is observable for each subsystem. If the system in (2.53) is linear then letting $W_O^i(t_1', t_2')$ denote the observability Gramian for mode i,

$$\begin{aligned}\int_{t_1'}^{t_2'} \|y_i(\tau; x, t_2') - y_i(\tau; \bar{x}, t_2')\|^2 d\tau &= (x - \bar{x})^\top W_O^i(t_1', t_2')(x - \bar{x}) \\ &\geq \lambda_{min}(W_O^i(t_1', t_2'))\|x - \bar{x}\|^2.\end{aligned}$$

So (2.56) reduces to the observability Gramian in each mode being positive definite if the system is linear. Hence, the state is observable if (2.56) is satisfied. Combining state observability of each subsystem with the mode sequence observability guaranteed by condition (2.55) proves Theorem 2.23.

2.4.4 Conclusions

This section extends the existing observability conditions for LTI switched systems to those for LTV switched systems with and without input. Using the notion of strong observability, sufficient conditions for observability of nonlinear switched state models are also set forth, and represent a basis for continued research. The next chapter explores a robustness metric for state and mode sequence observability for SLTI systems with additive perturbations to each of the system matrices.

2.5 Appendix for Section 2.3

Lemma 2.14.

Proof. Let $\tilde{\Sigma} = \{\tilde{A}, \tilde{C}\}$ be an LTI system

$$\begin{aligned}\dot{\tilde{x}}(t) &= \tilde{A}\tilde{x} \\ \tilde{y} &= \tilde{C}\tilde{x}\end{aligned}$$

where

$$\tilde{A} = \begin{bmatrix} A_{v(t')} & 0 \\ 0 & A_{\bar{v}(t')} \end{bmatrix}, \qquad \tilde{C} = \begin{bmatrix} C_{v(t')} & -C_{\bar{v}(t')} \end{bmatrix}$$

By calculation, $\tilde{\Sigma}$ has observability matrix $\tilde{\mathcal{O}} = [\mathcal{O}_{2n}(v(t')), -\mathcal{O}_{2n}(\bar{v}(t'))]$. Thus, $\tilde{\Sigma}$ has an identically zero output for $t \geq t'$ exactly when $\tilde{x}(t') \in Null(\tilde{\mathcal{O}})$ since $Null(\tilde{\mathcal{O}})$ is exactly the unobservable subspace for $\tilde{\Sigma}$. Defining $\tilde{x}(t') = [x(t')^T, \bar{x}(t')^T]^T$, the preceding sentence can be expressed as

$$0 = \tilde{\mathcal{Y}}_\infty(t') = \mathcal{Y}_\infty(t') - \bar{\mathcal{Y}}_\infty(t')$$

if and only if

$$[\mathcal{O}_{2n}(v(t')), -\mathcal{O}_{2n}(\bar{v}(t'))] \begin{bmatrix} x(t') \\ \bar{x}(t') \end{bmatrix}$$

by the Cayley-Hamilton theorem proving the result. □

Theorem 2.15.

Proof. For sufficiency assume that Σ is ST unobservable, each pair (C_i, A_i) is observable for $i \in SM$, and (2.36) is satisfied. Since Σ is ST unobservable, there exists indistinguishable SMS $\{x_0, v(t)\}$ and $\{\bar{x}_0, \bar{v}(t)\}$ with x_0 or $\bar{x}_0$ nonzero, $v(t_0)$, $\bar{v}(t_0) \in SM$, and a switching time t_1 such that $v(t_1^-), v(t_1^+), \bar{v}(t_1^-) \in SM$ and $\bar{v}(t_1^+) \in FM$. By the definition of indistinguishable SMS

$$\mathcal{Y}_\infty(t_1^-) = \bar{\mathcal{Y}}_\infty(t_1^-) \tag{2.57a}$$

$$\mathcal{Y}_\infty(t_1^+) = \bar{\mathcal{Y}}_\infty(t_1^+) \tag{2.57b}$$

Define $k_{s1} = v(t_1^-)$, $k_{s2} = v(t_1^+)$, $k_{s3} = \bar{v}(t_1^-)$, and $k_f = \bar{v}(t_1^+)$. Since by assumption there are no state jumps at the transition, (2.57) and lemma 2.14 imply that

$$\begin{bmatrix} \mathcal{O}_{2n}(k_{s1}) & \mathcal{O}_{2n}(k_{s3}) \\ \mathcal{O}_{2n}(k_{s2}) & \mathcal{O}_{2n}(k_f) \end{bmatrix} \begin{bmatrix} x(t_1) \\ -\bar{x}(t_1) \end{bmatrix} = 0 \tag{2.58}$$

The contradiction of (2.36) is nearly achieved. Equation (2.58) differs from (2.36) only by the condition that $k_{s1} \neq k_{s2}$ in (2.36). So if $v(t_1^-) \neq v(t_1^+)$ then (2.36) is contradicted and sufficiency follows.

On the other hand if $k_{s1} = v(t_1^-) = v(t_1^+) = k_{s2}$, the output at t_1 is smooth, i.e.

$$\mathcal{Y}_\infty(t_1^-) = \mathcal{Y}_\infty(t_1^+) \tag{2.59}$$

Combining (2.59) with (2.57) we observe

$$\bar{\mathcal{Y}}_\infty(t_1^-) = \bar{\mathcal{Y}}_\infty(t_1^+) \tag{2.60}$$

After substitution and manipulation, the first $2np$ rows of (2.60) imply

$$\begin{bmatrix} O_{2n}(\bar{v}(t_1^-)) & \mathcal{O}_{2n}(\bar{v}(t_1^+)) \end{bmatrix} \begin{bmatrix} \bar{x}(t_1) \\ -\bar{x}(t_1) \end{bmatrix} = 0 \tag{2.61}$$

Then we must show that there is another combination of modes which contradicts the ”for all” statement in (2.36).

Since both x_0 and $\bar{x}_0$ cannot both be zero we can assume without loss of generality $\bar{x}(t_1) \neq 0$ (due to the nonsingularity of the state transition matrix for the switched LTI system).

To contradict the ”for all” statement in (2.36) we assign $k'_{s2} = \bar{v}(t_1^-)$ and $k'_f = \bar{v}(t_1^+)$. Now let $k'_{s1} \in SM$ be a safe mode such that $k'_{s1} \neq k'_{s2}$. Assign $k'_{s3} = k'_{s1}$. Combining these assignments with (2.61) results in

$$\begin{bmatrix} \mathcal{O}_{2n}(k'_{s1}) & \mathcal{O}_{2n}(k'_{s3}) \\ \mathcal{O}_{2n}(k'_{s2}) & \mathcal{O}_{2n}(k'_f) \end{bmatrix} \begin{bmatrix} \bar{x}(t_1) \\ -\bar{x}(t_1) \end{bmatrix} = 0 \tag{2.62}$$

providing the needed contradiction for sufficiency.

For the necessity of (C_i, A_i) observable for $i \in SM$, assume there exists $k_s \in SM$ such that (C_{k_s}, A_{k_s}) is not observable. Then there exists nonzero $\tilde{x} \in \mathbb{R}^n$ such that $\mathcal{O}_{2n}(k_s)\tilde{x} = 0$. Consider the two SMS $\{x_0 = \tilde{x}, v(t) = k_s\ \forall t\}$ and $\{\bar{x}_0 = 0, \bar{v}(t)\}$ where $\bar{v}(t)$ is

$$\bar{v}(t) = \begin{cases} k_s, & \text{if } t \leq t_1 \\ k_f, & \text{if } t > t_1 \end{cases}$$

and $t_1 \in (t_0, t_f)$. Then $\{x_0, v(t)\}$ and $\{\bar{x}_0, \bar{v}(t)\}$ produce outputs $y(t) = \bar{y}(t) \equiv 0$, which contradicts that Σ is ST observable.

For necessity of (2.36), assume (2.36) is not satisfied. Thus there exists integers $k_{s1} \neq k_{s2}, k_{s3} \in SM$, $k_f \in FM$ such that

$$rank \begin{bmatrix} \mathcal{O}_{2n}(k_{s1}) & \mathcal{O}_{2n}(k_{s3}) \\ \mathcal{O}_{2n}(k_{s2}) & \mathcal{O}_{2n}(k_f) \end{bmatrix} = rank[M] \leq 2n \tag{2.63}$$

Thus there exists a nonzero $z \in \mathbb{R}^{2n}$ such that $Mz = 0$. Defining $z = [x_1^T, -\bar{x}_1^T]$ implies that x_1 or $\bar{x}_1$ is nonzero (or both) and

$$\begin{bmatrix} \mathcal{O}_{2n}(k_{s1}) & \mathcal{O}_{2n}(k_{s3}) \\ \mathcal{O}_{2n}(k_{s2}) & \mathcal{O}_{2n}(k_f) \end{bmatrix} \begin{bmatrix} x_1 \\ -\bar{x}_1 \end{bmatrix} = 0 \tag{2.64}$$

This half of the proof will proceed by constructing two indistinguishible SMS based on (2.64) which will imply Σ is ST unobservable. First consider t_1 to be a time in (t_0, t_f) and define two mode sequences as

$$\begin{aligned} v(t) &= \begin{cases} k_{s1}, & \text{if } t \leq t_1 \\ k_{s2}, & \text{if } t > t_1 \end{cases} \qquad (2.65) \\ \bar{v}(t) &= \begin{cases} k_{s3}, & \text{if } t \leq t_1 \\ k_f, & \text{if } t > t_1 \end{cases} \end{aligned}$$

Since for (2.65) there is only one switch at t_1 in $[t_0, t_f)$ the initial states can be constructed as

$$\begin{aligned} x_0 &= e^{A_{k_{s1}}(t_0 - t_1)} x_1 \\ \bar{x}_0 &= e^{A_{k_{s3}}(t_0 - t_1)} \bar{x}_1 \end{aligned}$$

We can now construct the SMS $\{x_0, v(t)\}$ and $\{\bar{x}_0, \bar{v}(t)\}$.

First consider the interval $t \in [t_0, t_1]$. Let $\mathcal{Y}_{2n}(t)$ and $\bar{\mathcal{Y}}_{2n}(t)$ be the outputs and $2n-1$ derivatives of $\{x_0, v(t)\}$ and $\{\bar{x}_0, \bar{v}(t)\}$, respectively. Then

$$\begin{aligned} &\mathcal{Y}_{2n}(t) - \bar{\mathcal{Y}}_{2n}(t) = \\ &\qquad \begin{bmatrix} \mathcal{O}_{2n}(k_{s1}) & \mathcal{O}_{2n}(k_{s3}) \end{bmatrix} \begin{bmatrix} e^{A_{k_{s1}}(t-t_1)} x_1 \\ e^{A_{k_{s3}}(t-t_1)}(-\bar{x}_1) \end{bmatrix} \end{aligned} \tag{2.66}$$

If (2.66) is equal to zero for all $t \in [t_0, t_1)$ then lemma 2.14 guarantees $\mathcal{Y}_{2n}(t) = \bar{\mathcal{Y}}_{2n}(t)$, i.e. the outputs are identically equal. To prove (2.66) is equal to zero, we define a new LTI system $(\tilde{C}, \tilde{A})$ where

$$\tilde{C} = \begin{bmatrix} C_{k_{s1}} & C_{k_{s3}} \end{bmatrix}, \qquad \tilde{A} = \begin{bmatrix} A_{k_{s1}} & 0 \\ 0 & A_{k_{s3}} \end{bmatrix}$$

Note that the observability matrix of $(\tilde{C}, \tilde{A})$ is exactly the matrix

$$\tilde{R} = \begin{bmatrix} \mathcal{O}_{2n}(k_{s1}) & \mathcal{O}_{2n}(k_{s3}) \end{bmatrix}$$

in (2.66). From linear system theory, the null space of the observability matrix is $\tilde{A}$-invariant. That is $\tilde{A}x \in Null(\tilde{R})$ if $x \in Null(\tilde{R})$. By the Cayley-Hamilton theorem, $\tilde{A}$ satisfies its own differential equation and $\tilde{A}^k$ for $k \geq 2n$ can be written as a linear combination of lower powers of $\tilde{A}$. Since $e^{\tilde{A}t} = \sum_{j=0}^{\infty} 1/j! \tilde{A}^j t^j$ by definition, the Cayley-Hamilton theorem combined with the $\tilde{A}$-invariance of $\tilde{R}$ implies

$$\begin{aligned} 0 &= \begin{bmatrix} \mathcal{O}_{2n}(k_{s1}) & \mathcal{O}_{2n}(k_{s3}) \end{bmatrix} e^{\tilde{A}(t-t_1)} \begin{bmatrix} x_1 \\ -\bar{x}_1 \end{bmatrix} \\ &= \begin{bmatrix} \mathcal{O}_{2n}(k_{s1}) & \mathcal{O}_{2n}(k_{s3}) \end{bmatrix} \begin{bmatrix} e^{A_{k_{s1}}(t-t_1)} x_1 \\ e^{A_{k_{s3}}(t-t_1)}(-\bar{x}_1) \end{bmatrix} \end{aligned} \tag{2.67}$$

By (2.66) and (2.67), lemma 2.14 implies that the outputs of $\{x_0, v(t)\}$ and $\{\bar{x}_0, \bar{v}(t)\}$ are identically equal over $[t_0, t_1]$. By the same argument, it can be shown that the outputs are identically equal over the interval (t_1, t_f). Thus $\{x_0, v(t)\}$ and $\{\bar{x}_0, \bar{v}(t)\}$ are indistinguishible implying Σ is ST unobservable completing the proof of necessity. □

3. STRUCTURED ROBUST PROPERTY METRIC: $\mathcal{P}$-ROBUSTNESS

3.1 Introduction

System properties, such as controllability and observability, are often characterized by binary labels, e.g., controllable or uncontrollable and stable or unstable. These binary labels fail to capture the robustness of these properties. For example, consider the LTI system

$$\dot{x}(t) = Ax(t) + Bu(t), \tag{3.1}$$

where $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$. The pair (A, B) is controllable if and only if for each $\lambda \in \mathbb{C}$

$$\operatorname{rank}\begin{bmatrix} A - \lambda I_n & B \end{bmatrix} = n, \tag{3.2}$$

where I_n denotes the $n \times n$ identity matrix [23]. The set of uncontrollable pairs $(A, B) \in \mathbb{R}^{n\times(n+m)}$, i.e., pairs failing to satisfy (3.2), is an algebraic variety of lower dimension and hence has measure zero in $\mathbb{R}^{n\times(n+m)}$. Since LTI models are only approximations of physical systems, it is also necessary to characterize the robustness of the controllability property, e.g., by determining the distance to the nearest uncontrollable system [24, 25]

$$\mu_{\mathbb{R}}(A, B) = \inf_{(\delta A, \delta B)\in\mathcal{C}} \|[\delta A, \delta B]\|_F, \tag{3.3}$$

where $\mathcal{C} = \{(\delta A, \delta B) : \exists\lambda \in \mathbb{C}, \operatorname{rank}[A - \delta A - \lambda I_n, B - \delta B] < n\}$. Similar metrics can be constructed for system properties including, but not limited to reachability, stabilizability, observability, and detectability.

Computing metrics such as $\mu_{\mathbb{R}}(A, B)$ and the associated minimizing perturbations has been an active area of research over the past 40 years [24–37]. The norm used to measure robustness separates the robust system property literature. The Frobenius norm metric in (3.3) is based on the work of [24] and is used in [25–31]. The primary alternative to the Frobenius norm metric is the spectral norm, i.e., the largest singular value of the matrix $[\delta A, \delta B]$. The spectral norm metric, usually referred to by the names controllability radius or observability radius, is explored in several works including [31–36]. This paper utilizes a Frobenius norm metric because a perturbation on each entry of a system matrix affects the Frobenius norm in a strong and direct way.

The primary challenge to either robustness metric is developing an algorithm to compute the minimum distance and associated perturbation matrices. In [25], the algorithm for computing (3.3) for real but otherwise unstructured perturbations is based on computing a coordinate transformation into a "nearly" Kalman uncontrollable form. Another approach for computing (3.3) is considered in [29] wherein one constructs a large $n(n+1) \times n(n+m)$ matrix X_{n-1} consisting of a structured arrangement of blocks of matrices

$$\begin{bmatrix} A & B \\ I & 0 \end{bmatrix}.$$

The "Structured Total Least Norm" algorithm then computes a low rank approximation to X_{n-1} where only the A and B matrices are potentially perturbed. The low rank approximation also provides the smallest perturbations $\delta A \in \mathbb{R}^{n\times n}$ and $\delta B \in \mathbb{R}^{n\times m}$ causing uncontrollability.

Reference [32] develops an algorithm for computing the controllability radius for real but otherwise unstructured perturbations utilizing a constrained optimization problem; the perturbations causing uncontrollability are constructed from singular vectors. In [36], a fast algorithm for computing the controllability radius is developed for unstructured complex perturbations. Extensions to higher-order LTI systems with affine perturbations are considered in [33]. Additional extensions including descriptor and time-delay LTI systems are considered in [34]. Finally, [31] develops an upper bound on the spectral distance to uncontrollability of a switched LTI system.

Reference [38] formulates the problem of structured rank reducing perturbations, belonging to a subspace $\mathcal{S} \subset \mathbb{C}^{n \times m}$, on a rectangular matrix $M \in \mathbb{C}^{n \times m}$ which cause the failure of a system property, $\mathcal{P}$, such as controllability, observability, or stability. This general $\mathcal{P}$–robustness framework encompasses many of the robustness problems previously addressed in the literature. No prior work has extended the $\mathcal{P}$–robustness framework, proven the necessary conditions for $\mathcal{P}$–robustness, or completed and proven convergence of the algorithm suggested in [38]; this list constitutes the main contributions of the current paper.

Section 3.2 introduces the $\mathcal{P}$–robustness framework. Section 3.3 establishes necessary conditions for solving the $\mathcal{P}$–robustness problem. The necessary conditions motivate an algorithm for solving the $\mathcal{P}$–robustness problem introduced in Section 3.4. The algorithm converges to a point satisfying the necessary conditions and is demonstrated with numerical examples in Section 3.6.

In this paper, the following notation will be used:

$\|\|M\|\|_F$	Frobenius norm of a matrix M.
$\|\|\nu\|\|_2$	Euclidean norm of a vector ν.
$M^\top$, M^H	Transpose and conj. transpose of M.
$\sigma_n(M)$	n^{th} singular value of matrix M.
I_m	$m \times m$ identity matrix.
$\text{diag}(\nu)$	Diag. matrix with diag. entries in ν.
$\text{vec}(M)$	Vectorizes M by stacking the columns.
$M \otimes N$	The kronecker product of M and N.
$M^\dagger$	The Moore-Penrose pseudoinverse.
$\text{Im}(M)$	Imag. component of M.
$\text{Re}(M)$	Real component of M.
$\sigma_i(M)$	i^{th} largest singular value of M.
$\text{cl}(U)$	The closure of the set U.
$\langle A_1, A_2 \rangle$	Inner product of $A_1, A_2 \in \mathbb{C}^{n \times m}$ defined as $\langle A_1, A_2 \rangle = \text{Re}(\text{vec}(A_1))^\top \text{Re}(\text{vec}(A_2)) + \text{Im}(\text{vec}(A_1))^\top \text{Im}(\text{vec}(A_2))$

3.2 $\mathcal{P}$-Robustness Problem

This section specifies the details of the $\mathcal{P}$–robustness problem. Later sections restate and rigorously prove necessary conditions for its solution (Theorem 3.2) and develop a complete algorithm (Algorithm 1) that converges to a point satisfying those necessary conditions, under appropriate assumptions.

Definition 3.1. *[38] Let $M \in \mathbb{C}^{n\times m}$ with $n \leq m$ (without loss of generality); let $\mathcal{P} \subset \mathbb{C}^{n\times m}$ and $\mathcal{S} \subset \mathbb{C}^{n\times m}$ be linear spaces over $\mathbb{R}$. The $\mathcal{P}$–robustness of M with respect to parameter variations in $\mathcal{S}$ is defined as*

$$r(M;\mathcal{S},\mathcal{P}) = \inf_{\delta M\in\mathcal{T}} \|\delta M\|_F \tag{3.4}$$

where

$$\mathcal{T} = \{\delta M \in \mathcal{S} : \exists R \in \mathcal{P}, \operatorname{rank}[M - \delta M - R] < n\} \tag{3.5}$$

As mentioned, the Frobenius norm metric, used herein, directly measures the magnitude of the parameter variations and thus appears to more accurately represent the robustness of the system property. This is in contrast to the controllability (and observability) radius which measures the largest singular value of the perturbation causing uncontrollability (unobservability), a metric that may not reflect some parameter variations: for a fixed largest singular value, changes in the smaller singular values due to parameter variations go unnoticed.

It is useful to consider bases for $\mathcal{S}$ and $\mathcal{P}$ (which are linear subspaces over the field $\mathbb{R}$). Let $\{S_1, S_2, \cdots, S_k\}$ be an orthonormal basis for $\mathcal{S}$ and $\{P_1, P_2, \cdots, P_r\}$ be an orthonormal basis for $\mathcal{P}$, where by orthonormal we mean that $\langle S_i, S_j\rangle \triangleq \operatorname{Re}(\operatorname{vec}(S_i))^\top \operatorname{Re}(\operatorname{vec}(S_j)) + \operatorname{Im}(\operatorname{vec}(S_i))^\top \operatorname{Im}(\operatorname{vec}(S_j))$ is 0 if $i \neq j$ and 1 if $i = j$. Each perturbation $\delta M \in \mathcal{S}$ can be represented by an associated vector $\zeta \in \mathbb{R}^k$ in this basis $\{S_1, S_2, \cdots, S_k\}$, i.e., $\delta M = \sum_{i=1}^{k} \zeta_i S_i$. Similarly, each $R \in \mathcal{P}$ is represented by a vector $\rho \in \mathbb{R}^r$. Using the fixed bases for $\mathcal{S}$ and $\mathcal{P}$, we can reformulate the $\mathcal{P}$–robustness of M with respect to perturbations in $\mathcal{S}$.

Definition 3.2. *Let $M \in \mathbb{C}^{n\times m}$; let $\mathcal{P} \subset \mathbb{C}^{n\times m}$ and $\mathcal{S} \subset \mathbb{C}^{n\times m}$ be linear spaces over $\mathbb{R}$ with orthonormal bases $\{S_1, S_2, \cdots, S_k\}$ and $\{P_1, P_2, \cdots, P_r\}$, respectively. The $\mathcal{P}$–robustness of M with respect to parameter variations in $\mathcal{S}$ is*

$$r(M; \mathcal{S}, \mathcal{P}) = \inf_{\zeta\in\mathbb{R}^k,\, \rho\in\mathbb{R}^r} \|\zeta\|_2 \tag{3.6}$$

subject to:

$$0 = \sigma_n\left(M - \sum_{i=1}^{k} \zeta_i S_i - \sum_{j=1}^{r} \rho_j P_j\right) \triangleq H(\zeta, \rho). \tag{3.7}$$

Note, Definition 3.2 is equivalent to Definition 3.1. Also, the $\mathcal{P}$–robustness of M with respect to parameter variations in $\mathcal{S}$ is independent of the orthonormal basis. Of course, the representation of the minimizing pair (ζ_*, ρ_*) depends on the selected bases for $\mathcal{S}$ and $\mathcal{P}$.

For use later in the paper, we define

$$f(\zeta, \rho) = 0.5\|\zeta\|_2^2. \tag{3.8}$$

A minimizer (ζ_*, ρ_*) to (3.6) subject to (3.7) (when it exists) is the same when $\|\zeta\|_2$ is replaced by $f(\zeta, \rho)$ in (3.6). The function $f(\zeta, \rho)$ is preferable because it simplifies the proofs later in the paper.

Example 3.1. *Applying the $\mathcal{P}$–robustness formulation to controllability of an LTI state model (A, B, C), we set $M = [A, B]$, $R = [\lambda I, 0]$, and $\delta M = [\delta A, \delta B]$, where δM has a specific perturbation structure defined by a basis for $\mathcal{S}$.*

The original motivation for this work stems from the need for specific structured real perturbations for the state and mode sequence (SMS) observability problem of switched LTI (SLTI) systems with safety applications described in [20, 30]. SMS observability of SLTI systems can also be formulated as a $\mathcal{P}$–robustness problem:

Example 3.2. *For the problem of computing the distance to the nearest SMS unobservable SLTI system (which can model transitions from safe to unsafe operation), the*

$\mathcal{P}$–robustness framework can be applied to each pair of modes i (safe) and j (unsafe) by defining

$$M_{ij} = \begin{bmatrix} A_i^\top & 0 & C_i^\top \\ 0 & A_j^\top & C_j^\top \end{bmatrix} \in \mathbb{R}^{2n\times(2n+p)}$$

$$\delta M_{ij} = -\begin{bmatrix} \delta A_i^\top & 0 & \delta C_i^\top \\ 0 & \delta A_j^\top & \delta C_j^\top \end{bmatrix} \in \mathbb{R}^{2n\times(2n+p)}$$

$$R = \begin{bmatrix} \lambda I & 0 & 0 \\ 0 & \lambda I & 0 \end{bmatrix} \in \mathbb{C}^{2n\times(2n+p)}.$$

Clearly, the perturbation δM_{ij} has a specialized structure that is problematic for most existing approaches. The $\mathcal{P}$–robustness of M_{ij} with respect to parameter variations $\delta M_{ij} \in \mathcal{S}$ provides $r_{ij} \triangleq r(M_{ij}, \mathcal{S}, \mathcal{P})$, see (3.4). Then $\min_{i,j}\{r_{ij}\}$ is exactly the distance to the nearest SMS unobservable SLTI system.

The rank reduction in the $\mathcal{P}$–robustness problem is characterized by the n^{th} singular value of $M - \delta M - R$ becoming zero. To analyze the n^{th} singular value, we define the following linear operator:

Definition 3.3. *Each pair of matrices $u \in \mathbb{C}^n$ and $V \in \mathbb{C}^{m\times(m-n+1)}$ induces a linear operator $L_{uV} : \mathbb{C}^{n\times m} \to \mathbb{C}^{1\times(m-n+1)}$ given by*

$$L_{uV}(N) = u^H N V. \tag{3.9}$$

Proposition 3.1. *Let $N = \widehat{U}\widehat{\Sigma}\widehat{V}^H$ be the singular value decomposition of N. Define u to be the last column of $\widehat{U}$ and V to be the last $m - n + 1$ columns of $\widehat{V}$. Then*

$$L_{uV}(N) = \begin{bmatrix} \sigma_n(N) & 0 & \cdots & 0 \end{bmatrix}.$$

Consequently, $\|L_{uV}(N)\|_F = \sigma_n(N)$.

The linear operator L_{uV} is defined for any u and V, independent of the argument. For example, u and V can be related to the singular value decomposition of a matrix $M - \delta M - R$ and operate on any matrix $N' \in \mathbb{C}^{n\times m}$. Since the perturbations and property matrices belong to lower dimensional subspaces $\mathcal{S}$ and $\mathcal{P}$, we define additional linear operators that have domains restricted to these subspaces.

Definition 3.4. *For $u \in \mathbb{C}^n$ and $V \in \mathbb{C}^{m\times(m-n+1)}$, the linear operators $L_{uV\mathcal{S}} : \mathcal{S} \to \mathbb{C}^{1\times(m-n+1)}$ and $L_{uV\mathcal{P}} : \mathcal{P} \to \mathbb{C}^{1\times(m-n+1)}$ are defined as*

$$L_{uV\mathcal{S}}(\delta M) \triangleq L_{uV}|_{\mathcal{S}}(\delta M) = u^H \delta M V$$
$$L_{uV\mathcal{P}}(R) \triangleq L_{uV}|_{\mathcal{P}}(R) = u^H R V.$$

The distinctions of the operator domains are pertinent when considering the pseudoinverses $L^{\dagger}_{uV\mathcal{S}} : \mathbb{C}^{1\times(m-n+1)} \to \mathcal{S}$ and $L^{\dagger}_{uV\mathcal{P}} : \mathbb{C}^{1\times(m-n+1)} \to \mathcal{P}$. The map $L_{uV\mathcal{S}}$ is surjective if for each $y \in \mathbb{C}^{1\times(m-n+1)}$ there exists $\delta M \in \mathcal{S}$ such that $L_{uV\mathcal{S}}(\delta M) = y$. When $L_{uV\mathcal{S}}$ is surjective, the pseudoinverse map $L^{\dagger}_{uV\mathcal{S}}(y) = \delta M$ is the smallest matrix $\delta M \in \mathcal{S}$ (in the Frobenius norm sense) solving the equation $L_{uV\mathcal{S}}(\delta M) = y$.

Fundamental to the solution of the $\mathcal{P}$–robustness problem is the surjectivity of a family of maps $\{L_{uV\mathcal{S}}\}$ as per the following assumption:

Assumption 3.1. *Let $\delta M \in \mathcal{S}$ and $R \in \mathcal{P}$. Let $M - \delta M - R$ have singular value decomposition $\widehat{U}\widehat{\Sigma}\widehat{V}^H$. Define u to be the n^{th} column of $\widehat{U}$ and V to be the last $m - n + 1$ columns of $\widehat{V}$. Then we assume $L_{uV\mathcal{S}}$ is surjective for each $\delta M \in \mathcal{S}$ and every $R \in \mathcal{P}$.*

We would like to explain why Assumption 3.1 is appropriate. Using Kronecker product notation, $L_{uV\mathcal{S}}$ is surjective if and only if

$$\operatorname{rank} \begin{bmatrix} \operatorname{Re}\left[(V^\top \otimes u^H)B_{\mathcal{S}}\right] \\ \operatorname{Im}\left[(V^\top \otimes u^H)B_{\mathcal{S}}\right] \end{bmatrix} = 2(m-n+1). \tag{3.10}$$

where $\{S_1, \cdots, S_k\}$ is a basis for $\mathcal{S}$ and $B_{\mathcal{S}} \triangleq [\operatorname{vec}(S_1), \cdots, \operatorname{vec}(S_k)]$. Clearly, $B_{\mathcal{S}}$ must have at least $2(m-n+1)$ columns for (3.10) to be satisfied, i.e., $\mathcal{S}$ as a vector space over $\mathbb{R}$ must have dimension no less than $2(m-n+1)$. Consequently, for the problem to be solvable, we require that the perturbation space $\mathcal{S}$ be sufficiently rich. The proof of (3.10) and related surjectivity results are included in the appendix.

As described in [38], the surjectivity of $L_{uV\mathcal{S}}$ ensures a certain regularity condition on a rank reducing perturbation/property matrix pair $(\delta M, R) \in \mathcal{S} \times \mathcal{P}$. This regularity condition guarantees that there are neighboring perturbation/property matrix pairs $(\delta M', R')$ which are also rank reducing, i.e., δM is not an isolated rank reducing perturbation.

If a rank reducing perturbation is isolated it is naturally a local minimum. A $\mathcal{P}$–robustness problem with a finite number of isolated extrema is much easier to solve: one can find $\zeta \in \mathbb{R}^k$ and $\rho \in \mathbb{R}^r$ such that $\det[M(\zeta,\rho)M(\zeta,\rho)^\top] = 0$, where

$$M(\zeta,\rho) = M - \sum_{i=1}^{k} \zeta_i S_i - \sum_{i=1}^{r} \rho_i P_i. \tag{3.11}$$

Consequently, we focus on $\mathcal{P}$–robustness problems satisfying the surjectivity assumption. In addition, we can exclude $\mathcal{P}$–robustness problems where $\operatorname{rank}(M - R) < n$ for some $R \in \mathcal{P}$ (i.e., M does not satisfy the property $\mathcal{P}$) since $r(M; \mathcal{S}, \mathcal{P}) = 0$ in this case. The next section sets up necessary conditions for the solution to the $\mathcal{P}$–robustness problem.

3.3 Necessary Conditions

The objective of this section is proving the necessary conditions on a minimum norm rank-reducing perturbation $\delta M_* \in \mathcal{S}$ and the associated property matrix $R_* \in \mathcal{P}$ (when they exist), i.e., $\|\delta M_*\|_F = r(M; \mathcal{S}, \mathcal{P})$ and $\operatorname{rank}(M - \delta M_* - R_*) < n$. We next provide some intuition for the necessary conditions.

Let us first assume that the property matrix R_* is fixed. For δM_* to be the minimum norm rank-reducing perturbation for $M - R_*$, the tangent plane to the hypersurface $\Upsilon_1 = \{\delta M \in \mathcal{S} : \sigma_n(M - R_* - \delta M) = 0\}$ must be perpendicular to the line connecting $M - R_*$ and $M - R_* - \delta M_*$, see Figure 3.1. For u_* the n^{th} left singular vector (lsv) and V_* having columns equal to the n^{th} through m^{th} right singular vectors (rsv) of $M - R_* - \delta M_*$, $\|L_{u_* V_*}(M - R_* - \delta M_*)\|_F = \sigma_n(M - R_* - \delta M_*) = 0$. As will be seen in the proof of Theorem 3.2, the hyperplane $\Upsilon_2 = \{\delta M \in \mathcal{S} : L_{u_* V_*}(M - R_* - \delta M) = 0\}$ is related to the tangent plane to Υ_1 at δM_*. Note, elements of Υ_2 are precisely the minima of $\|L_{u_* V_* \mathcal{S}}(\delta M) - L_{u_* V_*}(M - R_*)\|_F$ over $\delta M \in \mathcal{S}$. If δM_* is the smallest rank reducing perturbation on $M - R_*$, then δM_* has the smallest norm of any perturbation in Υ_2, i.e., δM_* minimizes $\|L_{u_* V_* \mathcal{S}}(\delta M_*) - L_{u_* V_*}(M - R_*)\|_F$

and $\|\delta M_*\|_F$ has the least norm of all such matrices. Since $L_{u_*V_*\mathcal{S}}$ is surjective by Assumption 3.1, the minimum is given by

$$\delta M_* = (L^\dagger_{u_*V_*\mathcal{S}} \circ L_{u_*V_*})(M - R_*), \tag{3.12}$$

the first necessary condition in Theorem 3.2.

The second necessary condition addresses the locally optimal property matrix R_*. As per the discussion above, the optimal rank reducing perturbation satisfies (3.12). Let $\Delta R \in \mathcal{P}$ be an alteration to R_*. Define

$$\delta M_0(\Delta R) \triangleq (L^\dagger_{u_0V_0\mathcal{S}} \circ L_{u_0V_0})(M - R_* - \Delta R) \tag{3.13}$$

where u_0 is the n^{th} lsv and V_0 has columns equal to the n^{th} through m^{th} rsv of $M - R_* - \Delta R - \delta M_0(\Delta R)$. It is difficult to directly minimize the norm of (3.13) with respect to ΔR because the matrices u_0 and V_0 change with ΔR. However, for sufficiently small ΔR, we will approximate $\delta M_0(\Delta R)$ with

$$\delta M_0(\Delta R) \approx (L^\dagger_{u_*V_*\mathcal{S}} \circ L_{u_*V_*})(M - R_* - \Delta R), \tag{3.14}$$

where u_* is the n^{th} lsv and V_* has columns equal to the n^{th} through m^{th} rsv of $M - R_* - \delta M_*$. Thus in a sufficiently small neighborhood of R_*, minimizing the norm of (3.13) with respect to ΔR is equivalent to minimizing

$$\|(L^\dagger_{u_*V_*\mathcal{S}} \circ L_{u_*V_*\mathcal{P}})\Delta R - (L^\dagger_{u_*V_*\mathcal{S}} \circ L_{u_*V_*})(M - R_*)\|_F. \tag{3.15}$$

The least square minimum of (3.15) is given by

$$\Delta R = (L^\dagger_{u_*V_*\mathcal{S}} \circ L_{u_*V_*\mathcal{P}})^\dagger (L^\dagger_{u_*V_*\mathcal{S}} \circ L_{u_*V_*})(M - R_*). \tag{3.16}$$

If R_* is the optimal property matrix, then (3.15) is minimized at $\Delta R = 0$. Hence, the right-hand side of (3.16) equals zero, the second necessary condition.

Theorem 3.2. *Suppose there exists $\delta M_* \in \mathcal{S}$ that is a local minimum norm element of the set $\mathcal{T} = \{\delta M \in \mathcal{S} : \exists R \in \mathcal{P}, \mathrm{rank}[M - \delta M - R] < n\}$; choose $R_* \in \{R \in \mathcal{P} : \mathrm{rank}[M - \delta M_* - R] < n\}$ and let u be a non-trivial element of $\ker[(M - \delta M_* - R_*)^H]$ and let V be a matrix whose columns span $\ker[M - \delta M_* - R_*]$. If $L_{uV\mathcal{S}}$ is surjective, then the following two necessary conditions both hold:*

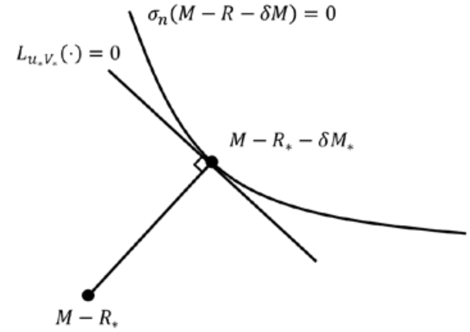

Fig. 3.1. This figure illustrates the first necessary condition. Given appropriate assumptions, the surface $L_{u_* V_*}(\cdot) = 0$ is tangent to the curve $\sigma_n(M - R_* - \delta M) = 0$ for $\delta M \in \mathcal{S}$ at $M - R_* - \delta M_*$, where u_* is the n^{th} lsv and V_* is the n^{th} through m^{th} rsv of $M - R_* - \delta M_*$. For δM_* to be a local minimum rank reducing perturbation for the property matrix R_* the line connecting $M - R_*$ to $M - R_* - \delta M_*$ must be perpendicular to the tangent surface $L_{u_* V_*}(\cdot) = 0$.

1a) $\delta M_* \in \mathcal{S}$ *is a minimum norm matrix minimizing*

$$\|L_{uV\mathcal{S}}(\delta M_*) - L_{uV}(M - R_*)\|_F.$$

2a) $0 = \Delta R_*$, *where* $\Delta R_* \in \mathcal{P}$ *is the minimum norm matrix minimizing*

$$\|(L^{\dagger}_{uV\mathcal{S}} \circ L_{uV\mathcal{P}})(\Delta R_*) - (L^{\dagger}_{uV\mathcal{S}} \circ L_{uV})(M - R_*)\|_F.$$

Equivalently,

1b) $\delta M_* = (L^{\dagger}_{uV\mathcal{S}} \circ L_{uV})(M - R_*)$ *and*

2b) $0 = \Delta R_* = (L^{\dagger}_{uV\mathcal{S}} \circ L_{uV\mathcal{P}})^{\dagger}(L^{\dagger}_{uV\mathcal{S}} \circ L_{uV})(M - R_*)$.

Remark 3.1. *Conditions 1a and 2a could be generalized to other norms, such as the spectral norm. On the other hand, conditions 1b and 2b are Frobenius-norm specific.*

Note, condition 1a essentially requires δM_* to be the smallest matrix minimizing $\sigma_n(M - R_* - \delta M)$ for $\delta M \in \mathcal{S}$. So even when no rank reducing perturbation exists, condition 1a provides the “best” solution.

The proof of Theorem 3.2 requires some machinery and four technical lemmas. As will be seen, proving the necessary conditions in Theorem 3.2 requires the application of the inverse function theorem[1] which in turn requires Fréchet differentiability of the equality constraint $H(\zeta, \rho) = 0$ in (3.7). Unfortunately, there are points at which H is only directionally differentiable. These non-Fréchet differentiable points are caused by two structural components of the svd: i) the ordering of the singular values and ii) the requirement that the singular values be positive. We observe that, in general, perturbation and property matrices δM and R for which $M - \delta M - R$ has a repeated smallest singular value or a zero smallest singular value is an algebraic variety of lower dimension in $\mathcal{S} \times \mathcal{P}$. Consequently, the function $H(\zeta, \rho)$ is Fréchet differentiable almost

[1] A simpler proof of the necessary conditions which does not require the inverse function theorem can be developed if the conditions in Proposition 3.16 (in the appendix) are satisfied. This simpler proof is consistent with the proof outlined in [38], but requires conditions stronger than surjectivity.

everywhere. Since we are concerned with rank reducing perturbations, we need to resolve the non-Fréchet differentiability when $H(\zeta,\rho)=0$.

De Moor and Boyd in [39] suggest an alternative svd that relaxes the reordering of the singular values/vectors and positivity of the singular values. The focus of [39] is computing analytic unsigned and unordered singular value decompositions along an analytic path. These results on analytic paths are extended herein to an open set in $\mathbb{R}^k \times \mathbb{R}^r$; in this way, we can construct a Fréchet differentiable function $\widetilde{H}(\zeta,\rho)$ which is zero exactly when $H(\zeta,\rho)=0$.

Let $(\delta M_0, R_0) \in \mathcal{S} \times \mathcal{P}$ be a pair matrices which satisfy

$$\operatorname{rank}(M - \delta M_0 - R_0) = n - 1.$$

Let $(\zeta_0,\rho_0) \in \mathbb{R}^k \times \mathbb{R}^r$ represent $(\delta M_0, R_0)$ in the bases $\{S_1,\cdots,S_k\}$ and $\{P_1,\cdots,P_r\}$, respectively. To simplify the notation, define the map $\sigma_n : \mathbb{R}^k \times \mathbb{R}^r \times \mathbb{C}^{n\times m} \to \mathbb{R}$ given by

$$\sigma_n(\zeta,\rho;M) \triangleq \sigma_n\left(M(\zeta,\rho)\right), \tag{3.17}$$

where $M(\zeta,\rho)$ has been defined in (3.11). Since $\sigma_n(M-\delta M_0-R_0)$ is distinct and $\sigma_n(\cdot)$ is continuous everywhere, there exists a simply-connected and open neighborhood $W \subset \mathbb{R}^k \times \mathbb{R}^r$ of (ζ_0,ρ_0) sufficiently small such that

1. for each $(\zeta,\rho) \in W$, $\sigma_n(\zeta,\rho;M)$ (the smallest singular value) is distinct,
2. there exists simply-connected and open subsets $W_1, W_2 \subset W$ such that
   (a) $W \subset \operatorname{cl}(W_1 \cup W_2)$,
   (b) W_1 and W_2 are disjoint, and
   (c) for each $(\zeta,\rho) \in W_1 \cup W_2$, $\sigma_n(\zeta,\rho;M) > 0$.

The simply-connected and open subsets W, W_1, and W_2 are illustrated in Figure 3.2. Note that the open set W includes the common boundary of the open sets W_1 and W_2.

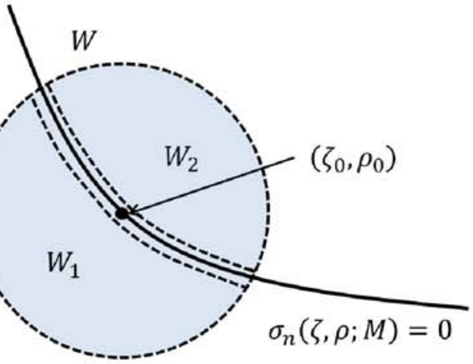

Fig. 3.2. This figure illustrates the simply-connected neighborhood W of (ζ_0, ρ_0) partitioned three simply connected regions: W_1 and W_2 disjoint open sets with $W \subset \text{cl}(W_1 \cup W_2)$ and the surface $W \cap \{\sigma_n(\cdot) = 0\}$.

Let $g : [0,1] \to W$ be an analytic function with $g(s) \in W_1$ for $s < 0.5$ and $g(s) \in W_2$ for $s > 0.5$. According to [39, Theorem 1], there exists an analytic function $f_g : [0,1] \to \mathbb{R}$, called the unsigned n^{th} singular value function, such that

$$|f_g(s)| = \sigma_n(g(s); M), \quad s \in [0,1].$$

The function f_g can only change sign as it transitions through the common boundary of W_1 and W_2, i.e., at $s = 0.5$. By [39, Theorem 3], there exists analytic singular vector functions $u_g : [0,1] \to \mathbb{C}^n$ and $v_g : [0,1] \to \mathbb{C}^m$ such that for each $s \in [0,1]$, $u_g(s)$ and $v_g(s)$ are the unsigned n^{th} lsv and rsv associated with $f_g(s)$, i.e., $u_g(s)$ and $v_g(s)$ are unit vectors satisfying

$$u_g^H(s)\left(M - \sum_{i=1}^{k} \zeta_{gi}(s) S_i - \sum_{j=1}^{r} \rho_{gj}(s) P_j\right) = f_g(s) v_g^H(s),$$

for each $s \in [0,1]$, where $\zeta_g : [0,1] \to \mathbb{R}^k$ and $\rho_g(s) \to \mathbb{R}^r$ are defined by the relation $g \triangleq (\zeta_g, \rho_g)$.

Let $\widetilde{H} : W \to \mathbb{R}$ be the extension of the unsigned singular value function f_g to the set W, i.e.,

$$\widetilde{H}(\zeta, \rho) = \begin{cases} \operatorname{sign}(f_g(0))\sigma_n(\zeta, \rho; M) & (\zeta, \rho) \in W_1 \\ \operatorname{sign}(f_g(1))\sigma_n(\zeta, \rho; M) & \text{otherwise} \end{cases}. \tag{3.18}$$

Note that, $\widetilde{H}(g(s)) = f_g(s)$ for each $s \in [0,1]$ since by construction of $g(s)$, f_g can change sign only at $s = 0.5$. In addition, the form of $\widetilde{H}$ implies that for each $(\zeta, \rho) \in W$, $|\widetilde{H}(\zeta, \rho)| = \sigma_n(\zeta, \rho)$. We will show that $\widetilde{H}$ is Fréchet differentiable on W, as per the following lemma.

Lemma 3.3. *Let $(\zeta_0, \rho_0) \in \mathbb{R}^k \times \mathbb{R}^r$ satisfy* $\operatorname{rank}[M(\zeta_0, \rho_0)] = n-1$*. Let $W \subset \mathbb{R}^k \times \mathbb{R}^r$ be as defined above. Let $\widetilde{H} : W \to \mathbb{R}$ be as in* (3.18)*. Then $\widetilde{H}$ is Fréchet differentiable on W with partial derivatives given by*

$$\begin{aligned} \frac{\partial \widetilde{H}(\zeta, \rho)}{\partial \zeta_i} &= -\operatorname{Re}(u^H S_i v) \\ \frac{\partial \widetilde{H}(\zeta, \rho)}{\partial \rho_i} &= -\operatorname{Re}(u^H P_i v), \end{aligned} \tag{3.19}$$

where u and v are unsigned n^{th} lsv and rsv of $M(\zeta_0, \rho_0)$, i.e., $u^H M(\zeta_0, \rho_0) = \widetilde{H}(\zeta, \rho)v^H$ and $M(\zeta_0, \rho_0)v = \widetilde{H}(\zeta, \rho)u$.

Proof. See appendix. □

As per Lemma 3.3, we now consider replacing the equality constraint $H(\zeta, \rho) = 0$ with $\widetilde{H}(\zeta, \rho) = 0$ on the set W. For each $(\zeta_0, \rho_0) \in W$, $\widetilde{H}$ has Fréchet derivative $\widetilde{H}'(\zeta_0, \rho_0)$ given by

$$-\operatorname{Re}\left[u^H S_1 v, \cdots, u^H S_k v, u^H P_1 v, \cdots, u^H P_r v\right], \tag{3.20}$$

where u and v are unsigned n^{th} lsv and rsv of $M - \delta M_0 - R_0$ with $\delta M_0 \in \mathcal{S}$ and $R_0 \in \mathcal{P}$ the matrices represented by ζ_0 and ρ_0, respectively.

The next lemma proves that if the condition 1a (or equivalently 1b) of Theorem 3.2 is not satisfied, then there exists a direction $\Delta M \in \mathcal{S}$ to change the perturbation δM_* on the tangent plane $L_{uV}(\cdot) = 0$ (see Figure 3.1). This new perturbation $\delta M_* + \widetilde{\Delta M}$ may not be rank reducing, but will allow us to prove the existence of rank reducing perturbations with norms smaller than $\|\delta M_*\|_F$.

Lemma 3.4. *Let $M \in \mathbb{C}^{n\times m}$, $\delta M_0 \in \mathcal{S}$, $R_0 \in \mathcal{P}$ satisfy* $\operatorname{rank}[M - \delta M_0 - R_0] < n$ *with $L_{uV\mathcal{S}}$ surjective, where u is the n^{th} lsv and V has columns equal to the n^{th} through m^{th} rsv of $M - \delta M_0 - R_0$. Suppose $\delta M_0 \neq (L^\dagger_{uV\mathcal{S}} \circ L_{uV})(M - R_0)$. Then there exists a matrix $\Delta M \in \mathcal{S}$ such that*

$$L_{uV\mathcal{S}}(\Delta M) = 0 \tag{3.21}$$

and

$$\langle \delta M_0, \Delta M\rangle < 0. \tag{3.22}$$

Proof. See appendix. □

Similar to Lemma 3.4, the following lemma proves that if the second necessary condition of Theorem 3.2 is not satisfied, then there exist directions ΔM and ΔR for changing perturbation δM_* and the property matrix R_*, respectively, reducing the norm of the perturbation on the tangent surface $L_{uV}(\cdot) = 0$. This will allow us to prove to existence of a rank reducing perturbation with smaller norm (and associated property matrix).

Lemma 3.5. *Let $M \in \mathbb{C}^{n\times m}$, $\delta M_0 \in \mathcal{S}$, $R_0 \in \mathcal{P}$ satisfy* $\operatorname{rank}[M - \delta M_0 - R_0] < n$ *with $L_{uV\mathcal{S}}$ surjective, where u is the n^{th} lsv and V has columns equal to the n^{th} through m^{th} rsv of $M - \delta M_0 - R_0$. Suppose $\delta M_0 = (L^{\dagger}_{uV\mathcal{S}} \circ L_{uV})(M - R_0)$ and*

$$0 \neq \Delta R \triangleq (L^{\dagger}_{uV\mathcal{S}} \circ L_{uV\mathcal{P}})^{\dagger}(L^{\dagger}_{uV\mathcal{S}} \circ L_{uV})(M - R_0).$$

Then there exist $\Delta M \in \mathcal{S}$ such that

$$L_{uV}(\Delta M + \Delta R) = 0 \tag{3.23}$$

and

$$\langle \delta M_0, \Delta M \rangle < 0. \tag{3.24}$$

Proof. See appendix. □

The last technical lemma provides the machinery for using Lemma's 3.4 and 3.5 to prove the existence of rank reducing perturbations with smaller Frobenius norm given that one of the two necessary conditions is not satisfied.

Lemma 3.6. *Let $(\zeta_0, \rho_0) \in \mathbb{R}^k \times \mathbb{R}^r$ satisfy* $\operatorname{rank}(M(\zeta_0, \rho_0)) = n - 1$ *and $W \subset \mathbb{R}^k \times \mathbb{R}^r$ be a neighborhood of (ζ_0, ρ_0) as in Lemma 3.3 where the n^{th} singular value is distinct. Let $T : W \to \mathbb{R}^2$ be the function*

$$T(\zeta, \rho) = \begin{bmatrix} f(\zeta, \rho) - f(\zeta_0, \rho_0) \\ \widetilde{H}(\zeta, \rho) \end{bmatrix}, \tag{3.25}$$

where f and $\widetilde{H}$ are defined in (3.8) *and Lemma 3.3, respectively. If i) $\frac{\partial}{\partial \zeta}\widetilde{H}(\zeta, \rho_0)|_{\zeta=\zeta_0}$ is surjective and ii) there exists $\rho_\Delta \in \mathbb{R}^r$ and $\zeta_\Delta \in \mathbb{R}^k$ such that $\zeta_0^\top \zeta_\Delta < 0$ and $\widetilde{H}'(\zeta_0, \rho_0)[\zeta_\Delta^\top, \rho_\Delta^\top]^\top = 0$, then T is Fréchet differentiable and $T'(\zeta_0, \rho_0)$ is surjective.*

Proof. See appendix. □

We can now prove necessary conditions as per the following proof.

Proof of Theorem 3.2.

Condition 1: For contradiction, assume that δM_0 is a minimum norm element in $\mathcal{T}$ (defined in (3.5)) with associated property matrix R_0, but condition 1 is not satisfied, i.e.,

$$\delta M_0 \neq (L^{\dagger}_{uV\mathcal{S}} \circ L_{uV})(M - R_0) \tag{3.26}$$

where $u \in \mathbb{C}^n$ and $V \in \mathbb{C}^{m\times(m-n+1)}$ are the n^{th} lsv and n^{th} through m^{th} rsv of $M - \delta M_0 - R_0$, respectively. By Lemma 3.4, there exists $\Delta M \in \mathcal{S}$ such that $L_{uV\mathcal{S}}(\Delta M) = 0$ and $\langle \delta M_0, \Delta M_0 \rangle < 0$. Let ζ_0 and ζ_Δ in $\mathbb{R}^k$ represent δM_0 and ΔM in the orthonormal basis $\{S_1, \cdots, S_k\}$, respectively. Let $\rho_0 \in \mathbb{R}^r$ represent R_0 in the orthonormal basis $\{P_1, \cdots, P_r\}$. Since the basis $\{S_1, \cdots, S_k\}$ is orthonormal, $\zeta_0^\top \zeta_\Delta = \langle \delta M_0, \Delta M \rangle < 0$. By the form of $\widetilde{H}'(\zeta_0, \rho_0)$ in Lemma 3.3,

$$\widetilde{H}'(\zeta_0, \rho_0) \begin{bmatrix} \zeta_\Delta \\ 0 \end{bmatrix} = -\operatorname{Re}(u^H \Delta M v),$$

where v is the n^{th} unsigned rsv of $M - \delta M_0 - R_0$. Since $L_{uV\mathcal{S}}(\Delta M) = 0$ and v is a linear combination of the columns of V, $u^H \Delta M v = 0$. This implies $\widetilde{H}'(\zeta_0, \rho_0)[\zeta_\Delta^\top, 0]^\top = 0$. In addition, since $L_{uV\mathcal{S}}$ is surjective, then $\frac{\partial}{\partial \zeta}\widetilde{H}(\zeta, \rho_0)|_{\zeta=\zeta_0}$ is also surjective. Hence by Lemma 3.6, $T(\zeta, \rho)$ given in (3.25) is Fréchet differentiable and $T'(\zeta_0, \rho_0)$ is surjective. Thus by the inverse function theorem [40], there exists an open set $W \subset \mathbb{R}^2$ containing zero such that for all $y \in W$, there exists $\zeta_y \in \mathbb{R}^k$ and $\rho_y \in \mathbb{R}^r$ such that $T(\zeta_y, \rho_y) = y$. Hence for all sufficiently small neighborhoods of 0 in $\mathbb{R}^2$, there exists $\delta > 0$, $\zeta_* \in \mathbb{R}^k$, and $\rho_* \in \mathbb{R}^r$ such that $T(\zeta_*, \rho_*) = [-\delta, 0]^\top$. This implies that $\delta M_* = \sum_{i=1}^k \zeta_{*i} S_i \in \mathcal{T}$, i.e., a rank reducing perturbation, with associated property matrix $R_* = \sum_{j=1}^r \rho_{*j} R_j$. Since $f(\zeta_*, \rho_*) - f(\zeta_0, \rho_0) = -\delta < 0$, $\|\delta M_*\|_F < \|\delta M_0\|_F$ contradicting that δM_0 is a local minimum norm element in $\mathcal{T}$.

Condition 2: For contradiction, assume that δM_0 is a minimum norm element in $\mathcal{T}$ with associated property matrix R_0 and condition 1 is satisfied, but condition 2 is not, i.e.,

$$\delta M_0 = (L^{\dagger}_{uV\mathcal{S}} \circ L_{uV})(M - R_0), \text{ and} \tag{3.27}$$

$$0 \neq (L^{\dagger}_{uV\mathcal{S}} \circ L_{uV\mathcal{P}})^{\dagger}(L^{\dagger}_{uV\mathcal{S}} \circ L_{uV})(M - R_0) \tag{3.28}$$

where $u \in \mathbb{C}^n$ and $V \in \mathbb{C}^{m\times(m-n+1)}$ are the n^{th} lsv and n^{th} through m^{th} rsv of $M - \delta M_0 - R_0$, respectively. By Lemma 3.5, there exists $\Delta M \in \mathcal{S}$ and $\Delta R \in \mathcal{P}$ such that $L_{uV}(\Delta M + \Delta R) = 0$ and $\langle \delta M_0, \Delta M_0\rangle < 0$. Let ζ_0 and ζ_Δ in $\mathbb{R}^k$ represent δM_0 and ΔM in the orthonormal basis $\{S_1, \cdots, S_k\}$, respectively. Let ρ_0 and ρ_Δ represent R_0 and ΔR in the orthonormal basis $\{P_1, \cdots, P_r\}$, respectively. Since the basis $\{S_1, \cdots, S_k\}$ is orthonormal, $\zeta_0^\top \zeta_\Delta = \langle \delta M_0, \Delta M\rangle < 0$. By the form of $\widetilde{H}'(\zeta_0, \rho_0)$ in Lemma 3.3,

$$\widetilde{H}'(\zeta_0, \rho_0)\begin{bmatrix}\zeta_\Delta \\ \rho_\Delta\end{bmatrix} = -\operatorname{Re}(u^H \Delta M v).$$

Since $L_{uV}(\Delta M + \Delta R) = 0$, $u^H(\Delta M + \Delta R)v = 0$ and this implies $\widetilde{H}'(\zeta_0, \rho_0)[\zeta_\Delta^\top, \rho_\Delta^\top]^\top = 0$. In addition, since $L_{uV\mathcal{S}}$ is surjective, then $\frac{\partial}{\partial\zeta}\widetilde{H}(\zeta, \rho_0)|_{\zeta=\zeta_0}$ is also surjective. Hence by Lemma 3.6 $T(\zeta, \rho)$ given in (3.25) is Fréchet differentiable and $T'(\zeta_0, \rho_0)$ is surjective. Using the same arguments as proving condition 1, this implies that there exists $\delta M_* \in \mathcal{T}$ smaller than δM_0 contradicting δM_0 is a local minimum element. □

The next section sets forth an algorithm which is proven to converge to a perturbation and property matrix pair $(\delta M_*, R_*)$ satisfying the necessary conditions of Theorem 3.2.

3.4 $\mathcal{P}$-Robustness Algorithm

We precede the proof of Algorithm 1 with a qualitative discussion on its scope and construction. Generically, a perturbation is unlikely to cause a two–dimensional drop in rank, i.e. $\operatorname{rank}[M - \delta M - R] < n - 1$. Hence, we focus on the most common problem structure where $\operatorname{rank}[M - \delta M - R] \geq n - 1$. This condition is formally captured by Assumption 3.4 in the next section. Modifications to this algorithm can be made to account for the more general case, but such is not included. Additional assumptions that guarantee convergence of the algorithm are introduced after the algorithm is delineated. It is important to note that the steps in the algorithm are chosen to compute norm reducing and rank reducing directions of search at each iteration. The algorithm proceeds along the direction of the vector sum with a step size α_k chosen to reduce a discrete step-dependent Lyapunov function.

Algorithm 1. $\mathcal{P}$–Robustness.

1. $k = 0$

2. Initialize $\delta M_0 \in \mathcal{S}$ and $R_0 \in \mathcal{P}$. Set $g_0 = 1$.

3. Repeat

4. Let u and V be the n^{th} lsv and n^{th} through m^{th} rsv of $M - \delta M_k - R_k$, respectively[2]. Define $[\sigma_n]_k \triangleq \sigma_n(M - \delta M_k - R_k)$.

5. **Norm reducing direction:**
   Let $\widetilde{\phi}_k = \min_{\delta M \in \mathcal{S}, \Delta R \in \mathcal{P}} \|L_{uV}(\delta M + \Delta R) - L_{uV\mathcal{S}}(\delta M_k)\|_F$ and

$$\begin{aligned} \widetilde{Z} = \{&(\delta M, \Delta R) : \\ &\|L_{uV}(\delta M + \Delta R - \delta M_k)\|_F = \widetilde{\phi}_k\} \end{aligned} \tag{3.29}$$

   Compute $\widetilde{\delta M_k}$ to be the first component of $\operatorname{argmin}_{(\delta M, \Delta R) \in \widetilde{Z}} \|\delta M\|_F$ and

$$\Delta \widetilde{R}_k = \operatorname*{argmin}_{\Delta R' \in \{\Delta R \in \mathcal{P} : (\widetilde{\delta M}_k, \Delta R) \in \widetilde{Z}\}} \|\Delta R'\|_F. \tag{3.30}$$

[2] We suppress the k-dependence of u and V to prevent overburdening the notation, i.e., the singular vectors change in each iteration.

If $L_{uV\mathcal{S}}$ is surjective then $\delta\widetilde{M}_k$ and $\Delta\widetilde{R}_k$ are given by

$$\Delta\widetilde{R}_k = (L_{uV\mathcal{S}}^{\dagger} \circ L_{uV\mathcal{P}})^{\dagger}(L_{uV\mathcal{S}}^{\dagger} \circ L_{uV\mathcal{S}})(\delta M_k) \tag{3.31}$$

$$\delta\widetilde{M}_k = (L_{uV\mathcal{S}}^{\dagger} \circ L_{uV})(\delta M_k - \Delta\widetilde{R}_k). \tag{3.32}$$

6. **Rank reducing direction:**

   Let $\overline{\phi}_k = \min_{\delta M \in \mathcal{S}, \Delta R \in \mathcal{P}} \|L_{uV}(\delta M + \Delta R) - L_{uV}(M - R_k - \delta M_k)\|_F$ and

$$\begin{aligned} \overline{Z} = \{&(\delta M, \Delta R) : \\ &\|L_{uV}(\delta M + \Delta R - (M - R_k - \delta M_k))\|_F = \overline{\phi}_k\} \end{aligned} \tag{3.33}$$

   Compute $\delta\overline{M}_k$ to be the first component of $\text{argmin}_{(\delta M, \Delta R) \in \overline{Z}} \|\delta M\|_F$ and

$$\Delta\overline{R}_k = \underset{\Delta R' \in \{\Delta R \in \mathcal{P} : (\delta\overline{M}_k, \Delta R) \in \widetilde{Z}\}}{\text{argmin}} \|\Delta R'\|_F. \tag{3.34}$$

   If $L_{uV\mathcal{S}}$ is surjective then $\delta\overline{M}_k$ and $\Delta\overline{R}_k$ are given by

$$\Delta\overline{R}_k = (L_{uV\mathcal{S}}^{\dagger} \circ L_{uV\mathcal{P}})^{\dagger}(L_{uV\mathcal{S}}^{\dagger} \circ L_{uV})(M - R_k - \delta M_k) \tag{3.35}$$

$$\delta\overline{M}_k = (L_{uV\mathcal{S}}^{\dagger} \circ L_{uV})(M - R_k - \Delta\overline{R}_k - \delta M_k) \tag{3.36}$$

7. **Lyapunov function reducing direction:**

   $\Delta R_k = \Delta\widetilde{R}_k + \Delta\overline{R}_k$ and $\delta\widehat{M}_k = \delta\widetilde{M}_k + \delta\overline{M}_k$

8. **Normalizing weights:**

   $g_k = \min\big(g_{k-1}, [\sigma_n]_k/(2\|\delta\overline{M}_k\|_F)\big)$, $b_k = 0$ if $\delta M_k = 0$, otherwise $b_k = \frac{1}{2}\|\delta M_k\|_F^{-1}$

9. **Choosing a step size:** Define

$$\begin{aligned} f_{ub}^{(k)}(\alpha) = \frac{-[\sigma_n]_k}{2}\alpha + a_k\alpha^2 - g_k b_k \|\delta M_k\|_F^2 \\ + g_k b_k \|(1-\alpha)\delta M_k + \alpha\delta\widetilde{M}_k\|_F^2 \end{aligned} \tag{3.37}$$

   where

$$\begin{aligned} a_k = \|[u_n]_k^H(\delta\widehat{M}_k - \delta M_k - \Delta R_k)(I - V_k V_k^H) \\ * (M - \delta M_k - R_k)^{\dagger}(\delta\widehat{M}_k - \delta M_k + \Delta R_k)\|_2. \end{aligned} \tag{3.38}$$

```
    Compute
        $\alpha_k = \operatorname*{argmin}_{\alpha\in[0,1]} f_{ub}^{(k)}(\alpha)$        (3.39)
10. Update estimates: $R_{k+1} = R_k + \alpha_k \Delta R_k$ and $\delta M_{k+1} = (1-\alpha_k)\delta M_k + \alpha_k \delta \widehat{M}_k$.
11. $k \to k+1$
12. UNTIL $\|\Delta R_k\|_F < \epsilon$, $\|\delta M_k - \delta\widetilde{M}_k\|_F < \epsilon$, $[\sigma_n]_k < \epsilon$
```

Several steps in Algorithm 1 require some elaboration. For the initialization in step 2, the initial guesses δM_0 and R_0 can be chosen as the best estimate for δM_* and R_*. For example, algorithms which compute upper and lower bounds (e.g. [25, 30]) can provide the initial estimates δM_0 and R_0. Alternatively, one can always choose $\delta M_0 = 0$ and $R_0 = 0$.

As mentioned, the $\mathcal{P}$–robustness algorithm is designed to reduce a Lyapunov energy function, which has the form $P_k = [\sigma_n]_k + g_k \|\delta M_k\|_F$, where $[\sigma_n]_k = \sigma_n(M - R_k - \delta M_k)$ and g_k is a nonzero adaptive weight computed in step 8. A direction for reducing the Lyapunov energy function is found by moving along the vector sum of the directions $(\delta\widetilde{M}_k, \Delta\widetilde{R}_k)$ (step 5) and $(\delta\overline{M}_k, \Delta\overline{R}_k)$ (step 6), which reduce $\|\delta M_{k+1}\|_F$ and $[\sigma_n]_{k+1}$, respectively. To illustrate how these directions affect the Lyapunov energy function, consider first the optimization problem in step 5. To find a pair $(\delta\widetilde{M}_k, \Delta\widetilde{R}_k)$ reducing $\|\delta M_{k+1}\|_F$, we search for the smallest pair that does not change the n^{th} singular value by approximating the function $\sigma_n(\cdot)$ with L_{uV}. Specifically, we require $(\delta\widetilde{M}_k, \Delta\widetilde{R}_k)$ to be the pair minimizing $\|\delta\widetilde{M}_k\|_F$ subject to

$$\begin{aligned} L_{uV}(M - R_k - \Delta\widetilde{R}_k - \delta\widetilde{M}_k) &= L_{uV}(M - R_k - \delta M_k) \\ &= \begin{bmatrix} [\sigma_n]_k & 0 & \cdots & 0 \end{bmatrix}. \end{aligned} \tag{3.40}$$

Subtracting $L_{uV}(M - R_k)$ from both sides of (3.40) and using the linearity of L_{uV}, we observe that

$$L_{uV}(\Delta\widetilde{R}_k + \delta\widetilde{M}_k) - L_{uV\mathcal{S}}(\delta M_k) = 0. \tag{3.41}$$

Hence the pairs $(\delta\widetilde{M_k}, \Delta\widetilde{R}_k)$ satisfying (3.40) constitute the set $\widetilde{Z}$ in (3.29) if $L_{uV\mathcal{S}}$ is surjective. In other words, if $L_{uV\mathcal{S}}$ is surjective then $\widetilde{\phi}_k = 0$, since for any $\Delta R \in \mathcal{P}$ setting

$$\delta M = (L_{uV\mathcal{S}}^{\dagger} \circ L_{uV})(\delta M_k - \Delta R) \tag{3.42}$$

results in

$$0 = \|L_{uV\mathcal{S}}(\delta M) - L_{uV}(\delta M_k - \Delta R)\|_F \geq \widetilde{\phi}_k \geq 0. \tag{3.43}$$

Moreover, δM defined by (3.42) is the matrix with the smallest Frobenius norm in $\mathcal{S}$ such that (3.43) is zero. So any pair $(\delta M, \Delta R) \in \widetilde{Z}$ for which $\|\delta M\|_F$ is minimized, will satisfy (3.42), i.e., for a yet unspecified $\Delta\widetilde{R}_k$,

$$\delta\widetilde{M_k} = (L_{uV\mathcal{S}}^{\dagger} \circ L_{uV})(\delta M_k - \Delta\widetilde{R}_k). \tag{3.44}$$

Choosing $\Delta\widetilde{R}_k$ to minimize $\|\delta\widetilde{M_k}\|_F$ (for pairs in $\widetilde{Z}$) is then equivalent to minimizing the norm of the right hand side of (3.44), i.e.,

$$\Delta\widetilde{R}_k = \operatorname*{argmin}_{\Delta R \in \mathcal{P}} \|\widetilde{\psi}(\Delta R)\|_F, \tag{3.45}$$

where

$$\psi(\Delta R) \triangleq (L_{uV\mathcal{S}}^{\dagger} \circ L_{uV\mathcal{P}})(\Delta R) - (L_{uV\mathcal{S}}^{\dagger} \circ L_{uV\mathcal{S}})(\delta M_k). \tag{3.46}$$

The matrix $\Delta\widetilde{R}_k$ with smallest Frobenius norm minimizing (3.45) is given by

$$\Delta\widetilde{R}_k = (L_{uV\mathcal{S}}^{\dagger} \circ L_{uV\mathcal{P}})^{\dagger} (L_{uV\mathcal{S}}^{\dagger} \circ L_{uV\mathcal{S}})(\delta M_k). \tag{3.47}$$

Since δM_k is known from the previous step, when $L_{uV\mathcal{S}}$ is surjective $\Delta\widetilde{R}_k$ can be computed first using (3.47) which is identical to (3.31) prior to computing $\delta\widetilde{M_k}$ using (3.44) which is identical to (3.32). This justifies the statements of step 5.

Step 6 computes a direction $(\delta\overline{M}_k, \Delta\overline{R}_k)$ for reducing $[\sigma_n]_{k+1}$. Namely, the objective to choose $(\delta\overline{M}_k, \Delta\overline{R}_k)$ minimizing $\|\delta\overline{M}_k\|_F$ subject to

$$\begin{aligned} L_{uV}(\delta\overline{M}_k + \Delta\overline{R}_k) &= L_{uV}(M - R_k - \delta M_k). \\ &= \begin{bmatrix} [\sigma_n]_k & 0 & \cdots & 0 \end{bmatrix}. \end{aligned} \tag{3.48}$$

As in step 5, the linear operator L_{uV} approximates the smallest singular value function $\sigma_n(\cdot)$. Hence (3.48) is an approximation of the constraint $\sigma_n(M - R_k - \Delta\overline{R}_k - \delta M_k - \delta\overline{M}_k) = 0$. Using the same arguments used above for step 5, pairs $(\delta\overline{M}_k, \Delta\overline{R}_k)$ satisfying (3.48) are in $\overline{Z}$ if $L_{uV\mathcal{S}}$ is surjective. In other words, if $L_{uV\mathcal{S}}$ is surjective, then $\overline{\phi}_k = 0$ and the pair $(\delta\overline{M}_k, \Delta\overline{R}_k) \in \overline{Z}$ minimizing $\|\delta\overline{M}_k\|_F$ satisfies

$$\delta\overline{M}_k = (L^{\dagger}_{uV\mathcal{S}} \circ L_{uV})(M - R_k - \Delta\overline{R}_k - \delta M_k), \tag{3.49}$$

where $\Delta\overline{R}_k$ is chosen to be the smallest norm matrix in $\mathcal{P}$ minimizing the norm of the right side of (3.49), i.e.,

$$\Delta\overline{R}_k = \operatorname*{argmin}_{\Delta R \in \mathcal{P}} \|\overline{\psi}(\Delta R)\|_F, \tag{3.50}$$

where

$$\begin{aligned} \overline{\psi}(\Delta R) = {} & (L^{\dagger}_{uV\mathcal{S}} \circ L_{uV\mathcal{P}})(\Delta R) \\ & - (L^{\dagger}_{uV\mathcal{S}} \circ L_{uV})(M - R_k - \delta M_k). \end{aligned} \tag{3.51}$$

The matrix $\Delta\overline{R}_k$ with smallest Frobenius norm minimizing (3.50) is given by

$$\Delta\overline{R}_k = (L^{\dagger}_{uV\mathcal{S}} \circ L_{uV\mathcal{P}})^{\dagger} (L^{\dagger}_{uV\mathcal{S}} \circ L_{uV})(M - R_k - \delta M_k). \tag{3.52}$$

This completes the justification of step 6.

What remains is to specify the step size α_k. To choose α_k, we would like to minimize the Lyapunov function $P_{k+1} = [\sigma_n]_{k+1} + g_{k+1}\|\delta M_{k+1}\|_F$ in the direction of ΔR_k and $\delta\widehat{M_k}$ in step 7. Due to differentiability issues of the singular value function $\sigma_n(\cdot)$, we will instead minimize a quadratic function $f^{(k)}_{ub}(\alpha)$ in (3.37) which upper bounds (verified in the proof of Theorem 3.10) the decrease in the Lyapunov function, i.e., $P_{k+1}(\alpha) - P_k \leq f^{(k)}_{ub}(\alpha)$. We will show that choosing α_k to be the minimum of this upper bound $f^{(k)}_{ub}$ will imply that the sequence of Lyapunov functions $\{P_k\}$ converges. Unlike the usual definition of Lyapunov energy functions, we will not guarantee that $\{P_k\}$ converges to zero, but rather a positive constant, $\{P_k\} \to d = g_*\|\delta M_*\|_F$. This will be sufficient for guaranteeing the necessary conditions are met at the terminating values for δM_* and R_* if Assumption 3.1 and two additional assumptions to be described in Section 3.5. The next subsection addresses details for the implementation of Algorithm 1 in software.

3.4.1 Algorithm 1 Implementation

Implementing a few steps of Algorithm 1 require some explanation. To implement steps 5 and 6 of Algorithm 1, the pseudoinverse $L_{uV\mathcal{S}}$ is computed via Kronecker products and the vec operator [41,42]. Applying the vec operator to $L_{uV\mathcal{S}}$ we obtain

$$\begin{aligned}\operatorname{vec}(L_{uV\mathcal{S}}(\delta M)) &= (V^\top \otimes u^H)\operatorname{vec}(\delta M)\\ &= (V^\top \otimes u^H)B_{\mathcal{S}}\zeta,\end{aligned}$$

where ζ respresents δM in the orthonormal basis $\{S_1, \cdots, S_k\}$ and

$$B_{\mathcal{S}} \triangleq [\operatorname{vec}(S_1), \cdots, \operatorname{vec}(S_k)].$$

Taking the real and imaginary components of $\operatorname{vec}(L_{uV\mathcal{S}}(\delta M))$ we obtain

$$\begin{bmatrix}(\operatorname{Re}[L_{uV\mathcal{S}}(\delta M)])^\top\\ (\operatorname{Im}[L_{uV\mathcal{S}}(\delta M)])^\top\end{bmatrix} = \begin{bmatrix}\operatorname{Re}[(V^\top \otimes u^H)B_{\mathcal{S}}]\\ \operatorname{Im}[(V^\top \otimes u^H)B_{\mathcal{S}}]\end{bmatrix}\zeta.$$

Let $N_{\mathcal{S}} \in \mathbb{C}^{2(m-n+1)\times k}$ and $N_{\mathcal{P}} \in \mathbb{C}^{2(m-n+1)\times r}$ be given by

$$\begin{aligned}N_{\mathcal{S}} &= \begin{bmatrix}\operatorname{Re}[(V_k^\top \otimes [u_n]_k^H)B_{\mathcal{S}}]\\ \operatorname{Im}[(V_k^\top \otimes [u_n]_k^H)B_{\mathcal{S}}]\end{bmatrix},\\ N_{\mathcal{P}} &= \begin{bmatrix}\operatorname{Re}[(V_k^\top \otimes [u_n]_k^H)B_{\mathcal{P}}]\\ \operatorname{Im}[(V_k^\top \otimes [u_n]_k^H)B_{\mathcal{P}}]\end{bmatrix},\end{aligned}$$

where $B_{\mathcal{P}} \triangleq [\operatorname{vec}(P_1), \cdots, \operatorname{vec}(P_r)]$. With this notation, $\Delta\widetilde{R}_k$ of (3.31) and $\delta\widetilde{M}_k$ of (3.32) satisfy

$$\begin{aligned}\operatorname{vec}(\Delta\widetilde{R}_k) &= B_{\mathcal{P}}(N_{\mathcal{S}}^\dagger N_{\mathcal{P}})^\dagger N_{\mathcal{S}}^\dagger N_{\mathcal{S}}\zeta_k\\ \operatorname{vec}(\delta\widetilde{M}_k) &= B_{\mathcal{S}}N_{\mathcal{S}}^\dagger\begin{bmatrix}\operatorname{Re}[(V_k^\top \otimes [u_n]_k^H)\operatorname{vec}(\delta M_k - \Delta\widetilde{R}_k)]\\ \operatorname{Im}[(V_k^\top \otimes [u_n]_k^H)\operatorname{vec}(\delta M_k - \Delta\widetilde{R}_k)]\end{bmatrix},\end{aligned}$$

where ζ_k represents δM_k in $\{S_1, \cdots, S_k\}$. Similarly, $\Delta \overline{R}_k$ in (3.35) and $\delta \overline{M}_k$ in (3.36) satisfy

$$
\begin{aligned}
\mathrm{vec}(\Delta \overline{R}_k) &= B_{\mathcal{P}}(N_{\mathcal{S}}^{\dagger} N_{\mathcal{P}})^{\dagger} N_{\mathcal{S}}^{\dagger} * \\
&\qquad \begin{bmatrix} \mathrm{Re}[(V_k^{\top} \otimes [u_n]_k^H) \, \mathrm{vec}(M - R_k - \delta M_k)] \\ \mathrm{Im}[(V_k^{\top} \otimes [u_n]_k^H) \, \mathrm{vec}(M - R_k - \delta M_k)] \end{bmatrix} \\
\mathrm{vec}(\delta \widetilde{M}_k) &= B_{\mathcal{S}} N_{\mathcal{S}}^{\dagger} * \\
&\qquad \begin{bmatrix} \mathrm{Re}[(V_k^{\top} \otimes [u_n]_k^H) \, \mathrm{vec}(M - R_k - \Delta \overline{R}_k - \delta M_k)] \\ \mathrm{Im}[(V_k^{\top} \otimes [u_n]_k^H) \, \mathrm{vec}(M - R_k - \Delta \overline{R}_k - \delta M_k)] \end{bmatrix}
\end{aligned}
$$

Now we consider step 9 that requires the minimization of the function $f_{ub}^{(k)}$ in (3.37) with respect to the step size α_k. One such method for this minimization is a one dimensional constrained line search for $\alpha_k \in [0, 1]$. Since a decrease in the Lyapunov energy function P_{k+1} is guaranteed for α sufficiently small, an appropriate initial guess for α_k is 0. Alternatively, one can analytically solve for the minimizer since $f_{ub}^{(k)}$ is a quadratic function of α. Namely

$$
f_{ub}^{(k)}(\alpha) = \alpha(c_1^{(k)} + c_2^{(k)} \alpha),
$$

where $c_1^{(k)}$ and $c_2^{(k)}$ are given by

$$
\begin{aligned}
c_1^{(k)} &= -\left(\frac{[\sigma_n]_k}{2} + 2 \, \mathrm{Re}[\langle \delta M_k, \delta \widetilde{M}_k - \delta M_k \rangle] \right) \\
c_2^{(k)} &= a_k + \| \delta \widetilde{M}_k - \delta M_k \|_F^2 .
\end{aligned}
$$

The minimizer of $f_{ub}^{(k)}$ in the interval $[0, 1]$ is given by

$$
\alpha_k = \min \left\{ \frac{c_1^{(k)}}{2 c_2^{(k)}}, 1 \right\},
$$

if $c_2^{(k)} \neq 0$ and 0 otherwise. Either a line search or the analytical solution to minimizing $f_{ub}^{(k)}$ over the interval $[0, 1]$ can be used for step 9.

3.5 Convergence of Algorithm 1

Convergence of the $\mathcal{P}$–robustness problem requires a structural condition on the property space $\mathcal{P}$:

Assumption 3.2. *Each nonzero property matrix $R \in \mathcal{P}$ is full rank, i.e.,* $\operatorname{rank} R = n$*.*

If there exists a nonzero $R \in \mathcal{P}$ which is not full rank, $\sigma_n(M - \delta M - \eta R)$ may be finite (and possibly optimal) as $\eta \to \infty$. Convergence to a finite property matrix is guaranteed to exist if for all nonzero $R \in \mathcal{P}$, $\operatorname{rank}(R) = n$, i.e., R is full row rank; hence, the infimum in (3.4) and (3.6) can be replaced with the minimum since the associated optimal property matrix is bounded.

Two additional assumptions aid in proving convergence of Algorithm 1.

Assumption 3.3. *The sequence $\{g_k\}$ computed by Algorithm 1 is bounded away from zero.*

Assumption 3.4. *The sequence $\{[\sigma_{n-1}]_k\}$ computed by Algorithm 1 is bounded away from zero.*

The sequence $\{g_k\}$ in Assumption 3.3 essentially measures the surjectivity of $L_{uV\mathcal{S}}$. When $\{g_k\}$ is bounded away from zero, the algorithm converges to a minimizer at which $L_{uV\mathcal{S}}$ is surjective. Assumption 3.4 requires that the $(n-1)^{th}$ singular value, $[\sigma_{n-1}]_k$, is nonzero. In addition, we will assume that $L_{uV\mathcal{S}}$ is surjective for each unit vector $u \in \mathbb{C}^n$ and each matrix $V \in \mathbb{C}^{m \times (m-n+1)}$ with columns which are orthogonal unit vectors which can appear as singular vectors of $M - \delta M_k - R_k$ as introduced in Assumption 3.1. Note, this set of u and V satisfy $u^H u = u^\dagger u = 1$ and $V^H V = V^\dagger V = I$. To prove convergence of Algorithm 1, we need to establish a few necessary lemmas. The first lemma bounds $[\sigma_n]_{k+1}(\alpha)$ as a function of α.

Lemma 3.7. *Let u and V contain lsv and rsv of $M - \delta M_k - R_k$, respectively, as in Algorithm 1. If $L_{uV\mathcal{S}}$ is surjective, then for all $\alpha \in (0, 1)$,*

$$\sigma_n\left(M - \delta M_k - R_k - \alpha(\widehat{\delta M_k} - \delta M_k + \Delta R_k)\right)$$

is bounded from above by

$$(1-\alpha)[\sigma_n]_k + \alpha^2 a_k \tag{3.53}$$

where $[\sigma_n]_k \triangleq \sigma_n(M - \delta M_k - R_k)$ *and* a_k *is given in* (3.38).

Proof. See appendix. □

The next lemma constructs an upper bound on the norm $\|\delta M_{k+1}\|_F$ as a function of α. To state the upper bound, we require the following linear orthogonal projection operators from the proofs of Lemmas 3.4 and 3.5:

$$Q_1 \triangleq (L^{\dagger}_{uV\mathcal{S}} \circ L_{uV\mathcal{S}}) \tag{3.54}$$

$$Q_2 \triangleq I - (L^{\dagger}_{uV\mathcal{S}} \circ L_{uV\mathcal{P}})(L^{\dagger}_{uV\mathcal{S}} \circ L_{uV\mathcal{P}})^{\dagger}. \tag{3.55}$$

Lemma 3.8. *Let* u *and* V *contain lsv and rsv of* $M - \delta M_k - R_k$*, respectively, as in Algorithm 1. If* $L_{uV\mathcal{S}}$ *is surjective, then for all* $\alpha \in [0,1]$

$$\begin{aligned}\|(1-\alpha)\delta M_k + \alpha\delta\widehat{M_k}\|_F &\leq \|\delta M_k\|_F + \alpha\|\delta\overline{M}_k\|_F \\ &+ b_k\left(\|(1-\alpha)\delta M_k + \alpha(Q_2\circ Q_1)(\delta M_k)\|_F^2 - \|\delta M_k\|_F^2\right),\end{aligned} \tag{3.56}$$

where b_k *and* $\delta\overline{M}_k$ *are given in Algorithm 1 and* Q_1 *and* Q_2 *are given in* (3.54) *and* (3.55)*, respectively.*

Proof. See appendix. □

Proving convergence of Algorithm 1 will be achieved by appealing to a Lyapunov function $P_k = [\sigma_n]_k + g_k\|\delta M_k\|_F$. Given Assumptions 3.2 and 3.3, we will show i) $P_{k+1} - P_k \leq f_{ub}^{(k)}(\alpha_k) \leq 0$ and ii) that $f_{ub}^{(k)}(\alpha_k) < 0$ if the necessary conditions of δM_k and δR_k in Theorem 3.2 are not satisfied. Since P_k is nonnegative for each k, proving that $\{P_k\}$ is nonincreasing implies it is a bounded monotone function so the sequence converges, i.e., $P_{k+1} - P_k \to 0$. Hence this will prove that $f_{ub}^{(k)}(\alpha_k) \to 0$ which implies convergence to a pair δM_* and R_* satisfying the necessary conditions in Theorem 3.2. The next lemma is the key step in relating $f_{ub}^{(k)}$ and the necessary conditions in Theorem 3.2.

Lemma 3.9. *Let $R \in \mathcal{P}$ and $\delta M \in \mathcal{S}$ satisfy* $\operatorname{rank}[M - R - \delta M] < n$*, i.e., a candidate solution. R and δM satisfy the necessary conditions in Theorem 3.2 if and only if*

$$(Q_2 \circ Q_1)(\delta M) = \delta M,$$

where Q_1 and Q_2 are given in (3.54) *and* (3.55)*, respectively, with u the n^{th} lsv of $M - R - \delta M$ and V has columns containing the n^{th} through m^{th} rsv of $M - R - \delta M$.*

Proof. See appendix. □

The next Theorem proves that if Algorithm 1 is carried out to infinite precision, then the algorithm converges to a necessary condition for an optimal solution R_* and δM_*. The stopping conditions in Algorithm 1 guarantee that the algorithm terminates. The parameter ϵ determines how far the terminal points δM_k and R_k are from satisfying the necessary conditions.

Theorem 3.10. *If Assumptions 3.2-3.4 hold, then the sequence $\{P_k\}$ computed by Algorithm 1 converges, where*

$$P_k \triangleq [\sigma_n]_k + g_k \|\delta M_k\|_F. \tag{3.57}$$

Further, the sequences $\{\delta M_k\}$ and $\{R_k\}$ have limit points δM_* and R_* satisfying the necessary conditions of Theorem 3.2.

Proof. First we show that $P_{k+1} - P_k \leq f_{ub}^{(k)}(\alpha_k)$. Because g_k is nonincreasing, Lemma 3.7 and 3.8 imply that

$$\begin{aligned}
&P_{k+1} - P_k \\
&= [\sigma_n]_{k+1} + g_{k+1}\|\delta M_{k+1}\|_F - ([\sigma_n]_k + g_k\|\delta M_k\|_F) \\
&\leq -\alpha_k[\sigma_n]_k + a_k\alpha_k^2 + \alpha_k g_k\|\delta\overline{M}_k\|_F - g_k b_k\|\delta M_k\|_F^2 \\
&\quad + g_k b_k\|(1-\alpha_k)\delta M_k + \alpha_k(Q_2 \circ Q_1)(\delta M_k)\|_F^2
\end{aligned}$$

Since $g_k \leq [\sigma_n]_k/(2\|\delta\overline{M}_k\|_F)$,

$$\begin{aligned}
&P_{k+1} - P_k \\
&\leq \frac{-[\sigma_n]_k}{2}\alpha_k + a_k\alpha_k^2 - g_k b_k \|\delta M_k\|_F^2 \\
&\quad + g_k b_k \|(1-\alpha_k)\delta M_k + \alpha_k(Q_2 \circ Q_1)(\delta M_k)\|_F^2 \\
&\triangleq f_{ub}^{(k)}(\alpha_k).
\end{aligned}$$

Note that $f_{ub}^{(k)}(\alpha)$ is a quadratic function of α and $f_{ub}^{(k)}(0) = 0$. So there exists constants $c_1^{(k)}$ and $c_2^{(k)}$ such that

$$f_{ub}^{(k)}(\alpha) = c_1^{(k)}\alpha + c_2^{(k)}\alpha^2.$$

Careful inspection of $f_{ub}^{(k)}(\alpha)$ shows that $c_2^{(k)} \geq 0$, i.e., $f_{ub}^{(k)}(\alpha)$ admits a global minimum. Since g_k, b_k, and $[\sigma_n]_k$ are all nonnegative, $c_1^{(k)} \leq 0$ if the coefficient of the linear term in the quadratic

$$\|(1-\alpha_k)\delta M_k + \alpha_k(Q_2 \circ Q_1)(\delta M_k)\|_F^2 - \|\delta M_k\|_F^2$$

is nonpositive. This is clearly the case since $\|(Q_2 \circ Q_1)(\delta M_k)\|_F \leq \|\delta M_k\|_F$. Further, $c_1^{(k)} = 0$ if and only if $[\sigma_n]_k = 0$ and $(Q_2 \circ Q_1)(\delta M_k) = \delta M_k$ since $g_k > 0$ by Assumption 3.3. Equivalently, Lemma 3.9 implies that $c_1^{(k)} = 0$ if and only if the necessary conditions are satisfied. Since α_k is chosen to minimize $f_{ub}^{(k)}$ over the interval $[0, 1]$, $f_{ub}^{(k)}(\alpha_k) < 0$ so long as the necessary conditions are not satisfied.

Thus $\{P_k\}$ is nonnegative and decreasing since $P_{k+1} - P_k \leq f_{ub}^{(k)}(\alpha_k) \leq 0$. By the monotone convergence theorem, $\{P_k\}$ converges, i.e., $P_{k+1} - P_k \to 0$. To prove that we converge to a necessary condition, we will prove that the sequence $\{c_1^{(k)}\}$ converges to zero. Based on Lemma 3.9, this implies that a necessary condition is satisfied.

Since $P_{k+1} - P_k \leq f_{ub}^{(k)}(\alpha_k) \leq 0$ and $P_{k+1} - P_k \to 0$, the sequence $\{f_{ub}^{(k)}(\alpha_k)\} \to 0$. As long as $\{c_2^{(k)}\}$ is bounded, this implies that $\{c_1^{(k)}\} \to 0$ as desired. The sequence $\{c_2^{(k)}\}$ is unbounded only if the quadratic coefficient of the function

$$\begin{aligned}
&a_k\alpha_k^2 - g_k b_k \|\delta M_k\|_F^2 \\
&\quad + g_k b_k \|(1-\alpha_k)\delta M_k + \alpha_k(Q_2 \circ Q_1)(\delta M_k)\|_F^2
\end{aligned} \tag{3.58}$$

goes unbounded. The quadratic term coefficient in (3.58) is given by $a_k + g_k b_k \|(I - Q_2 \circ Q_1)\delta M_k\|^2$ and by construction $b_k \|(I - Q_2 \circ Q_1)\delta M_k\|_F^2 \leq b_k \|\delta M_k\|_F^2 \leq \|\delta M_k\|_F$. Since $\{P_k\}$ converges and $\{g_k\} > 0$, $\|\delta M_k\|_F$ is bounded. By (3.38), $a_k \leq \|\widehat{\delta M_k} - \delta M_k - \Delta R_k\|_F^2 / [\sigma_{n-1}]_k$, which by Assumption 3.4 is bounded if ΔR_k is bounded, or equivalently if R_k is bounded.

Assume for contradiction that $\{R_k\}$ is unbounded. Since $\{P_k\}$ converges, $[\sigma_n]_k$ is bounded. Since $\{g_k\} > 0$, $\{\delta M_k\}$ is bounded as well. Let R_{min} be the norm one property matrix minimizing σ_n, i.e., $R_{min} = \operatorname{argmin}_{R \in \mathcal{P}, \|R\|_F = 1} \sigma_n(R)$. By Assumption 3.2, $\sigma_n(R_{min}) > 0$ and for any $R \in \mathcal{P}$, $\sigma_n(R) \geq \|R\|_F \sigma_n(R_{min})$. Hence, letting u_k be the n^{th} lsv of $M - R_k - \delta M_k$, for sufficiently large k

$$\begin{aligned}
[\sigma_n]_k &= \|u_k^H (M - R_k - \delta M_k)\|_2 \\
&\geq \|u_k^H R_k\|_2 - \|u_k^H (M - \delta M_k)\|_2 \\
&\geq \sigma_n(R_k) - \|u_k^H (M - \delta M_k)\|_2 \\
&\geq \|R_k\|_F \sigma_n(R_{min}) - \|u_k^H (M - \delta M_k)\|_2 .
\end{aligned}$$

Hence, if $\|R_k\|_F \to \infty$, then $[\sigma_n]_k \to \infty$ contradicting that $\{P_k\}$ converges. Hence $\{\Delta R_k\}$ is bounded and thus $\{a_k\}$ and $\{c_2^{(k)}\}$ are bounded. This implies that as $k \to \infty$, $(Q_2 \circ Q_1)(\delta M_k) \to \delta M_k$, i.e., the two necessary conditions in Theorem 3.2 are satisfied as $k \to \infty$. Finally, since $\{g_k\} > 0$ and $\{P_k\}$ converges, the sequence of perturbations $\{\delta M_k\}$ has a bounded accumulation point δM_*. Since δM_* satisfies the two necessary conditions, the sequence $\{\Delta R_k\}$ has an accumulation point $\Delta R_* = 0$, i.e., $R_k \to R_*$, completing the proof. □

3.6 Numerical Examples

3.6.1 Example 1

Consider the third example in [25] (also appears in [32] and [29]), which in the $\mathcal{P}$–robustness framework has system matrix M, structured perturbations δM, and property matrices R given by

$$\begin{aligned} M &= \begin{bmatrix} A & B \end{bmatrix} \\ \delta M &= \begin{bmatrix} \delta A & \delta B \end{bmatrix} \in \mathbb{R}^{3\times 4} \\ R &= \lambda \begin{bmatrix} I & 0 \end{bmatrix} \in \mathbb{C}^{3\times 4}, \end{aligned}$$

where

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 0.1 & 3 & 5 \\ 0 & -1 & -1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 0.1 \\ 0 \end{bmatrix}.$$

With initial guesses $\delta M_0 = 0$ and $R_0 = [jI, 0]$, and $\epsilon = 10^{-10}$, Algorithm 1 terminates in 9 iterations. The $\mathcal{P}$–robustness of M with respect to parameter variations in $\mathcal{S}$ is computed to be $r(M; \mathcal{S}, \mathcal{P}) = 0.057737$. The minimizing property and perturbation matrices are $R_* = (0.9824 + 0.9731j)[I, 0]$ and $\delta M_* = [\delta A_*, \delta B_*]$, respectively, where

$$\begin{aligned} \delta A_* &= 10^{-4} \begin{bmatrix} -5.8878 & -0.49659 & 0.29287 \\ 168.48 & 14.210 & -8.3803 \\ 167.31 & 14.111 & -8.3221 \end{bmatrix} \\ \delta B_*^\top &= 10^{-3} \begin{bmatrix} 1.1427 & 15.754 & 49.685 \end{bmatrix}. \end{aligned}$$

Upon termination, $\sigma_n(M - \delta M_* - R_*) = 2.8944 \times 10^{-16}$ which is approximately zero. These results are consistent with [25] and [29]. As noted in [29], we cannot compare the results of this example to [32] due to the different norm used therein (largest singular value of δM versus the Frobenius norm).

3.6.2 Example 2

Consider the example in [30], which in the $\mathcal{P}$–robustness framework has a fixed system matrix M, structured perturbations δM, and property matrices R are given by

$$M = \begin{bmatrix} A_0^\top & 0 & C_0^\top \\ 0 & A_1^\top & C_1^\top \end{bmatrix}$$
$$\delta M = \begin{bmatrix} \delta A_0^\top & 0 & \delta C_0^\top \\ 0 & \delta A_1^\top & \delta C_1^\top \end{bmatrix} \in \mathbb{R}^{4\times 5}$$
$$R = \lambda \begin{bmatrix} I & 0 \end{bmatrix} \in \mathbb{C}^{4\times 5},$$

where

$$A_0 = \begin{bmatrix} -1 & 2 \\ 0 & -2 \end{bmatrix}, A_1 = \begin{bmatrix} -3 & 0.1 \\ 5 & -1 \end{bmatrix}$$
$$C_0 = \begin{bmatrix} 1 & 0 \end{bmatrix}, \quad C_1 = \begin{bmatrix} 1 & 1 \end{bmatrix}.$$

The perturbation space $\mathcal{S}$ is real and does not allow perturbations of the off-diagonal entries of M. In [30], the distance to the nearest SMS SLTI system is computed to satisfy

$$0.0506 \leq r(M; \mathcal{S}, \mathcal{P}) \leq 0.4570.$$

Setting the terminating condition for $\epsilon = 10^{-15}$ and initial guesses $\delta M_0 = 0$ and $R_0 = 0$, the Algorithm 1 terminated in 13 iterations. The distance $r(M; \mathcal{S}, \mathcal{P})$ is computed to be $r(M; \mathcal{S}, \mathcal{P}) = 0.071821$ where $R_* = -0.9065 \begin{bmatrix} I & 0 \end{bmatrix}$,

$$\delta M_* = 10^{-3} \begin{bmatrix} -21.5 & -39.3 & 0 & 0 & 0 \\ -0.8 & -1.4 & 0 & 0 & 0 \\ 0 & 0 & 1.2 & 0.5 & 0 \\ 0 & 0 & 51.8 & 21.7 & 0 \end{bmatrix}, \tag{3.59}$$

and $\sigma_n(M - \delta M_* - R_*) = 5.256 \times 10^{-16} \approx 0$. Note that $\phi_{\mathbb{R}}(\Sigma) = 0.071821$ is between 0.0506 and 0.4570.

3.7 Conclusion

In this work, the $\mathcal{P}$–robustness framework developed in [38] is used to solve a family of robustness problems. Specifically, the Frobenius norm metric is used to measure the $\mathcal{P}$–robustness of M with respect to perturbations in $\mathcal{S}$. Necessary conditions for a minimal rank reducing perturbation are proven in Theorem 3.2. The necessary conditions motivate Algorithm 1 for computing both the metric $r(M\ \mathcal{S}, \mathcal{P})$ and the minimizing property matrix R_* and perturbation matrix δM_*.

In future work, we will modify Algorithm 1 to solve $\mathcal{P}$–robustness problems with singular property matrices, i.e., $\text{rank}(R) < n$. This modification will address the case where the norm of the optimal property matrix R_* is unbounded. In addition, we expect that Algorithm 1 can be modified to compute the $\mathcal{P}$–robustness of M using the spectral norm metric, i.e., minimizing $\sigma_1(\delta M_*)$. Although the Frobenius norm may be a more accurate measure of robustness, extending to the spectral norm metric unifies the robustness property literature.

3.8 Chapter 3 Appendix

3.8.1 Surjectivity

This section explores conditions for surjectivity of maps L_{uV} and $L_{uV\mathcal{S}}$. The first result is that $L_{uV} : \mathbb{C}^{n\times m} \to \mathbb{C}^{1\times(m-n+1)}$ is surjective if u is a unit vector and V has mutually orthonormal columns.

Proposition 3.11. *If $u \in \mathbb{C}^m$ is a unit vector and $V \in \mathbb{C}^{m\times(m-n+1)}$ V has mutually orthonormal columns then $L_{uV} : \mathbb{C}^{n\times m} \to \mathbb{C}^{1\times(m-n+1)}$ is surjective.*

Proof. Represent the map L_{uV} as the $(V^\top \otimes u^H)$ which maps $\text{vec}(M)$ into $\mathbb{C}^{m-n+1}$. As such, $\text{rank}(V^\top \otimes u^H) = \text{rank}(V) * \text{rank}(u) = m-n+1$ (see [43, Corollary 13.11] for the rank of Kronecker product). Since the matrix representing L_{uV} has rank $m-n+1$, the range $L_{uV}(\mathbb{C}^{n\times m})$ has dimension $m-n+1$ (with respect to complex coefficients), i.e., $L_{uV}(\mathbb{C}^{n\times m}) = \mathbb{C}^{1\times(m-n+1)}$. □

Since $\mathcal{S}$ is a subspace of $\mathbb{C}^{n\times m}$, u being a unit vector and V having mutually orthonormal columns is insufficient for surjectivity of $L_{uV\mathcal{S}}$. The surjectivity of $L_{uV\mathcal{S}}$ is now investigated in the following two results. To avoid confusion when discussing the dimension of a complex subspace viewed as a subspace over the field of real numbers, we define $\dim_{\mathbb{C}}$ and $\dim_{\mathbb{R}}$ to denote the dimension of the subspace over the field of complex and real numbers, respectively. The following example illustrates the distinction.

Example 3.3. *Consider the subspace $\mathcal{E} \in \mathbb{C}$ given by $\mathcal{E}_1 = \{\alpha(1+i) : \alpha \in \mathbb{R}\}$. $\mathcal{E}_1$ has exactly one basis vector when viewed as a subspace over the field of real numbers, hence* $\dim_{\mathbb{R}}(\mathcal{E}_1) = 1$. *$\mathcal{E}_1$ is not a subspace over the field of complex numbers. For comparison,* $\dim_{\mathbb{C}}(\mathbb{C}) = 1$ *and* $\dim_{\mathbb{R}}(\mathbb{C}) = 2$ *since $\mathbb{C}$ can be expressed as $\mathbb{C} = \{\alpha_1 + \alpha_2 j : \alpha_i \in \mathbb{R}\}$.*

Proposition 3.12. *Let $u \in \mathbb{C}^n$, and $V \in \mathbb{C}^{m\times(m-n+1)}$. Let $\{S_1, S_2, \ldots, S_k\}$ be a basis for $\mathcal{S}$. $L_{uV\mathcal{S}}$ is surjective if and only if*

$$\operatorname{rank}\left(\begin{bmatrix}\mathrm{Re}[(V^\top \otimes u^H)B_{\mathcal{S}}] \\ \mathrm{Im}[(V^\top \otimes u^H)B_{\mathcal{S}}]\end{bmatrix}\right) = 2(m-n+1) \tag{3.60}$$

where $B_{\mathcal{S}} = [\mathrm{vec}(S_1), \mathrm{vec}(S_2), \ldots, \mathrm{vec}(S_k)]$.

Proof. Using the $\mathrm{vec}(\cdot)$ operator, $L_{uV\mathcal{S}}$ is surjective if and only if $\dim(L_{uV\mathcal{S}}(\mathcal{S})) = m-n+1$. Let $\zeta_0 \in \mathbb{R}^k$ satisfy $\mathrm{vec}(\delta M) = B_{\mathcal{S}}\zeta_0$. Then,

$$\begin{aligned}\mathrm{vec}(L_{uV\mathcal{S}}(\delta M)) &= (V^\top \otimes u^H)\,\mathrm{vec}(\delta M) \\ &= (V^\top \otimes u^H)B_{\mathcal{S}}\zeta_0. \end{aligned} \tag{3.61}$$

Let $y \in \mathbb{C}^{m-n+1}$ be an arbitrary vector. A matrix $\delta M \in \mathcal{S}$ satisfies $L_{uV\mathcal{S}}(\delta M) = y^\top$ if and only if

$$\begin{bmatrix}\mathrm{Re}(y) \\ \mathrm{Im}(y)\end{bmatrix} = \begin{bmatrix}\mathrm{Re}(\mathrm{vec}(L_{uV\mathcal{S}}(\delta M))) \\ \mathrm{Im}(\mathrm{vec}(L_{uV\mathcal{S}}(\delta M)))\end{bmatrix}.$$

Using (3.61), $L_{uV\mathcal{S}}(\delta M) = y^\top$ where δM has a real basis vector ζ_0 if and only if

$$\begin{bmatrix}\mathrm{Re}(y) \\ \mathrm{Im}(y)\end{bmatrix} = \begin{bmatrix}\mathrm{Re}((V^\top \otimes u^H)B_{\mathcal{S}}) \\ \mathrm{Im}((V^\top \otimes u^H)B_{\mathcal{S}})\end{bmatrix}\zeta_0. \tag{3.62}$$

$L_{uV\mathcal{S}}$ is surjective if and only if for each $y \in \mathbb{C}^{m-n+1}$, there exists $\zeta_0 \in \mathbb{R}^k$ such that (3.62) holds. Hence, $L_{uV\mathcal{S}}$ is surjective if and only if

$$\begin{bmatrix} \mathrm{Re}((V^\top \otimes u^H)B_{\mathcal{S}}) \\ \mathrm{Im}((V^\top \otimes u^H)B_{\mathcal{S}}) \end{bmatrix}$$

is full row rank, i.e., (3.60) is satisfied. □

Corollary 3.13. *Let $u \in \mathbb{C}^n$, and $V \in \mathbb{C}^{m\times(m-n+1)}$. Then $L_{uV\mathcal{S}}$ is surjective only if $\dim_{\mathbb{R}} \mathcal{S} \geq 2(m-n+1)$.*

Proposition 3.12 requires at least $2(m-n+1)$ columns of the basis matrix $B_{\mathcal{S}}$ to lie outside of the null space of $(V^\top \otimes u^H)$. It is both the dimension of $\mathcal{S}$ (i.e., the number of columns of $B_{\mathcal{S}}$) and the null space of $(V^\top \otimes u^H)$ that determines surjectivity of $L_{uV\mathcal{S}}$. Not all pairs of matrices $u \in \mathbb{C}^n$ and $V \in \mathbb{C}^{m\times(m-n+1)}$ induce a linear operator $L_{uV\mathcal{S}}$ which is surjective. For example, if $u = 0$ or $V = 0$ clearly $L_{uV\mathcal{S}}$ is not surjective.

We complete this discussion on surjectivity of the linear operators $L_{uV\mathcal{S}}$ by considering the special case of real perturbations, i.e., $\mathcal{S} \subset \mathbb{R}^{n\times m}$. In this subset of $\mathcal{P}$–robustness problems, the added structure leads to several strongly sufficient conditions for surjectivity that have analogous results for the particular problem considered in [28].

Corollary 3.14. *Let $\mathcal{S} = \mathbb{R}^{n\times m}$ and $\mathrm{rank}(Im(M-R)) = n$ for a fixed $R \in \mathcal{P}$. Let $\delta M \in \mathcal{S}$ be any perturbation such that $\mathrm{rank}[M - R - \delta M] < n$ and let $u \in \mathbb{C}^n$ be the n^{th} lsv of $M - R - \delta M$ and $V \in \mathbb{C}^{m\times(m-n+1)}$ have columns equal to the last $m-n+1$ rsv of $M - R - \delta M$. Then $L_{uV\mathcal{S}}$ is surjective.*

Proof. Since $\mathcal{S} = \mathbb{R}^{n\times m}$, without loss of generality let $B_{\mathcal{S}} = I_{mn}$. Let $y_1, y_2 \in \mathbb{C}^{m-n+1}$ be any vectors such that $y_1^\top \mathrm{Re}[V^\top \otimes u^H] + y_2^\top \mathrm{Im}[V^\top \otimes u^H] = 0$. Using appropriate Kronecker product identities, one can verify $(V^\top \otimes u^H) = V^\top(I_m \otimes u^H)$. Recall that

$(A+B)\otimes(C+D) = A\otimes C + B\otimes C + A\otimes D + B\otimes D$ for matrices A, B, C, D of appropriate dimension. Hence, applying $B_{\mathcal{S}} = I_{mn}$, we observe

$$\begin{aligned}
&\begin{bmatrix} \mathrm{Re}[(V^\top \otimes u^H)B_{\mathcal{S}}] \\ \mathrm{Im}[(V^\top \otimes u^H)B_{\mathcal{S}}] \end{bmatrix} \\
&\qquad = \begin{bmatrix} \mathrm{Re}(V^\top) & -\mathrm{Im}(V^\top) \\ \mathrm{Im}(V^\top) & \mathrm{Re}(V^\top) \end{bmatrix} \begin{bmatrix} I_m \otimes \mathrm{Re}(u^H) \\ I_m \otimes \mathrm{Im}(u^H) \end{bmatrix} \\
&\qquad \triangleq \overline{V}^\top \begin{bmatrix} I_m \otimes \mathrm{Re}(u^H) \\ I_m \otimes \mathrm{Im}(u^H) \end{bmatrix}.
\end{aligned}$$

Since V has orthonormal columns,

$$\begin{aligned}
V^H V &= \mathrm{Re}(V^\top)\,\mathrm{Re}(V) - \mathrm{Im}(V^\top)\,\mathrm{Im}(V) + j(\mathrm{Re}(V^\top)\,\mathrm{Im}(V) + \mathrm{Im}(V^\top)\,\mathrm{Re}(V)) \\
&= I_{m-n+1}
\end{aligned}$$

implying

$$\overline{V}^\top \begin{bmatrix} \mathrm{Re}(V) & -\mathrm{Im}(V) \\ \mathrm{Im}(V) & \mathrm{Re}(V) \end{bmatrix} = I_{2(m-n+1)}.$$

Thus $\overline{V}^\top$ has a right inverse and must have full row rank. Consequently, $[y_1^\top, y_2^\top]\overline{V}^\top = 0$ only if $y_1 = y_2 = 0$. Let $[\nu_1^\top, \nu_2^\top] = [y_1^\top, y_2^\top]\overline{V}^\top$. Then $\nu_1^\top(I_m \otimes \mathrm{Re}(u^H)) + \nu_2(I_m \otimes \mathrm{Im}(u^H)) = 0$. However, this implies

$$\begin{bmatrix} \nu_1 & \nu_2 \end{bmatrix} \begin{bmatrix} \mathrm{Re}(u^H) \\ \mathrm{Im}(u^H) \end{bmatrix} = 0. \tag{3.63}$$

By [28, Proposition 3.3], since $\mathrm{rank}(\mathrm{Im}(M-R)) = n$ and $\mathcal{S} = \mathbb{R}^{n\times m}$ any left null vector u of $M - R - \delta M$ satisfies $\mathrm{rank}[\mathrm{Re}(u), \mathrm{Im}(u)] = 2$ which by (3.63) implies $\nu_1 = \nu_2 = 0$ and thus $y_1 = y_2 = 0$, i.e., (3.60) is full row rank. Hence $L_{uV\mathcal{S}}$ is surjective by Proposition 3.12. □

Corollary 3.14 provides conditions sufficient for surjectivity. The next two results use a structural condition which is stronger than surjectivity of $L_{uV\mathcal{S}}$. Proposition 3.15 proves that for any unit vector u and real perturbation space $\mathcal{S}$ satisfying (3.64) a perturbation matrix $\delta M \in \mathcal{S}$ exists such that u is a left null vector of $M - \delta M$.

Proposition 3.15. *Let $u \in \mathbb{C}^n$ be unit vector and $m \geq n \geq 2$. Let $\mathcal{S} \subset \mathbb{R}^{n \times m}$ be a real vector space having basis $\{S_1, S_2, \ldots, S_k\}$ and let $B_{\mathcal{S}} = [vec(S_1), vec(S_2), \ldots, vec(S_k)]$. If*

$$\text{rank}[Z(u)] \triangleq \text{rank}\left(\begin{bmatrix} I_m \otimes \text{Re}(u^H) \\ I_m \otimes \text{Im}(u^H) \end{bmatrix} B_{\mathcal{S}}\right) = 2m, \tag{3.64}$$

then for each $M \in \mathbb{C}^{n \times m}$, there exists a perturbation $\delta M \in \mathcal{S}$ such that u is a left null vector of $M - \delta M$, i.e., $u^H(M - \delta M) = 0$.

Proof. Let $M \in \mathbb{C}^{n \times m}$. Let $x_0^H \triangleq -u^H M$. If $x_0 = 0$, then $\delta M = 0$ satisfies the required conditions trivially. Assume $x_0 \neq 0$. Since $Z(u)$ is full row rank, $Z^\dagger(u)$ is a right inverse of $Z(u)$. Let $\delta M \in \mathcal{S}$ be the matrix satisfying $\text{vec}(\delta M) = B_{\mathcal{S}}\zeta$ where $\zeta \in \mathbb{R}^k$ satisfies

$$\zeta = Z^\dagger(u) \begin{bmatrix} \text{Re}(\text{cj}(x)) \\ \text{Im}(\text{cj}(x)) \end{bmatrix},$$

where $\text{cj}(x) = \text{Re}(x) - j\,\text{Im}(x)$ is the complex conjugate. Then since $\text{vec}(u^H \delta M) = (u^H \delta M)^\top = (I_m \otimes u^H)\,\text{vec}(\delta M)$ and $\text{vec}(\delta M) = B_{\mathcal{S}}\zeta$,

$$\begin{bmatrix} \text{Re}((u^H \delta M)^\top) \\ \text{Im}((u^H \delta M)^\top) \end{bmatrix} = Z(u)\zeta = \begin{bmatrix} \text{Re}(\text{cj}(x)) \\ \text{Im}(\text{cj}(x)) \end{bmatrix}.$$

Hence $u^H(M - \delta M) = -x_0^H + x_0^H = 0$ implying u is a left null vector of $M - \delta M$. □

Proposition 3.16 proves that the condition in (3.64) guarantees that if $\text{rank}[M - \delta M - R] = n - 1$ then there exists a neighborhood of $M - \delta M - R$ for which rank reducing perturbations exist with a Frobenius norm bounded by the growth of the n^{th} singular value in this neighborhood. A similar unproven result is proposed in [38].

Proposition 3.16. *Let $\delta M' \in \mathcal{S}$ and $R' \in \mathcal{P}$ be such that $\widetilde{M} \triangleq M - \delta M' - R' \in \mathbb{C}^{n \times m}$ satisfies $\text{rank}[\widetilde{M}] = n - 1$ and let u be the n^{th} lsv of $\widetilde{M}$. Let $\mathcal{S} \subset \mathbb{R}^{n \times m}$ be a real vector space with orthonormal basis $\{S_1, S_2, \ldots, S_k\}$ with*

$$B_{\mathcal{S}} = [vec(S_1), vec(S_2), \ldots, vec(S_k)].$$

By orthonormal we mean $\langle S_i, S_j \rangle = \delta_{ij}$*, where* $\delta_{ij} = 1$ *if* $i = j$ *and* 0 *otherwise. If*

$$\operatorname{rank}[Z(u)] \triangleq \operatorname{rank}\left(\begin{bmatrix} I_m \otimes \operatorname{Re}(u^H) \\ I_m \otimes \operatorname{Im}(u^H) \end{bmatrix} B_{\mathcal{S}}\right) = 2m, \tag{3.65}$$

there exists constants c *and* K *such that for every* $N \in \mathbb{C}^{n\times m}$ *with* $\|N\|_F < c$ *there is a* $\delta M \in \mathcal{S}$ *satisfying* $\|\delta M\|_F \le K\sigma_n(\widetilde{M} - N)$ *and* $\operatorname{rank}(\widetilde{M} - N - \delta M) < n$*.*

Proof. Since the last singular value of $\widetilde{M}$ is distinct there exists a neighborhood of $\widetilde{M}$ (call it $\widetilde{M} - N$ for $\|N\|_F < d$) wherein all matrix valued functions $N(\alpha)$ which depend analytically on the real scalar α and satisfy $\|N(\cdot)\|_F < d$ have a n^{th} lsv function $\widetilde{u}(\alpha)$ of $\widetilde{M} - N(\alpha)$ which can be chosen to be an analytic function of α, (See [39, 44] for more details). As a result, there exists $c > 0$ small enough for which there exists $\epsilon = \epsilon(c) > 0$ such that for each N with $\|N\|_F < c$ the n^{th} lsv $\widetilde{u}$ of $\widetilde{M} - N$ satisfies

$$\sigma_{2m}(Z(\widetilde{u})) \ge \epsilon > 0. \tag{3.66}$$

Consider one specific N satisfying $\|N\|_F < c$ and n^{th} lsv $\widetilde{u}$ of $\widetilde{M} - N$. Let $x_0 \in \mathbb{C}^m$ be the unique vector satisfying

$$\widetilde{u}^H(\widetilde{M} - N) - x_0^H = 0.$$

We will now construct a perturbation $\delta M \in \mathcal{S}$ such that $\widetilde{u}^H \delta M = x_0^H$. If $x_0 = 0$, then $\operatorname{rank}(\widetilde{M} - N) < n$ and $\delta M = 0$ satisfies the conditions of the lemma. If $x_0 \neq 0$, we note that $\widetilde{u}^H \delta M = x_0^H$ if and only if $\operatorname{Re}(x_0^H) = \operatorname{Re}(\widetilde{u}^H \delta M)$ and $\operatorname{Im}(x_0^H) = \operatorname{Im}(\widetilde{u}^H \delta M)$. Equivalently, $\widetilde{u}^H \delta M = x_0^H$ if and only if

$$\begin{bmatrix} \operatorname{vec}(\operatorname{Re}(x_0^H)) \\ \operatorname{vec}(\operatorname{Im}(x_0^H)) \end{bmatrix} = \begin{bmatrix} \operatorname{Re}(\operatorname{vec}(\widetilde{u}^H \delta M)) \\ \operatorname{Im}(\operatorname{vec}(\widetilde{u}^H \delta M)) \end{bmatrix}. \tag{3.67}$$

Since $\operatorname{vec}(\widetilde{u}^H \delta M) = (I_m \otimes u^H)\operatorname{vec}(\delta M)$ and $\operatorname{vec}(\delta M) = B_{\mathcal{S}}\zeta_0$ for some $\zeta_0 \in \mathbb{R}^k$, $\widetilde{u}^H \delta M = x_0^H$ if and only if

$$\begin{bmatrix} \operatorname{vec}(\operatorname{Re}(x_0^H)) \\ \operatorname{vec}(\operatorname{Im}(x_0^H)) \end{bmatrix} = Z(\widetilde{u})\zeta_0.$$

Since $Z(\widetilde{u})$ has full row rank by (3.66), $Z(\widetilde{u})^\dagger$ is a right inverse. Then $\delta M \triangleq B_{\mathcal{S}}\zeta_0$ with

$$\zeta_0 \triangleq Z(\widetilde{u})^\dagger \begin{bmatrix} \operatorname{vec}(\operatorname{Re}(x_0^H)) \\ \operatorname{vec}(\operatorname{Im}(x_0^H)) \end{bmatrix}$$

satisfies $\widetilde{u}^H \delta M = x_0^H$. Thus $u^H(\widetilde{M} - N - \delta M) = 0$, i.e., $\operatorname{rank}(\widetilde{M} - N - \delta M) < n$. What remains is the show that is the bound on $\|\delta M\|_F$. Since $B_{\mathcal{S}}$ has orthonormal columns $\|\delta M\|_F = \|\zeta_0\|_2$. Hence

$$\begin{aligned} \|\delta M\|_F &\leq \sigma_1(Z(\widetilde{u})^\dagger) \left\| \begin{bmatrix} \operatorname{vec}(\operatorname{Re}(x_0^H)) \\ \operatorname{vec}(\operatorname{Im}(x_0^H)) \end{bmatrix} \right\|_2 \\ &= \frac{1}{\epsilon} \|\widetilde{u}^H(\widetilde{M} - N)\|_2. \end{aligned}$$

Since $\widetilde{u}$ is the n^{th} lsv of $\widetilde{M} - N$, $\|\widetilde{u}^H(\widetilde{M} - N)\|_2 = \sigma_n(\widetilde{M} - N)$. Letting $K = 1/\epsilon$, we obtain the desired bound $\|\delta M\| \leq K\sigma_n(\widetilde{M} - N)$. □

Propositions 3.12, 3.15, and 3.16 demonstrate the usefulness of fixing a basis for $\mathcal{S}$ (and later for $\mathcal{P}$) for verifying system properties.

3.8.2 Additional Proofs

Proof of Lemma 3.3.

Step 1: First we show that every analytic path $g_a : [0,1] \to W$ from W_1 to W_2 has an analytic unsigned n^{th} singular value function $f_a : [0,1] \to \mathbb{R}$ such that $f_a(s) = \widetilde{H}(g_a(s))$, i.e., $\widetilde{H}$ is an unsigned n^{th} singular value function for each analytic path in W. Without loss of generality assume $f_a(0) = \widetilde{H}(g_a(0))$. Recall $\widetilde{H}$ was constructed to be consistent with f_g, the unsigned n^{th} singular value function associated with the curve g. Construct a continuous closed path by connecting $g_a(0)$ with $g(0)$ in W_1 and $g_a(1)$ with $g(1)$ in W_2. Since σ_n is continuous and nonzero in $W_1 \cup W_2$, σ_n is continuous on the closed path and thus $\operatorname{sign}(f_a(1)) = \operatorname{sign}(f_g(1))$, i.e., $f_a(s) = \widetilde{H}(g_a(s))$ as desired. Consequently, $\widetilde{H}$ can be used for an unsigned n^{th} singular value for any analytic curve in W.

Step 2: if $(\widetilde{\zeta},\widetilde{\rho}) \in W_1 \cup W_2$ then the n^{th} singular value of $M(\widetilde{\zeta},\widetilde{\rho})$ is distinct and hence the n^{th} lsv and rsv are unique up to multiplication by unitary scalars. By [39, Theorem 3], there exists a neighborhood $W_0 \subset W_1 \cup W_2$ of $(\widetilde{\zeta},\widetilde{\rho})$ and analytic (unsigned) singular vector functions $\widetilde{u} : W_0 \to \mathbb{C}^n$ and $\widetilde{v} : W_0 \to \mathbb{C}^m$ such that for all $(\zeta,\rho) \in W_0$,

$$\widetilde{u}^H(\zeta,\rho)M(\zeta,\rho)\widetilde{v}(\zeta,\rho) = \widetilde{H}(\zeta,\rho).$$

Since $\widetilde{u}$, $\widetilde{v}$, and $M(\zeta,\rho)$ are analytic, so is $\widetilde{H}$.

To take the derivative $\frac{\partial \widetilde{H}(\widetilde{\zeta},\widetilde{\rho})}{\partial \zeta_i}$, we consider replacing ζ with a complex argument z.[3] The complex partial derivative with respect to z_i at $z = \widetilde{\zeta}$ is

$$\begin{aligned}\frac{\partial \widetilde{H}(z,\widetilde{\rho})}{\partial z_i} &= \frac{\partial \widetilde{u}^H}{\partial z_i} M(\widetilde{\zeta},\widetilde{\rho})\widetilde{v} + \widetilde{u}^H \frac{\partial M(z,\widetilde{\rho})}{\partial z_i}\widetilde{v} \\ &\quad + \widetilde{u}^H M(\widetilde{\zeta},\widetilde{\rho})\frac{\partial \widetilde{v}}{\partial z_i} \\ &= \widetilde{H}(\widetilde{\zeta},\widetilde{\rho})\left(\frac{\partial \widetilde{u}^H}{\partial z_i}\widetilde{u} + \widetilde{v}^H \frac{\partial \widetilde{v}}{\partial z_i}\right) - \widetilde{u}^H S_i \widetilde{v},\end{aligned}$$

since $\widetilde{u}$ and $\widetilde{v}$ are singular value functions, i.e.,

$$M(\widetilde{\zeta},\widetilde{\rho})\widetilde{v}(\widetilde{\zeta},\widetilde{\rho}) = \widetilde{H}(\widetilde{\zeta},\widetilde{\rho})\widetilde{u}(\widetilde{\zeta},\widetilde{\rho})$$

and

$$\widetilde{u}^H(\widetilde{\zeta},\widetilde{\rho})M(\widetilde{\zeta},\widetilde{\rho}) = \widetilde{H}(\widetilde{\zeta},\widetilde{\rho})\widetilde{v}^H(\widetilde{\zeta},\widetilde{\rho}).$$

However, since $1 = \widetilde{u}^H\widetilde{u}$, we obtain $0 = \frac{\partial \widetilde{u}^H}{\partial z_i}\widetilde{u} + \left(\frac{\partial \widetilde{u}^H}{\partial z_i}\widetilde{u}\right)^H$. Hence $\mathrm{Re}\left(\frac{\partial \widetilde{u}^H}{\partial z_i}\widetilde{u}\right) = 0$. Similarly, $\mathrm{Re}\left(\widetilde{v}^H \frac{\partial \widetilde{v}}{\partial z_i}\right) = 0$. The real derivative of $\widetilde{H}$ with respect to ζ_i is given by

$$\begin{aligned}\frac{\partial \widetilde{H}(\widetilde{\zeta},\widetilde{\rho})}{\partial \zeta_i} &= \mathrm{Re}\left[\frac{\partial \widetilde{H}(z,\widetilde{\rho})}{\partial z_i}\right] \\ &= -\,\mathrm{Re}(\widetilde{u}^H(\widetilde{\zeta},\widetilde{\rho})S_i\widetilde{v}(\widetilde{\zeta},\widetilde{\rho})).\end{aligned}$$

The same argument holds for computing $\frac{\partial \widetilde{H}(\widetilde{\zeta},\rho)}{\partial \rho_i} = -\,\mathrm{Re}(\widetilde{u}^H(\widetilde{\zeta},\widetilde{\rho})P_i\widetilde{v}(\widetilde{\zeta},\widetilde{\rho}))$.

[3] Rigorously, we should define a new function with a complex domain, but we have chosen to keep the presentation more direct.

Step 3: what remains is to show that $\widetilde{H}$ is Fréchet differentiable for $(\zeta_0, \rho_0) \in W \setminus W_1 \cup W_2$, i.e., points where $M(\zeta_0, \rho_0)$ drops rank. Although $\sigma_n(M(\zeta_0, \rho_0)) = 0$, the n^{th} lsv is unique (up to unitary scalar multiplication) since the last singular value is distinct and $M(\zeta_0, \rho_0)$ has fewer rows than columns. The problem with the n^{th} (unsigned) rsv is that it is not unique since $M(\zeta, \rho_0)$ has a right null space of dimension $m - n + 1$. However, not all n^{th} (unsigned) rsv can be part of an analytic singular value function for analytic paths passing through (ζ_0, ρ_0). Let $g_b : [0, 1] \to W$ be an analytic curve from W_1 to W_2 with $g_b(0.5) = (\zeta_0, \rho_0)$. By [39], there exists analytic n^{th} (unsigned) lsv and rsv functions $u_b : [0, 1] \to \mathbb{C}^n$ and $v_b : [0, 1] \to \mathbb{C}^m$, respectively, associated with the unsigned n^{th} singular value function $\widetilde{H}(g_b(\cdot))$. Since for each $s \neq 0.5$, $M(g_b(s))$ is a fixed matrix with a nonzero n^{th} singular value, the product $u_b(s)v_b^H(s)$ is unique (even though $u_b(s)$ and $v_b(s)$ are not unique). The same uniqueness result holds for analytic curves from W_2 to W_1 passing through (ζ_0, ρ_0). For analytic paths in $W \setminus W_1 \cup W_2$ passing through (ζ_0, ρ_0), the right null space of $M(\zeta, \rho)$ changes analytically and hence one can choose $u_b(0.5)$ and $v_b(0.5)$ as the n^{th} unsigned lsv and rsv along these paths as well. Because W, W_1, and W_2 are simply connected, there exists a neighborhood W_0 of (ζ_0, ρ_0) and analytic singular vector functions $u_0 : W_0 \to \mathbb{C}^n$ and $v_0 : W_0 \to \mathbb{C}^m$ such that

$$u_0^H(\zeta, \rho)M(\zeta, \rho)v_0(\zeta, \rho) = \widetilde{H}(\zeta, \rho)$$

for all $(\zeta, \rho) \in W_0$. Using the same arguments as in Step 2 it follows that $\widetilde{H}$ is Fréchet differentiable at $(\zeta_0, \rho_0) \in W \setminus W_1 \cup W_2$ with partial derivatives given in (3.19). $\square$

Proof of Lemma 3.4.

Let $\delta\widetilde{M} \triangleq (L_{uV\mathcal{S}}^\dagger \circ L_{uV})(M - R_0)$. Since $L_{uV\mathcal{S}}$ is surjective, $L_{uV\mathcal{S}}(\delta\widetilde{M}) = L_{uV}(M - R_0)$. Since $\sigma_n(M - \delta M_0 - R_0) = 0$, $L_{uV\mathcal{S}}(\delta M_0) = L_{uV}(M - R_0)$. Hence, defining $\Delta M \triangleq \delta\widetilde{M} - \delta M_0$, we obtain

$$L_{uV\mathcal{S}}(\Delta M) = L_{uV\mathcal{S}}(\delta\widetilde{M}) - L_{uV\mathcal{S}}(\delta M_0) = 0.$$

Observe that by definition $\Delta M = -(I - L^{\dagger}_{uV\mathcal{S}} \circ L_{uV\mathcal{S}})(\delta M_0) = -(I - Q_1)(\delta M_0)$, where the linear operator Q_1 is

$$Q_1 \triangleq (L^{\dagger}_{uV\mathcal{S}} \circ L_{uV\mathcal{S}}). \tag{3.68}$$

By the properties of the Moore-Penrose pseudoinverse, Q_1 is an orthogonal projection on $\mathcal{S}$ with respect to the inner product $\langle \cdot, \cdot \rangle$, i.e., Q_1 is a self-adjoint linear operator and $Q_1^2 = Q_1$. In addition, $I - Q_1$ is also an orthogonal projection on $\mathcal{S}$. Hence,

$$\begin{aligned}\langle \delta M_0, \Delta M \rangle &= -\langle \delta M_0, (I - Q_1)(\delta M_0) \rangle \\ &= -\langle (I - Q_1)(\delta M_0), (I - Q_1)(\delta M_0) \rangle \\ &= -\|(I - Q_1)(\delta M_0)\|_F^2.\end{aligned}$$

Since $\Delta M = -(I - Q_1)(\delta M_0) \neq 0$ because $\delta M_0 \neq \delta \widetilde{M}$, $\|(I - Q_1)(\delta M_0)\|_F \neq 0$. Thus $\langle \delta M_0, \Delta M \rangle = -\|(I - Q_1)(\delta M_0)\|_F^2 < 0$. □

Proof of Lemma 3.5.

Note that since $\Delta R \neq 0$, $(L^{\dagger}_{uV\mathcal{S}} \circ L_{uV\mathcal{P}})(\Delta R) \neq 0$ by the properties of the Moore-Penrose pseudoinverse. Let $\Delta M \in \mathcal{S}$ be defined as $\Delta M \triangleq -(L^{\dagger}_{uV\mathcal{S}} \circ L_{uVP})(\Delta R)$. Since $L_{uV\mathcal{S}}$ is surjective, $L_{uV\mathcal{S}}(\Delta M) = -L_{uV\mathcal{P}}(\Delta R)$. Hence, by linearity

$$L_{uV}(\Delta M + \Delta R) = L_{uV\mathcal{S}}(\Delta M) + L_{uV\mathcal{P}}(\Delta R) = 0.$$

Since $\delta M_0 = (L^{\dagger}_{uV\mathcal{S}} \circ L_{uV})(M - R_0) =$ and δM_0 is rank reducing

$$\begin{aligned}\Delta M &= -(L^{\dagger}_{uV\mathcal{S}} \circ L_{uV\mathcal{P}})(L^{\dagger}_{uV\mathcal{S}} \circ L_{uV\mathcal{P}})^{\dagger} \delta M_0 \\ &= -(I - Q_2)(\delta M_0),\end{aligned}$$

where the linear operator Q_2 is

$$Q_2 \triangleq I - (L^{\dagger}_{uV\mathcal{S}} \circ L_{uV\mathcal{P}})(L^{\dagger}_{uV\mathcal{S}} \circ L_{uV\mathcal{P}})^{\dagger}. \tag{3.69}$$

By the properties of the Moore-Penrose pseudoinverse, Q_2 is an orthogonal projection on $\mathcal{S}$ with respect to the inner product $\langle \cdot, \cdot \rangle$, i.e., Q_2 is a self-adjoint linear operator and $Q_2^2 = Q_2$. In addition, $I - Q_2$ is also an orthogonal projection on $\mathcal{S}$. This implies

$$\begin{aligned}\langle \delta M_0, \Delta M \rangle &= -\langle \delta M_0, (I - Q_2)(M_0) \rangle \\ &= -\langle (I - Q_2)(\delta M_0), (I - Q_2)(M_0) \rangle \\ &= -\|(I - Q_2)(\delta M_0)\|_F^2 \\ &= -\|\Delta M\|_F^2\end{aligned}$$

Since $(L_{uV\mathcal{S}}^{\dagger} \circ L_{uV\mathcal{P}})(\Delta R) \neq 0$ by construction of ΔR, we conclude $\Delta M \neq 0$. Consequently, $\langle \delta M_0, \Delta M \rangle < 0$. □

Proof of Lemma 3.6.

By definition, f is differentiable with derivative $f'(\zeta_0, \rho_0) = [\zeta_0^\top, 0]$. Since the Fréchet differential $\widetilde{H}'(\zeta_0, \rho_0)$ exists by Lemma 3.3, T is Fréchet differentiable. Let $y = [y_1, y_2]^\top \in \mathbb{R}^2$ be an arbitrary vector. Since $\frac{\partial}{\partial \zeta} \widetilde{H}(\zeta, \rho_0)|_{\zeta = \zeta_0}$ is surjective, there exists $\zeta_1 \in \mathbb{R}^k$ such that $\frac{\partial}{\partial \zeta} \widetilde{H}(\zeta, \rho_0)|_{\zeta = \zeta_0} \zeta_1 = y_2$. This implies that

$$T'(\zeta_0, \rho_0) \begin{bmatrix} \zeta_1 \\ 0 \end{bmatrix} = \begin{bmatrix} \zeta_0^\top \zeta_1 \\ y_2 \end{bmatrix}.$$

By assumption ii), there exists ζ_Δ and ρ_Δ such that $\widetilde{H}'(\zeta_0, \rho_0)[\zeta_\Delta^\top, \rho_\Delta^\top]^\top = 0$ and $\zeta_0^\top \zeta_\Delta < 0$. Define $\zeta_2 \in \mathbb{R}^k$ and $\rho_2 \in \mathbb{R}^r$ by

$$\begin{aligned}\zeta_2 &= \left(\frac{y_1 - \zeta_0^\top \zeta_1}{\zeta_0^\top \zeta_\Delta} \right) \zeta_\Delta \\ \rho_2 &= \left(\frac{y_1 - \zeta_0^\top \zeta_1}{\zeta_0^\top \zeta_\Delta} \right) \rho_\Delta\end{aligned}$$

Then by linearity,

$$\begin{aligned}T'(\zeta_0, \rho_0) \begin{bmatrix} \zeta_1 + \zeta_2 \\ \rho_2 \end{bmatrix} &= \begin{bmatrix} \zeta_0^\top \zeta_1 \\ y_2 \end{bmatrix} + \begin{bmatrix} y_1 - \zeta_0^\top \zeta_1 \\ 0 \end{bmatrix} \\ &= y.\end{aligned}$$

Since y was arbitrary, $T'(\zeta_0, \rho_0)$ is surjective. □

Proof of Lemma 3.7.

Let $\widehat{v}_k^H = [u_n]_k^H(\delta\widehat{M}_k - \delta M_k + \Delta R_k)(I - V_k V_k^H)$ and let $\widehat{u}_k$ satisfy $\widehat{u}_k^H(M - \delta M_k - R_k) = \widehat{v}_k^H$; such a $\widehat{u}_k$ exists since $\widehat{v}_k$ is in the row space of $M - \delta M_k - R_k$. Consider the product

$$\begin{aligned}
([u_n]_k + \alpha\widehat{u}_k)^H&(M - \delta M_k - R_k - \alpha(\delta\widehat{M}_k - \delta M_k + \Delta R_k)) \\
&= [u_n]_k^H(M - \delta M_k - R_k) \\
&\quad - \alpha\left(\widehat{v}_k^H - [u_n]_k^H(\delta\widehat{M}_k - \delta M_k + \Delta R_k)\right) \\
&\quad - \alpha^2\widehat{u}_k^H(\delta\widehat{M}_k - \delta M_k + \Delta R_k) \\
&= (1-\alpha)L_{uV}(M - \delta M_k - R_k)V_k^H \\
&\quad + \alpha L_{uV}(M - \delta\widehat{M}_k - R_k - \Delta R_k)V_k^H \qquad (3.70) \\
&\quad - \alpha^2\widehat{u}_k^H(\delta\widehat{M}_k - \delta M_k + \Delta R_k).
\end{aligned}$$

Since $L_{uV\mathcal{S}}$ is surjective, step 6 of Algorithm 1 guarantees $L_{uV}(M - \delta\widehat{M}_k - R_k - \Delta R_k) = 0$. Recall that for all $u \in \mathbb{C}^n$, $\widetilde{M} \in \mathbb{C}^{n\times m}$, $u^H\widetilde{M} \geq \sigma_n(\widetilde{M})\|u\|$ (See [42, Corollary 9.6.7]). Combining this with the fact that $[u_n]_k$ and $\widehat{u}_k$ are orthogonal, implying $\|[u_n]_k + \alpha\widehat{u}_k\| \geq \|[u_n]_k\| = 1$, the norm of the left-hand side of (3.70) upper bounds $\sigma_n(M - \delta M_k - R_k - \alpha(\delta\widehat{M}_k - \delta M_k + \Delta R_k))$. Taking the norm of both sides of (3.70) and applying the triangle inequality results in the statement of the lemma. □

Proof of Lemma 3.8.

By definition of $\delta\overline{M}_k$, $\delta\widehat{M}_k = (Q_2 \circ Q_1)(\delta M_k) + \delta\overline{M}_k$. If $\delta M_k = 0$, $(Q_2 \circ Q_1)(\delta M_k) = 0 = \delta\overline{M}_k$ and (3.56) holds trivially. Assume $\delta M_k \neq 0$. Then since Q_1 and Q_2 are orthogonal projections $\|(Q_2 \circ Q_1)(\delta M_k)\|_F \leq \|\delta M_k\|_F$. Hence,

$$\begin{aligned}
&\|(1-\alpha)\delta M_k + \alpha(Q_2 \circ Q_1)(\delta M_k)\|_F - \|\delta M_k\|_F \\
&= \frac{\|(1-\alpha)\delta M_k + \alpha(Q_2 \circ Q_1)(\delta M_k)\|_F^2 - \|\delta M_k\|_F^2}{\|(1-\alpha)\delta M_k + \alpha(Q_2 \circ Q_1)(\delta M_k)\|_F + \|\delta M_k\|_F} \\
&\leq \frac{\|(1-\alpha)\delta M_k + \alpha(Q_2 \circ Q_1)(\delta M_k)\|_F^2 - \|\delta M_k\|_F^2}{2\|\delta M_k\|_F}. \qquad (3.71)
\end{aligned}$$

Since $\delta\widehat{M}_k = (Q_2 \circ Q_1)(\delta M_k) + \delta\overline{M}_k$, the triangle inequality implies that

$$\begin{aligned}
&\|(1-\alpha)\delta M_k + \alpha\delta\widehat{M}_k\|_F \\
&\quad \leq \|(1-\alpha)\delta M_k + \alpha(Q_2 \circ Q_1)(\delta M_k)\|_F + \alpha\|\delta\overline{M}_k\|_F.
\end{aligned}$$

Applying (3.71) yields the desired result. □

Proof of Lemma 3.9.

To prove necessity, assume R and δM satisfy necessary conditions i) and ii) of Theorem 3.2. Then by i)

$$\begin{aligned}
Q_1(\delta M) &= (L^\dagger_{uV\mathcal{S}} \circ L_{uV\mathcal{S}})\delta M \\
&= (L^\dagger_{uV\mathcal{S}} \circ L_{uV})(M - R) \\
&= \delta M,
\end{aligned}$$

i.e., $Q_1(\delta M) = \delta M$. By necessary condition ii),

$$0 = \Delta R \triangleq (L^\dagger_{uV\mathcal{S}} \circ L_{uV\mathcal{P}})^\dagger (L^\dagger_{uV\mathcal{S}} \circ L_{uV})(M - R).$$

Using the definitions of Q_1 and Q_2, we have

$$\begin{aligned}
(Q_2 \circ Q_1)(\delta M) &= Q_1(\delta M) + (L^\dagger_{uV\mathcal{S}} \circ L_{uV\mathcal{P}})\Delta R \\
&= \delta M.
\end{aligned}$$

Thus $(Q_2 \circ Q_1)(\delta M) = \delta M$ as desired.

Now for sufficiency, assume that $(Q_2 \circ Q_1)(\delta M) = \delta M$. Since Q_1 and Q_2 are orthogonal projections

$$\|\delta M\|_F = \|(Q_2 \circ Q_1)(\delta M)\|_F \leq \|Q_1(\delta M)\|_F \leq \|\delta M\|_F,$$

implying that equality holds. Thus $\|\delta M\|_F = \|Q_1(\delta M)\|_F$ and this implies that $\delta M = Q_1(\delta M)$ since Q_1 is an orthogonal projection. Similarly, we can show that $\delta M = Q_2(\delta M)$ since

$$\|\delta M\|_F = \|(Q_2 \circ Q_1)(\delta M)\|_F = \|Q_2(\delta M)\|_F \leq \|\delta M\|_F.$$

Since $Q_1(\delta M) = \delta M$, δM satisfies the first necessary condition in Theorem 3.2. What remains is to show that $\Delta R = 0$. Since $Q_2(\delta M) = \delta M$ and $L_{uV}(M - R) = L_{uV\mathcal{S}}\delta M$,

$$\begin{aligned} \Delta R &= (L^{\dagger}_{uV\mathcal{S}} \circ L_{uV\mathcal{P}})^{\dagger}\delta M \\ &= (L^{\dagger}_{uV\mathcal{S}} \circ L_{uV\mathcal{P}})^{\dagger} Q_2(\delta M). \end{aligned}$$

Thus by definition of Q_2, $\Delta R = T^{\dagger}\delta M - T^{\dagger}TT^{\dagger}\delta M = 0$, where $T = L^{\dagger}_{uV\mathcal{S}} \circ L_{uV\mathcal{P}}$ and $T^{\dagger}TT^{\dagger} = T^{\dagger}$ follows from the definition of the Moore Penrose pseudoinverse. □

4. SWITCHED SYSTEM OBSERVERS: REVIEW AND PROPOSED SOLUTION

In this chapter, relevant observer designs from the literature will be summarized. Following the literature review, the proposed embedded moving horizon observer will be introduced and preliminary results will be explored. For convenience, the embedded moving horizon observer will be reintroduced from Chapter 1.

The basic structure of a moving horizon is shown in Figure 1.3. The MHO problem is to consider a finite horizon $[t_f - T, t_f]$ of width T and choose an optimal state and mode estimate $\hat{x}(t)$ and $\hat{v}(t)$ to minimize the error between the measured output $y^M(t)$ and the estimator output $\hat{y}(t)$. Since a fixed initial condition and mode sequence uniquely describes a state trajectory which satisfies (1.2), the MHO problem over each horizon is to pick a single state $\hat{x}(t_f - h)$ for $0 \leq h \leq T$ (determining which time the state estimate is fixed) and the mode sequence $\hat{v}(t)$. To allow for continuous solvers, we embed the mode sequence into a larger class of trajectories.

For the two mode case, this means we expand the class range of $\hat{v}(t)$ from $\{0, 1\}$ which is original SLTI system to a range of $[0, 1]$. The embedded system estimator dynamics, again for a two mode SLTI system, has the form,

$$\begin{aligned}\dot{\hat{x}}(t) &= ((1 - \hat{v}(t))A_0 + \hat{v}(t)A_1)\hat{x}(t) \\ &\quad + ((1 - \hat{v}(t))B_0 + \hat{v}(t)B_1)u^M(t) \qquad (4.1a)\\ \hat{y}(t) &= ((1 - \hat{v}(t))C_0 + \hat{v}(t)C_1)\hat{x}(t). \qquad (4.1b)\end{aligned}$$

This embedding has shown promise in the area of switched optimal control and the trajectories of original switched system are dense in the set of embedded system trajectories [12]. This motivates the application of the embedded system in the moving horizon observer.

Embedded Moving Horizon Observer (EMHO) Problem: For each finite horizon $[t_f - T, t_f]$ and $0 \leq h \leq T$ the EHMO problem is given by

$$\min_{\substack{\hat{x}(t_f-h) \\ \hat{v}:[t_f-T,t_f]\to[0,1]}} \int_{t_f-T}^{t_f} \left\| y^M(t) - \hat{y}(t) \right\|^2 dt$$

subject to:

$$\begin{aligned}\dot{\hat{x}}(t) &= ((1-\hat{v}(t))A_0 + \hat{v}(t)A_1)\hat{x}(t) \\ &\quad + ((1-\hat{v}(t))B_0 + \hat{v}(t)B_1)u(t) \\ \hat{y}(t) &= ((1-\hat{v}(t))C_0 + \hat{v}(t)C_1)\hat{x}(t)\end{aligned}$$

where $u^M(t)$ is the measured input. The next horizon with final time t'_f is assumed to shift in time by δ, i.e. $t'_f = t_f + \delta$.

In addition to the EMHO, we will consider a modified EMHO scheme which adds a penalty for deviating from previous state estimates (if available). The modified EMHO scheme is given below.

Modified Embedded Moving Horizon Observer (MEMHO) Problem: For each finite horizon $[t_f - T, t_f]$ and $0 \leq h \leq T$ the EHMO problem is given by

$$\min_{\substack{\hat{x}(t_f-h) \\ \hat{v}:[t_f-T,t_f]\mapsto[0,1]}} \int_{t_f-T}^{t_f} \left\| y^M(t) - \hat{y}(t) \right\|^2 dt + \Gamma(\hat{x}(t_f - h))$$

subject to: $\gamma : \mathbb{R} \mapsto \mathbb{R}$ measurable penalty function

$$\begin{aligned}\dot{\hat{x}}(t) &= ((1-\hat{v}(t))A_0 + \hat{v}(t)A_1)\hat{x}(t) \\ &\quad + ((1-\hat{v}(t))B_0 + \hat{v}(t)B_1)u^M(t) \\ \hat{y}(t) &= ((1-\hat{v}(t))C_0 + \hat{v}(t)C_1)\hat{x}(t) \\ \Gamma(\hat{x}(t_f-h)) &= \int_{t_f-T}^{t_f-h} \gamma^2(t) \left\| \hat{x}(t) - \hat{x}_{prev}(t) \right\|^2 dt\end{aligned}$$

where $u^M(t)$ is the measured input and $\hat{x}_{prev}$ is the previous state estimate. If at any time t, $\hat{x}_{prev}(t)$ is unavailable, it is replaced with $\hat{x}(t)$ effectively removing it from the penalty term. The next horizon with final time t'_f is assumed to shift in time by δ, i.e. $t'_f = t_f + \delta$.

4.1 Switched System Observer Review

Many switched system observers in literature consider reconstructing the continuous state $x(t)$ but not the mode sequence $v(t)$ [45–47]. For example, in [47] a switched observer is constructed for a fixed and known mode sequence. Therein, the novelty is in using the knowledge of the switching sequence to reconstruct the state even when each subsystem may be unobservable. The main idea is to pick up states unobservable in one mode when one passes into another mode where these states may be observable. Many other techniques such as common Lyapunov functions for a Luenberger observer which can be used regardless of the mode sequence have also been proposed [45]. This review will emphasize switched system observers that simultaneously reconstruct the state and mode sequence and moving horizon observers.

4.1.1 Bank of State Observers

One popular method for reconstructing both the state and the mode is to construct classical observers for each subsystem. Then one determines the active mode by measuring which subsystem observer is tracking "most" effectively. This type of observer is explored for SLTI systems in [9–11]. The basic structure in these papers is summarized in Figure 1.3 in Chapter 1.

If the switched system is SMS observable with input u then the only subsystem observer which can accurately track the system output $y(t)$ is the correct mode. The complication comes from convergence rates. One may attempt to use a Luenberger observer for mode i which is a dynamic observer of the form

$$\dot{\hat{x}}_i = A_i\hat{x}_i + B_iu(t) + L_i(y - \hat{y}) \tag{4.2a}$$

$$\hat{y} = C_i\hat{x}_i, \tag{4.2b}$$

where L_i is the feedback gain matrix of dimension $n \times p$ designed such that $A_i - L_iC_i$ has eigenvalues in the open left-hand plane (which is possible if (A_i, C_i) is stabilizable). The issue with using a Luenberger observer is that the correct mode has error $e_i =$

$x - \hat{x}_i$ with dynamics $\dot{e}_i = (A_i - L_i C_i)e_i$ which converges exponentially, but when the mode changes from mode i to mode j the exponential convergence is no longer guaranteed globally. Moreover, exponential convergence still does not indicate perfect output tracking so determining which mode is active becomes a threshold problem. This potential issue is addressed in [11] by using thresholds for the residual signals $y - y_i$ for each subsystem to determine the active mode. Then under the assumption that the current mode is detected within a small enough delay δ in relation to the minimum dwell time T_{min}, appropriate conditions are developed in [11] to guarantee exponential convergence of the state estimation error.

In the case of a SLTI system, another form of observer can be used for each LTI subsystem which guarantees finite time convergence for the subsystem matching the active mode. This method is discussed in [9] where each subsystem observer in the bank of observers is a Super-Twisting observer. The Super-Twisting observer structure is beyond the scope of this review, but a few key ideas about how the Super-Twisting observer works can be made without excessive notation. The Super-Twisting observer is well known and uses a second order sliding mode algorithm. Here sliding mode refers a relay-like observer structure which is discontinuous. If each state has a uniformly bounded derivative and the LTI system is observable, the observer can be designed to converge in a given finite time τ (which is arbitrarily small).

The contribution in [9] is to point out that one can design these Super-Twisting observers for each mode k to have finite time convergence $\tau_k << T_{min}$, where T_{min} is the minimum dwell time. If each mode is observable and each pair of modes is distinguishable (generically), then only the observer corresponding to the active mode will converge (generically). The structure in Figure 1.3 with the Super-Twisting observer in each mode then guarantees state and mode reconstruction after $\min \tau_k$ seconds of every switching time.

Two main drawbacks arise when using the "bank of observers" structure in Figure 1.3. First and foremost is the added computation from running observers in each mode. If there are n states and M modes, then these methods will often require on the order of $n \times M$ integrators. The second difficulty in these methods is that the extension to the nonlinear case is more challenging since observers for each nonlinear mode will again be an increase in computation. A note here is that the second order sliding mode in [9] can double the number of integrators for the second order term.

4.1.2 Moving Horizon Observers

Nonlinear Case

The moving horizon observer reposes the estimation problem for a nonlinear system as an optimization problem. This method was popularized in [22]. This subsection will introduce the notation and the observer structure presented in [22]. Therein, the following nonlinear system was considered.

$$\dot{x}(t) = f(x(t), u(t)) \tag{4.3a}$$

$$y(t) = g(x(t)), \tag{4.3b}$$

where $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$, and $y(t) \in \mathbb{R}^p$ for all t, and f and g are known. Further it is assumed that the input $u(\cdot) \in L_\infty$, f and g are locally Lipschitz continuous with respect to both arguments, and $f(0,0) = 0$. For notation, a solution to (4.3a) at time t which passes through x_0 at time t_0 controlled with input u will be denoted $x^u(t; x_0, t_0)$.

For two times t_1 and t_2 and for a state estimate $w \in \mathbb{R}^n$ the estimation error will have measure $V_E(w; t_1, t_2)$ over the interval $[t_1, t_2]$ given by

$$V_E(w; t_1, t_2) = \int_{t_1}^{t_2} \left\| g(x^u(s; w, t_1)) - y^M(s) \right\|^2 ds, \tag{4.4}$$

where y^M is the measured output of the system. Note that the correct state estimate $w = x(t_1)$ will cause $V_E(x(t_1); t_2, t_2) = 0$ since the output of the estimator would match the measured output y^M. Guaranteeing that only the correct state estimate will cause the measure V_E to be zero is exactly the observability problem over the interval $[t_1, t_2]$. In [22], the reconstructibility assumption (assuming no finite escape times) has the following form.

Assumption 4.1. *There exists a horizon $T \in (0, \infty)$ and constant $\gamma \in (0, \infty)$ such that for any two boundary conditions (w_1, t), $(w_2, t) \in \mathbb{R}^n \times \mathbb{R}$ and any admissible control function $u \in L_\infty$ the L_2 norm of the difference between corresponding outputs given by*

$$W(w_1, w_2; t-T, t) \triangleq \int_{t-T}^{t} \left\| g(x^u(s; w_1, t)) - g(x^u(s; w_2, t)) \right\|^2 ds \tag{4.5}$$

satisfies

$$W(w_1, w_2; t-T, t) \geq \gamma \left\| w_1 - w_2 \right\|^2 . \tag{4.6}$$

The condition in (4.6) was shown in [22] to reduce to the observability gramian over the interval $[t - T, t]$ for the linear case. In the nonlinear case one can see that this condition follows locally if the local equivalent linear system is observable. When the system satisfies (4.6), exponential convergence of the subsequent starting horizon times can be achieved if at each new horizon the measure V_E satisfies a contraction with level $\beta \in (0, 1)$. This moving horizon observer algorithm is described in Algorithm1.

Switched System State Estimation with Known Mode

In the switched system literature, the paper [48] explores using a moving horizon observer for state estimation in a piecewise-affine (PWA) system with disturbances. The PWA system is a subclass of general switched systems. In [48] a PWA system without input is modeled in discrete-time in the following equations:

$$x(t+1) = A_i x(t) + f_i + d_a(t) \tag{4.8a}$$

$$y(t) = C_i x(t) + g_i + d_s(t), \text{ for } x(t) \in \mathcal{X}_i \tag{4.8b}$$

$$x \in \mathbb{X} = \mathbb{R}^{n_c} \times \{0, 1\}^{n_\ell} \tag{4.8c}$$

$$d_a \in \mathbb{W}, \tag{4.8d}$$

where x is a composite state containing n_c continuous states and n_ℓ logic states with $\mathbb{X} \subset \mathbb{R}^{n_c} \times \mathbb{R}^{n_\ell}$ a bounded polyhedron with polyhedral partition $\{\mathcal{X}_i\}_{i=1}^{s}$, sensor

Algorithm 1 Moving Horizon Observer

```
Data: $w_0 \in \mathbb{R}^n$, $T \in (0, \infty)$, $\beta \in (0, 1)$, the sampling time $\delta \in (0, T)$, the (measured) output function $y^M : [-T, 0] \mapsto \mathbb{R}^p$, and the control $u : [-T, \infty) \mapsto \mathbb{R}^m$.
Initialization: Set $t_0 = 0$.
Observer: For $i = 0, 1, 2, \cdots$
At time $t_i$, $t_{i+1} = t_i + \delta$.
At time $t_{i+1}$, $w_{i+1} \in \mathbb{R}^n$ (an improved estimate of $x(t_{i+1} - T)$) is calculated to satisfy
    $V_E(w_{i+1}; t_{i+1} - T, t_{i+1}) \leq \beta V_E(w_i; t_i - T, t_i)$    (4.7)
(the point $w_i' = x^u(t_{i+1} - T; w_i, t_i - T)$ is used as an initial point for this calculation).
At any time $t \in [t_i, t_{i+1})$, the estimate of the state $x(t)$ is $\hat{x}(t) = x^u(t; w_i, t_i - T)$. In particular $\hat{x}_{i+1} \triangleq \hat{x}(t_{i+1}) = x^u(t_{i+1}; w_{t+1}, t_{i+1} - T)$.
```

disturbance $d_s(t) \in \mathbb{R}^{p_c} \times \{0,1\}^{p_\ell}$, system noise constraint $\mathbb{W} \subset \mathbb{R}^{n_c} \times \{0,1\}^{n_\ell}$ is a bounded polyhedron containing the origin and f_i and g_i are constant vectors of appropriate dimension. The system noise d_a and sensor output disturbance d_s are assumed unmeasured and not that these disturbances can occur in the logic states and outputs respectively.

Remark 4.1. *Note that the form in* (4.8) *assumes that the switching between the different affine models is driven by the partition $\mathcal{X}_i$ which is assumed to be known. If all logical states and outputs are removed, the system in* (4.8) *is a switched system with a known and predefined switching rule. So in the problem considered in [48], the correct state estimate satisfying* (4.8) *uniquely describes a switching sequence (when no disturbances are present).*

The moving horizon observer structure depends on the cost functional given by

$$J(\tau, t, d_a, d_s, x(\tau), \Gamma_\tau) \triangleq \sum_{k=\tau}^{t-1} \|d_s(k)\|_R^2 + \|d_a(k)\|_Q^2 + \Gamma_\tau(x(\tau)) \tag{4.9}$$

where $\tau, t \in \mathbb{N}$, $\tau < t$, Γ_τ is a continuous function and Q and R are positive–definite matrices of suitable dimension. The function Γ_τ represents an initial penalty or arrival cost for a state estimate $x(\tau)$. When the problem is formulated as a fixed horizon optimization problem, the arrival cost Γ_τ is intended to capture all data preceding the fixed horizon into a simple continuous function. In the linear unconstrained case, this Γ_τ can be calculated with the Kalman filter covariance update recursion, but in general the penalty function will be challenging to construct. At a time t with fixed horizon T, the optimization problem to be solved is given by

$$\min_{x(t-T), d_a} J(t-T, t, d_a, d_s, x(t-T), \Gamma_{t-T}), \text{ subj. to (4.8).} \tag{4.10}$$

The key idea of this observer is to search the space of state estimates $x(t-T)$ which simultaneously reduces the measure of the disturbance estimate $d_a(k)$. Assuming appropriate observability notions for the PWA system guarantees convergence at each step when there is no disturbance. As mentioned in [48], this observer scheme can be used to reconstruct system faults if the faults can be represented in (4.8) using the binary-valued logical states. In conclusion, the moving horizon observer developed in [48] develops an observer scheme for PWA systems with known switching rules robust to system and output disturbance.

Remark 4.2. *For brevity, this review of [48] simplifies several constructions developed therein. In particular, much effort is put forth computational methods for designing bounds on the arrival cost* Γ_{t-T} *improved convergence. See [48] for these additional details.*

4.2 New Results Moving Horizon Observer

This section develops new results for Moving Horizon Observer (MHO) schemes on switched linear time-varying (SLTV) systems given in (1.2). Special cases including time-invariant subsystems and the presence or absence of the continuous input will be divided into several subsections. We begin with the simplest case of time-invariant subsystems without input.

4.2.1 Time-Invariant Switched MHO (SMHO)

In this subsection, we consider SLTI systems without input which have the form

$$\dot{x} = A_{v(t)}x(t) \tag{4.11a}$$

$$y = C_{v(t)}x(t), \tag{4.11b}$$

where the system matrices are the same dimension as the counterparts in (1.2). The goal of this subsection is to construct a SMHO for the SLTI system which will extend results in [22] to the switched case. A secondary objective is to use this simpler example to demonstrate issues which must be addressed for the general SLTV observer. The first assumption is the necessary and sufficient condition for SMS observability of SLTI systems from [6]

Assumption 4.2. *For each pair of modes $i, j \in S_M$,*

$$\operatorname{rank}\left(\begin{bmatrix}\mathcal{O}_{2n}(i) & \mathcal{O}_{2n}(j)\end{bmatrix}\right) \triangleq \operatorname{rank}\left(\begin{bmatrix} C_i & C_j \\ C_iA_i & C_jA_j \\ \vdots & \vdots \\ C_iA_i^{2n-1} & C_jA_j^{2n-1}\end{bmatrix}\right) = 2n. \tag{4.12}$$

For the SLTI observer problem without a continuous input, one can see that if the initial condition is zero, $x_0 = 0$, then the corresponding state and output trajectories are given by $x(t) \equiv 0$ and $y(t) \equiv 0$ for all mode sequences. This implies that the mode cannot be reconstructed in the case of a zero initial condition. The following assumption is restrictive, but excludes the zero initial condition case for each moving horizon.

Assumption 4.3. *The continuous state is bounded away from zero for all time, i.e. $\|x(t)\|^2 > \epsilon_x > 0$ for all t. It is assumed that ϵ_x is known.*

Following [22], the following notation represents the cost function over the horizon $t_i, t_i + 1$ when there is assumed to be no switching in the interior of this interval.

$$V_E\left(\hat{x}_{i+1}, \hat{v}_{i+1}; t_i, t_{i+1}\right) \triangleq \int_{t_i}^{t_{i+1}} \|y(\tau) - \hat{y}(\tau)\|^2 \, d\tau \tag{4.13}$$

$$\text{sub.to} : \dot{\hat{x}} = A_{\hat{v}_{i+1}}\hat{x}, \; \hat{x}(t_{i+1}) = \hat{x}_{i+1}$$

$$\hat{y} = C_{\hat{v}_{i+1}}\hat{x}$$

To simplify notation, we let $W_O^i(t_i, t_{i+1})$ denote the observability Gramian of (4.11) in mode i over the interval $[t_i, t_{i+1}]$. The notation $W_O^{i,j}(t_i, t_{i+1})$ denotes the observability Gramian of the extended system for modes i, j, a tuple denoted (A, C), given by

$$\dot{\bar{x}} = A\bar{x}$$

$$\bar{y} = C\bar{x}$$

where

$$A = \begin{bmatrix} A_i & 0 \\ 0 & A_j \end{bmatrix}, \quad C = \begin{bmatrix} C_i & -C_j \end{bmatrix}.$$

Given these Gramian notations, two important quantities γ_1 and γ_2 can be defined.

$$\gamma_1(i+1) = \operatorname*{argmin}_{i \in S_M} \lambda_{min}\left(W_O^i(t_i, t_{i+1})\right) \tag{4.14}$$

$$\gamma_2(i+1) = \operatorname*{argmin}_{i,j \in S_M, i \neq j} \lambda_{min}\left(W_O^{i,j}(t_i, t_{i+1})\right) \tag{4.15}$$

The MHO algorithm is described in Algorithm 2. The key modification of the MHO algorithm in [22] is in step 4 where the cost function is required to be small enough to guarantee accurate mode reconstruction. Once the correct mode is guaranteed, the convergence of the algorithm is similar to [22]. Exponential convergence of Algorithm 2 is proven in Theorem 4.1.

Algorithm 2 MHO for SLTI Systems with Nonzero State

```
1: init: Set $t_0 = 0$.
2: observer: For $i = 0, 1, 2, \dots$
3: At $t_i$, $t_{i+1}$ is the next time in the sequence $\{t_0, t_1, \cdots\}$ which contains all the
   switching times.
4: At time $t_{i+1}$, $\hat{x}_{i+1}$ and $\hat{v}_{k+1}$ are calculated to satisfy
       $V_E(\hat{x}_{i+1}, \hat{v}_{i+1}; t_i, t_{i+1}) \leq \min(\gamma_2(i+1)\epsilon_x, \beta V_E(\hat{x}_i, \hat{v}_i; t_{i-1}, t_i))$
   where $V_E$ and $\gamma_2$ are defined in (4.13) and (4.15), resp., and $\beta \in (0,1)$.
5: At any time $t \in [t_i, t_{i+1})$ the estimates of the state and mode are $\hat{x}(t) = x(t; \hat{x}_i, t_i)$
   and $\hat{v}(t) = v_i$.
```

Theorem 4.1. *Given Assumptions 1.2, 2.1, 4.2, and 4.3, Algorithm 2 using the set* $\{t_1, t_2, \cdots\}$ *converges exponentially at the discrete sample points, i.e.* $\exists M \in (0, \infty)$ *such that*

$$\|x(t_i) - \hat{x}(t_i)\| \leq M e^{-\zeta i} \|x_0 - \hat{x}_0\| \tag{4.16}$$

where $\zeta = -0.5\ln(\beta)$*, and* $\hat{v}(t) = v(t)$ *for all* t*.*

Proof. Assumption 4.2 guarantees observability of the state and mode. Since switching times occur at $\{t_i\}$ which are separated by the minimum dwell time from Assumption 1.2 we have

$$\begin{aligned}
V_E(\hat{x}_{i+1}, \hat{v}_{i+1}, t_i, t_{i+1}) &= \int_{t_i}^{t_{i+1}} \|y(\tau) - \hat{y}(\tau)\| \, d\tau \\
&= \begin{bmatrix} x(t_{i+1}) \\ -\hat{x}_{i+1} \end{bmatrix}^T W_O^{v(t_{i+1}),\hat{v}_{i+1}}(t_i, t_{i+1}) \begin{bmatrix} x(t_{i+1}) \\ -\hat{x}_{i+1} \end{bmatrix} \\
&\geq \begin{cases} \lambda_{min}(W_O^{v(t_{i+1}),v_{i+1}}(t_i, t_{i+1})) \left\| \begin{bmatrix} x(t_{i+1}) \\ \hat{x}_{i+1} \end{bmatrix} \right\|^2, & \text{if } v(t_{i+1}) \neq v_{i+1} \\ \lambda_{min}(W_O^{v_{i+1}}(t_i, t_{i+1})) \left\| x(t_{i+1}) - \hat{x}_{i+1} \right\|^2, & \text{if } v(t_{i+1}) = v_{i+1} \end{cases} \\
&\geq \begin{cases} \gamma_2(i+1)\epsilon_x, & \text{if } v(t_{i+1}) \neq v_{i+1} \\ \gamma_1(i+1) \left\| x(t_{i+1}) - \hat{x}_{i+1} \right\|^2, & \text{if } v(t_{i+1}) = v_{i+1} \end{cases}
\end{aligned} \tag{4.17}$$

Since $V_E(\hat{x}_{i+1}, \hat{v}_{i+1}, t_i, t_{i+1}) < \gamma_2(i+1)\epsilon_x$, step 4 in Algorithm 2 combined with the first case (4.17) implies that $v(t_{i+1}) = v_{i+1}$ for each time t_{i+1}. Further, since $0 < \beta < 1$,

$$V_E(\hat{x}_{i+1}, \hat{v}_{i+1}, t_i, t_{i+1}) \to 0, \quad \text{as } i \to \infty. \tag{4.18}$$

Using the definition in (4.14) we now have that

$$\gamma_1 \left\| x_{i+1} - \hat{x}_{i+1} \right\|^2 \leq V_E(\hat{x}_{i+1}, \hat{v}_{i+1}, t_i, t_{i+1}). \tag{4.19}$$

This implies that $\|x_{i+1} - \hat{x}_{i+1}\| \to 0$ as $i \to \infty$ which establishes global convergence.

Since $V_E(\hat{x}_0, \hat{v}_0; t_{-1}, t_0)$ is finite, $\exists M_1 \in (0, \infty)$ such that

$$V_E(\hat{x}_0, \hat{v}_0, t_{-1}, t_0) \leq M_1 \left\| x_0 - \hat{x}_0 \right\|^2. \tag{4.20}$$

Equations (4.19) and (4.20) yield

$$\|x_{i+1} - \hat{x}_{i+1}\| \leq \beta^{i/2} M \|x_0 - \hat{x}_0\| \leq M e^{-\eta i} \|x_0 - \hat{x}_0\|$$

where $M = (M_1)^{1.5}\gamma_1^{-0.5}$, and $\eta = -0.5\ln(\beta) \in (0, \infty)$. □

4.2.2 Embedded LTI without Input

Assumption 4.2 guarantees observability of the SLTI system without input, but this does not guarantee that the embedded system problem is solvable. To simplify the problem, we consider the time-invariant switched linear system with two modes where the switching times are known (Assumption 2.1). In this case the embedded system is given by

$$\begin{aligned} \dot{x}_e &= ((1-v_e)A_0 + v_e A_1)\, x_e && (4.21\text{a}) \\ &\triangleq A(v_e)x_e(t) \\ y_e &= ((1-v_e)C_0 + v_e C_1)x_e && (4.21\text{b}) \\ &\triangleq C(v_e)x_e(t) \end{aligned}$$

Consider two embedded mode values v_1 and v_2, i.e. $v_1, v_2 \in [0,1]$. If SMS (x_1, v_1) and (x_2, v_2) are indistinguishable for the embedded system (4.21), then one can show that this implies that

$$\begin{aligned} 2n > \operatorname{rank}&\left(\begin{bmatrix} C(v_1) & C(v_2) \\ C(v_1)A(v_1) & C(v_2)A(v_2) \\ \vdots & \vdots \\ C(v_1)A^{2n-1}(v_1) & C(v_2)A^{2n-1}(v_2) \end{bmatrix}\right) \qquad (4.22) \\ \triangleq \operatorname{rank}&\left(\begin{bmatrix} \mathcal{O}_{2n}(v_1) & \mathcal{O}_{2n}(v_2) \end{bmatrix}\right), \end{aligned}$$

where $A(v)$ and $C(v)$ are defined in (4.21). The embedded MHO (EMHO) problem is solvable if (4.22) is not satisfied for all $v_1 \in \{0,1\}$ and all $v_2 \in [0,1]\backslash v_1$. If this is not immediately apparent, recall that the inequality in (4.22) being satisfied implies that two linear systems $(A(v_1), C(v_1))$ and $(A(v_2), C(v_2))$ are not always distinguishable. The EMHO searches for an optimal state and mode estimate over the larger space where the mode v_e can take values between 0 and 1. If an embedded value produces the minimum cost then this implies that an embedded mode value is indistinguishable from the switched mode value. Unfortunately, Assumption 4.2 is not sufficient to guarantee indistinguishability of all embedded mode values. However, two properties will be proven about the embedded search space which will allow for EMHO convergence. First, the set of points (x_2, v_2) indistinguishable from (x_1, v_1) with $x_1 \neq 0$ is a set of codimension 2. Secondly, for a fixed (x_1, v_1) with $x_1 \neq 0$, the space of embedded mode and state values which are distinguishable from (x_1, v_1) is path connected. This implies that there always exists a path for the EMHO algorithm to reach the optimal solution.

The following three definitions come from [49, pg. 205]. The first two definition lead to the definition of covering dimension which is used to prove codimension 2.

Definition 4.1. *[49] A collection $\mathcal{A}$ of subsets of a space X is said to have order $m+1$ if some point of X lies in $m+1$ elements of $\mathcal{A}$, and no point of X lies in more that $m+1$ elements of $\mathcal{A}$.*

Definition 4.2. *[49] A space X has topological dimension m if m is the smallest integer such that for every open covering $\mathcal{A}$ of X, there is an open covering $\mathcal{A}'$ of X which refines $\mathcal{A}$ and has order at most $m+1$.*

Covering dimension provides a topological metric to give some handle on relative size of sets. For example, in a 2-dimensional plane, a line segment has codimension 1 and a point has codimension 2. Another concept related to codimension is path connected spaces. In the 2-dimensional plane, a set of line segments can cause some portion of the space to not be path connected, but no finite collection of points can cause the space not to be path connected. The formal definition is given below for reference.

Definition 4.3. *[49, pg. 155] Given points x and y of the space X, a path in X from x to y is a continuous map $f : [a, b] \to X$ of some closed interval in the real line into X, such that $f(a) = x$ and $f(b) = y$. A space X is path connected if every pair of points of X can be joined by a path in X.*

The following lemma will be used in proving the desired results.

Lemma 4.2. *Suppose Assumption 4.2 is satisfied for the SLTI system* (4.11). *Let* $(x_1, v_1) \in \mathbb{R}^n \times \{0, 1\}$ *be fixed with* $x_1 \neq 0$. *Let* L *be the set of all tuples* $(x_2, v_2) \in \mathcal{X} \triangleq \mathbb{R}^n \times (0, 1)$ *such that*

$$0 = \begin{bmatrix} \mathcal{O}_{2n}(v_1) & \mathcal{O}_{2n}(v_2) \end{bmatrix} \begin{bmatrix} x_1 \\ -x_2 \end{bmatrix} \triangleq M(v_2) \begin{bmatrix} x_1 \\ -x_2 \end{bmatrix}. \tag{4.23}$$

Let the projection map $\pi_1 : \mathcal{X} \to [0, 1]$ *be defined as* $\pi_1((x, v)) = v$ *for* $(x, v) \in \mathcal{X}$. *Then* $\pi_1(L) = \{w_1, \ldots, w_k\}$ *for some* $0 \leq k \in \mathbb{N}$, *i.e.* $\pi_1(L)$ *is finite.*

Proof. Without loss of generality, let $v_1 = 0$. By Assumption 4.2, $M(1)$ has full column rank. Defining $\gamma(v_2) = \det(M^\top(v_2)M(v_2))$, this implies that $\gamma(1) \neq 0$. Since γ is a nonzero finite-degree polynomial in v_2, there are at most $k \in \mathbb{N}^+$ distinct values $p_1, \ldots, p_l \in [0, 1]$ such that $M(p_i)$ drops rank. For a point (x_2, v_2) to be in L, the vector $[x_1^\top, -x_2^\top]^\top$ is in the null space of $M(v_2)$. Since $x_1 \neq 0$, this occurs only if $M(v_2)$ is not full rank. Thus $\pi_1(L) \subset \{p_1, \ldots, p_l\}$, thus $\pi_1(L)$ is finite as desired. □

Theorem 4.3. *Suppose the conditions in Lemma 4.2 are satisfied. Then* $\mathcal{X} = [0, 1] \times \mathbb{R}^n$ *with the subspace topology in* $\mathbb{R}^{2n+1}$ *(with the product topology) has Lebesgue covering dimension* $n + 1$ *and* L *has codimension* $\delta \geq 2$ *in* $\mathcal{X}$.

Proof. First, recall that the Lebesgue covering dimension of $\mathbb{R}^n$ is n. Since $[0, 1] \subset \mathbb{R}$ and we are considering the subspace topology on $[0, 1]$, open sets in $[0, 1]$ have the form $[0, a)$, $(b, 1]$, (a, b), and $[0, 1]$ for $a, b \in [0, 1]$. Since $\mathcal{A} = \{[0, 1), (0, 1]\}$ covers $[0, 1]$ and every refinement $\mathcal{A}'$ of $\mathcal{A}$ has some point $a \in [0, 1]$ in at least two elements of $\mathcal{B}$. This implies that the dimension of $[0, 1]$ is at least 1. Since $[0, 1] \subset \mathbb{R}$ and $\mathbb{R}$ has dimension 1, the dimension of $[0, 1]$ must be 1. Thus $\mathcal{X}$ has covering dimension $\dim([0, 1]) + \dim(\mathbb{R}^n) = n + 1$.

From Lemma 4.2, L is a finite union of sets $L_i \triangleq \{(v, x) \in L | v = w_i\}$ for $i = 1, \ldots, k$ constructed using the projection $\pi_1(L) = \{w_1, \ldots, w_k\}$. Further, each L_i is closed in L with the subspace topology (each L_i is actually both open and closed in L since it is a finite disjoint union). Thus from [49, Cor. 50.3]

$$\dim L = \max\{\dim L_1, \ldots, \dim L_k\}.$$

Thus proving that each L_i has dimension at most $n-1$ will complete the proof. To this end, let $\pi_x : \mathcal{X} \to \mathbb{R}^n$ be the projection map such that for $(v,x) \in \mathcal{X}$, $\pi_x((v,x)) = x$. Since $\pi_1(L_i) = w_i$ is finite, $\dim L_i = \dim \pi_x(L_i)$. Since $\pi_x(L_i)$ is a subset of $\mathbb{R}^n$, $\pi_x(L_i)$ has dimension at most n.

For each pair $(w_i, x_2) \in L_i$,

$$\mathcal{O}_{2n}(w_i)x_2 = \mathcal{O}_{2n}(v_1)x_1 \neq 0,$$

since $\operatorname{rank}(\mathcal{O}_{2n}(v_1)) = n$ for Assumption 4.2 to be satisfied and $x_1 \neq 0$. Let $y_1, \ldots, y_s$ form a basis for the range of $\mathcal{O}_{2n}(w_i)$. Extend this to a basis for $\mathbb{R}^n$ by adding $y_{s+1}, \ldots, y_n$ which span the null space of $\mathcal{O}_{2n}(w_i)$. Since $\mathcal{O}_{2n}(v_1)x_1 \neq 0$, there exists unique scalars $\alpha_1, \ldots, \alpha_s$ not all of which are zero such that

$$\sum_{j=1,\ldots,s} \alpha_j y_j = \mathcal{O}_{2n}(v_1)x_1.$$

Since these scalars are unique, $\pi_x(L_i)$ has codimension at least s in $\mathbb{R}^n$ (codimension both topologically and with respect to dimension of linear subspaces). Note that $s \geq 1$ because $\mathcal{O}_{2n}(w_i)$ has a range space. Thus $\dim L_i = \dim \pi_x(L_i)$ has dimension at most $n-1$, completing the proof. □

Theorem 4.4. *Suppose Assumption 4.2 is satisfied for the SLTI system* (4.11)*. Let* $(v_1, x_1) \in \{0,1\} \times \mathbb{R}^n$ *be fixed with* $x_1 \neq 0$*. Define* $L \subset \mathcal{X} \triangleq [0,1] \times \mathbb{R}^n$ *to be the points* (v_2, x_2) *such that*

$$M(v_2)\begin{bmatrix} x_1 \\ -x_2 \end{bmatrix} \triangleq \begin{bmatrix} \mathcal{O}_{2n}(v_1) & \mathcal{O}_{2n}(v_2) \end{bmatrix}\begin{bmatrix} x_1 \\ -x_2 \end{bmatrix} = 0. \tag{4.24}$$

Then $\mathcal{X} \backslash L$ *is path connected.*

Proof. Let (v_2, x_2) and (v_2', x_2') be two distinct points in $\mathcal{X}\backslash L$. Let $f : [0,1] \to \mathcal{X}$ be the continuous function in $\mathcal{X}$ given by

$$f(t) = (1-t)(v_2, x_2) + t(v_2', x_2').$$

Note that $f(t) \in \mathcal{X}$ for all $t \in [0,1]$ since $f(\{0,1\}) \subset \mathcal{X}$ and $\mathcal{X}$ is convex. If the range of f is in $\mathcal{X} \backslash L$ then f is the desired path connecting (v_2, x_2) and (v_2', x_2'). If this is not satisfied, two cases arise: when $v_2 = v_2'$ and $v_2 \neq v_2'$.

When $v_2 \neq v_2'$, let $v_2 < v_2'$ without loss of generality. From Lemma 4.2, there are only a finite number of distinct values for the first coordinate of points in L, i.e. $\pi_1(L) = \{w_1, \ldots, w_k\}$. For all potentially problematic embedded mode values $w_1, \ldots, w_l \in \pi_1(L)$ between v_2 and v_2', let $t_i \in [0,1]$ be the number such that

$$w_i = (1 - t_i)v_2 + t_i v_2'.$$

It is only at these points t_i such that the function f can enter L, i.e. $f(t) \in L \implies t \in \{t_1, \ldots, t_l\}$. The desired continuous function will be constructed by adjusting f in an interval $(t_i - \epsilon, t_i + \epsilon)$ around t_i to guarantee that $f(t_i)$ is not in L.

Let $\epsilon > 0$ be a number smaller than half the distance between two distinct points t_i and t_j for $i, j \in \{1, \ldots, l\}$ and between each t_i and the end points v_2 and v_2', i.e. defining $t_0 \triangleq v_2$ and $t_{l+1} \triangleq v_2'$ for convenience, ϵ satisfies

$$0 < \epsilon < \min_{i \neq j \in \{0, \ldots, l+1\}} \frac{|t_i - t_j|}{2}. \tag{4.25}$$

For each t_i where $i \in \{1, \ldots, l+1\}$, there exists nonzero vectors $z_i \in \mathbb{R}^n$ in the range space of $\mathcal{O}_{2n}(w_i)$ because $\mathcal{O}_{2n}(v_1)x_1$ is a nonzero and in the range of $\mathcal{O}_{2n}(w_i)$. If $f(t_i) \in L$, then adding $(0, z_i)$ to $f(t_i)$ is not in L because

$$\begin{aligned} \mathcal{O}_{2n}(w_i)(\pi_x(f(t_i)) + z_i) &= \mathcal{O}_{2n}(v_1)x_1 + \mathcal{O}_{2n}(w_i)z_i \\ &\neq \mathcal{O}_{2n}(v_1)x_1. \end{aligned}$$

If $f(t_i) \notin L$, define $z_i = 0$ and then $f(t_i) + (0, z_i) \notin L$ for both cases, when $f(t_i) \in L$ and $f(t_i) \notin L$. Using this notation, the necessary modifications of f can be written in a general fashion. The desired continuous map $g : [0,1] \to \mathcal{X}$ can be constructed as follows

$$g(t) = \begin{cases} \frac{t_i - t}{\epsilon} f(t_i - \epsilon) + \frac{t - t_i + \epsilon}{\epsilon}(f(t_i) + (0, z_i)), & \text{if } t \in [t_i - \epsilon, t_i) \\ \frac{t - t_i + \epsilon}{\epsilon}(f(t_i) + (0, z_i)) + \frac{t - t_i}{\epsilon} f(t_i + \epsilon), & \text{if } t \in [t_i, t_i + \epsilon) \\ f(t), & \text{Otherwise.} \end{cases}$$

Continuity of $g(t)$ is can be verified by inspection since the selection of ϵ implies that the intervals $[t_i - \epsilon, t_i + \epsilon]$ are pairwise disjoint. The range of g is in $\mathcal{X}\backslash L$ since $f(t_i) + (0, z_i) \notin L$ by construction of z_i and since only at points t_i for $i = 1, \ldots l$ can $f(t)$ be in L.

When $v_2 = v_2'$, choose $v_3 \in [0, 1]$ such that there is no element $w_i \in \pi_1(L)$ in the interval $(v_2, v_3]$. This point v_3 exists since $\pi_1(L)$ is finite from Lemma 4.2. The desired continuous function will be the composition of two functions $g_1, g_2 : [0, 1] \to \mathcal{X}$. The first function g_1 moves along the line between (v_2, x_2) and (v_3, x_2), i.e. only changing the first coordinate. The second line moves from (v_3, x_2) to the desired endpoint (v_2', x_2'). The functions are defined as follows:

$$g_1(t) = (1 - t)(v_2, x_2) + t(v_3, x_2)$$
$$g_2(t) = (1 - t)(v_3, x_2) + t(v_2', x_2')$$

Then the composition function $h : [0, 1] \to \mathcal{X}$ given by

$$h(t) = \begin{cases} g_1(2t), & \text{for } t \in [0, 0.5) \\ g_2(2t - 1)), & \text{for } t \in [0.5, 1] \end{cases}$$

is a path in $\mathcal{X}\backslash L$ connecting (x_2, v_2) and (x_2', v_2'). Note that $h([0, 1]) \notin L$ since $\pi_1\big(h([0, 1])\big) \subset \{v_2\}$, i.e. only at the starting and ending point of the path h can $h(t)$ enter L; however, since the end points (v_2, x_2) and (v_2', x_2') are not in L by assumption, the desired result follows. □

4.2.3 Embedded LTV without Input

The convergence of the EMHO for switched linear time-varying (SLTV) systems can be approached with the same techniques as the SMHO. However, this approach requires the computation of the extended observability Gramian for each pair embedded mode and switched mode values. This is intractable because there are an infinite number of Gramians which would need to be calculated. Another approach is to consider a classical LTV observability result which uses the time-varying observability matrix listed below for reference.

Proposition 4.5. *[13] The pair $(C(t), A(t))$ is observable (in the classical sense) over $[t_0, t_1]$ if there is some positive integer q and some point $t' \in [t_0, t_1]$ such that*

$$n = \operatorname{rank} \begin{bmatrix} C(t')\Phi(t', t_0) \\ D[C(t)\Phi(t, t_0)]|_{t=t'} \\ \vdots \\ D^q[C(t)\Phi(t, t_0)]|_{t=t'} \end{bmatrix} \tag{4.26}$$

where $D^q = d^k/dt^k$ is the derivative operator and $C(t)$ and $\Phi(t', t_0)$ are q times differentiable.

Remark 4.3. *The condition is satisfied if are functionally independent. The condition in (4.26) guarantees functional independence of the columns of $C(t)\Phi(t', t_0)$ when the $(C(t), A(t))$ matrices are smooth.*

The extension to the SLTV system is immediate. For two modes 0 and 1 the extended system $(C(t), A(t))$ is given by

$$A(t) \triangleq \begin{bmatrix} A_0(t) & 0 \\ 0 & A_1(t) \end{bmatrix} \tag{4.27a}$$

$$C(t) \triangleq \begin{bmatrix} C_0(t) & C_1(t) \end{bmatrix}. \tag{4.27b}$$

For notation, let

$$R(t, t_0, v) = ((1 - v)C_0(t) + vC_1(t))\Phi_v(t, t_0), \tag{4.28}$$

where $\Phi_v(t, t_0)$ is the state transition matrix for the system $\dot{x} = ((1 - v)A_0(t) + vA_1(t))x$. In addition, for any $q \geq 0$ let

$$\mathcal{R}_q(s, t_0, v) = \begin{bmatrix} R(s, t_0, v) \\ D[R(t, t_0, v)]|_{t=s} \\ \vdots \\ D^q[R(t, t_0, v)]|_{t=s} \end{bmatrix}. \tag{4.29}$$

If between each potential switching time in Assumption 2.1, there exists $t' \in [t_0, t_1]$ and integer q such that

$$\operatorname{rank}\begin{bmatrix}\mathcal{R}_q(t', t_0, 0) & \mathcal{R}_q(t', t_0, 1)\end{bmatrix} = 2n \tag{4.30}$$

then the pair of modes are SMS observable as guaranteed by Proposition 4.5 for the extended system. The following theorem extends results about the time-invariant EMHO search space in Theorems 4.3 and 4.4 to the time-varying case.

Theorem 4.6. *Suppose Assumption 2.1 holds for the SLTV system* (4.27) *and* (4.30) *is satisfied at a time t between each switching time. Fix $(x_1, v_1) \in \mathbb{R}^n \times \{0, 1\}$ with $x_1 \neq 0$. Further we assume that* $\operatorname{rank}\begin{bmatrix}\mathcal{R}_q(t_0, t_0, 0) & \mathcal{R}_q(t_0, t_0, 1)\end{bmatrix} = 2n$ *for some integer q. Let $L \subset \mathcal{X} \triangleq \mathbb{R}^n \times [0, 1]$ be defined as the set of points $(x_2, v_2) \in L$ which satisfy*

$$0 = \begin{bmatrix}\mathcal{R}_q(t_0, t_0, v_1) & \mathcal{R}_q(t_0, t_0, v_2)\end{bmatrix}\begin{bmatrix} x_1 \\ -x_2 \end{bmatrix}. \tag{4.31}$$

Then L (with the subspace topology) has codimension $\delta \geq 2$ in $\mathcal{X}$. In addition, the space $X \backslash L$ is path connected.

Proof. The proof follows the arguments in Theorem 4.3 and Theorem 4.4 replacing $\begin{bmatrix}\mathcal{O}_{2n}(v_1) & \mathcal{O}_{2n}(v_2)\end{bmatrix}$ with $\begin{bmatrix}\mathcal{R}_q(t_0, t_0, v_1) & \mathcal{R}_q(t_0, t_0, v_2)\end{bmatrix}$. □

4.2.4 Embedded LTI with Input

In [50], observability for SLTI systems for almost every input was reduced to the difference in Toeplitz matrices being nonzero. That is for modes 0 and 1,

$$0 \neq \Gamma_{2n}(0) - \Gamma_{2n}(1) \triangleq \begin{bmatrix} 0 & \cdots & 0 & 0 \\ C_0B_0 & \cdots & 0 & 0 \\ C_0A_0B_0 & \cdots & \vdots & \vdots \\ \vdots & \cdots & 0 & 0 \\ C_0A_0^{2n-1}B_0 & \cdots & C_0B_0 & 0 \end{bmatrix} - \begin{bmatrix} 0 & \cdots & 0 & 0 \\ C_1B_1 & \cdots & 0 & 0 \\ C_1A_1B_1 & \cdots & \vdots & \vdots \\ \vdots & \cdots & 0 & 0 \\ C_1A_1^{2n-1}B_1 & \cdots & C_1B_1 & 0 \end{bmatrix} \tag{4.32}$$

Let $u(t)$ and its first $q-1$ derivatives be denoted

$$\mathcal{U}_q(t) = \begin{bmatrix} u(t) \\ \dot{u}(t) \\ \vdots \\ u^{(q-1)}(t) \end{bmatrix}. \tag{4.33}$$

The condition in [50] guarantees that the input $u(\cdot)$ which distinguishes all initial conditions $x_0, x_1 \in \mathbb{R}^n$ satisfies the following equation for some integer $0 < q \leq 4n+1$ satisfies

$$0 \neq \begin{bmatrix} \mathcal{O}_q(0) & \mathcal{O}_q(1) & (\Gamma_{q-1}(0) - \Gamma_{q-1}(1))\,\mathcal{U}_q(t) \end{bmatrix} \begin{bmatrix} x_0 \\ -x_1 \\ 1 \end{bmatrix}. \tag{4.34}$$

The bound of $q \leq 4n+1$ comes from the observation that the structure of $\Gamma_{2n}(0) - \Gamma_{2n}(1) \neq 0$ implies that $\Gamma_{4n}(0) - \Gamma_{4n}(1)$ has a rank lower bounded by $2n+1$. Since $\begin{bmatrix} \mathcal{O}_q(0) & \mathcal{O}_q(1) \end{bmatrix}$ has $2n$ columns, its rank is bounded by $2n$. So at $q = 4n$ the input and its derivatives $\mathcal{U}_{4n}$ can enter the output difference in a way outside the range space of $\begin{bmatrix} \mathcal{O}_q(0) & \mathcal{O}_q(1) \end{bmatrix}$, i.e. forcing distinguishability for all initial conditions. So in this subsection we will consider the performance of the EMHO when an input satisfies (4.34). With this input distinguishing all initial conditions, we will be able to establish a new path connected result for the search space of EMHO.

Theorem 4.7. *Let Assumption 4.2 hold for the two-mode SLTI system in* (4.11) *with a fixed $x_1 \in \mathbb{R}^n$, $v_1 \in \{0,1\}$, and $u(\cdot) \in C^\infty$ such that there exists $q \geq 2n+1$ satisfying*

$$\operatorname{rank} \begin{bmatrix} \mathcal{O}_q(0) & \mathcal{O}_q(1) & (\Gamma_{q-1}(0) - \Gamma_{q-1}(1))\mathcal{U}_{q-1}(t_0) \end{bmatrix} = 2n+1, \tag{4.35}$$

and $\mathcal{O}_q(v_1)x_1 + \Gamma_{q-1}(v_1)\mathcal{U}_{q-1}(t_0) \neq 0$. Let $L \subset \mathbb{R}^n \times [0,1] \triangleq \mathcal{X}$ denote all pairs $(x_2, v_2) \in L$ such that

$$M(v_2) \begin{bmatrix} x_1 \\ -x_2 \\ 1 \end{bmatrix} \triangleq \begin{bmatrix} \mathcal{O}_q(v_1) & \mathcal{O}_q(v_2) & (\Gamma_{q-1}(v_1) - \Gamma_{q-1}(v_2))\mathcal{U}_{q-1} \end{bmatrix} \begin{bmatrix} x_1 \\ -x_2 \\ 1 \end{bmatrix} = 0. \tag{4.36}$$

Then the space $\mathcal{X} \setminus L$ is path connected and L has codimension at least 2.

Proof. Let $\pi_v : \mathcal{X} \mapsto [0,1]$ denote the projection such that for all $(x,v) \in X$, $\pi_v((x,v)) = v$. Without loss of generality let $v_1 = 0$. We begin with a claim.

Claim 1: $\pi_v(L)$ is finite.

Let $\rho(v_2) = \det(M(v_2)^\top M(v_2))$. From (4.35) with the assumption $v_1 = 0$ we have that $\operatorname{rank}(M(1)) = 2n+1$, i.e. full column rank. Thus $\rho(1) \neq 0$ and $\rho(v_2)$ is a finite degree polynomial in v_2 implying $\rho(v_2)$ has at most finite roots in $[0,1]$. From (4.36), $(x_2, v_2) \in L$ only if $\operatorname{rank}(M(v_2)) < 2n+1$ which occurs exactly when $\rho(v_2) = 0$. Thus $\pi_v(L)$ is finite competing Claim 1.

Let (x_2, v_2), $(x_2', v_2') \in L$ and let $f : [0,1] \mapsto \mathcal{X}$ be given by

$$f(s) = (1-s)(x_2, v_2) + s(x_2', v_2'). \tag{4.37}$$

If $f([0,1]) \subset \mathcal{X} \setminus L$, we have the desired path. If not, consider first the case $v_2 \neq v_2'$. In this case $\pi_v(f([0,1]))$ intersects L for at most a finite number points as per the preceding claim. Let $\{s_i\}_{i=1}^k \in [0,1]$ with $s_i < s_{i+1}$ denote the finite points such that $f(s_i) \in L$. This leads to the next claim.

Claim 2: For each s_i, $i = 1, \ldots, k$, there exists $z_i \in \mathbb{R}^n$ such that $f(s_i) + (z_i, 0) \notin L$.

Let $f(s_i) = (x_{s_i}, v_{s_1})$. Note that if $\mathcal{O}_q(v_{s_i}) = 0$, then $\Gamma_{q-1}(v_{s_i}) = 0$ as well, but this would imply

$$M(v_{s_i}) \begin{bmatrix} x_1 \\ -x_{s_i} \\ 1 \end{bmatrix} = \mathcal{O}_q(0)x_1 + \Gamma_{q-1}(0)\mathcal{U}_{q-1}(t_0) \neq 0$$

by assumption which contradicts that $f(s_i) \in L$. Thus $\mathcal{O}_q(v_{s_i}) \neq 0$ for each i. Let $z_i \in \mathbb{R}^n$ such that $\mathcal{O}_q(v_{s_i})z_i \neq 0$. Then from (4.36), since $f(s_i) \in L$

$$M(v_{s_i}) \begin{bmatrix} x_1 \\ -(x_{s_i} + z_i) \\ 1 \end{bmatrix} = -\mathcal{O}_q(v_{s_i})z_i \neq 0,$$

hence $f(s_i) + (z_i, 0) \notin L$ as claimed.

To construct disjoint intervals between each s_i and s_{i+1} we define

$$s_{min} = \min\left\{\min_{j=1,\ldots,k-1} \frac{s_{j+1}-s_j}{2}, \frac{s_1}{2}, \frac{1-s_k}{2}\right\}.$$

By construction of s_{min}, $f(s) \notin L$ for all $s \in [s_i - s_{min}, s_i) \cup (s_i, s_i + s_{min}]$. For this reason we can modify f in these intervals to pass through $f(s_i) + (z_i, 0)$ for each s_i which will give the desired result as per the following function.

$$\tilde{f}(s) = \begin{cases} \left(\frac{-s+s_i}{s_{min}}\right) f(s_i - s_{min}) + \left(\frac{s-s_i+s_{min}}{s_{min}}\right)(f(s_i) + (z_i, 0)) & \text{if } s \in [s_i - s_{min}, s_i] \\ \left(\frac{-s+s_i+s_{min}}{s_{min}}\right)(f(s_i) + (z_i, 0)) + \left(\frac{s-s_i}{s_{min}}\right) f(s_i + s_{min}) & \text{if } s \in [s_i, s_i + s_{min}] \\ f(s) & \text{otherwise.} \end{cases}$$

If $v_2 = v_2' \in \pi_v(L)$, there exists $s^-, s^+ \in [0, 1)$ such that $\max(s^-, s^+) > 0$ and

$$[v_2 - s^-, v_2) \cup (v_2, v_2 + s^+) \subset \pi_v(\mathcal{X} \setminus L)$$

If $s^+ \neq 0$, then the desired path is

$$g(s) = \begin{cases} (1-2s)(x_2, v_2) + 2s(x_2, v_2 + s^+) & \text{if } s \in [0, \frac{1}{2}] \\ 2(1-s)(0, v_2 + s^+) + (2s-1)(x_2', v_2') & \text{if } s \in [\frac{1}{2}, 1]. \end{cases}$$

The case when $s^+ = 0$ and $s^- \neq 0$ is can be constructed replacing $v_2 + s^+$ with $v_2 - s^-$ in the function g above.

The proof that L has codimension at least two follows arguments in Theorem 4.3. A brief sketch of this proof begins with the observation that $\pi_v(L)$ is finite and for each $v_2 \in \pi_v(L)$, the pairs $(x_2, v_2) \in L$ must satisfy

$$\mathcal{O}_q(v_2)x_2 = \mathcal{O}_q(0)x_1 + (\Gamma_{q-1}(0) - \Gamma_{q-1}(v_2))\mathcal{U}_{q-1}(t_0). \tag{4.38}$$

Only if $\mathcal{O}_q(v_2) = 0$, can L have codimension one in $\mathcal{X}$. However, if $\mathcal{O}_q(v_2) = 0$, then $\Gamma_{q-1}(v_2) = 0$ and the right side of (4.38) is nonzero since $\mathcal{O}_q(v_1)x_1 + \Gamma_{q-1}(v_1)\mathcal{U}_{q-1}(t_0) \neq 0$ by assumption. Thus L must have codimension at least two.

□

4.2.5 Embedded LTV with Input

The result in Theorem 4.6 can be extended to the two-mode time-varying case in a straightforward manner. First for an interval $[t_0, t_f]$, as in Theorem 4.6 we assume there exists a $q \geq 0$ such that

$$\operatorname{rank}\begin{bmatrix}\mathcal{R}_q(t_0,t_0,0) & \mathcal{R}_q(t_0,t_0,1)\end{bmatrix} = 2n.$$

The effect of the input $u(\cdot)$ on the output in any embedded mode $v \in [0,1]$ is given by

$$N(t,t_0,v,u) = \int_{t_0}^{t} C(v,t)\Phi_v(t,q)((1-v)B_0(q) + vB_1(q))u(q)dq, \tag{4.39}$$

where $\Phi_v(t,t_0)$ is the state transition matrix for the system $\dot{x} = ((1-v)A_0(t) + vA_1(t))x$. In addition, for any $q \geq 0$ let

$$\mathcal{N}_q(s,t_0,v,u) = \begin{bmatrix} N(s,t_0,v,u) \\ D[N(t,t_0,v,u)]|_{t=s} \\ \vdots \\ D^q[N(t,t_0,v,u)]|_{t=s} \end{bmatrix}. \tag{4.40}$$

The quantity $\mathcal{N}_q(t_0,t_0,v,u)$ reduces to $\Gamma_{q-1}(v)\mathcal{U}_{q-1}(t_0)$ which was used for the SLTI case in Theorem 4.7 and will be used in the subsequent theorem in much the same manner.

Theorem 4.8. *Let Assumption 4.2 hold for the two-mode SLTV system in* (1.2) *with a fixed $x_1 \in \mathbb{R}^n$, $v_1 \in \{0,1\}$, and $u(\cdot) \in C^\infty$ such that there exists $q \geq 2n+1$ satisfying*

$$\operatorname{rank}\begin{bmatrix}\mathcal{R}_q(t_0,t_0,0) & \mathcal{R}_q(t_0,t_0,1) & \mathcal{N}_q(t_0,t_0,0,u) - \mathcal{N}_q(t_0,t_0,1,u)\end{bmatrix} = 2n+1, \tag{4.41}$$

and $\mathcal{R}_q(t_0,t_0,v_1)x_1 + \mathcal{N}_q(t_0,t_0,v_1) \neq 0$. Let $L \subset \mathbb{R}^n \times [0,1] \triangleq \mathcal{X}$ denote all pairs $(x_2,v_2) \in L$ such that

$$
\begin{aligned}
0 &= M(v_2)\begin{bmatrix} x_1 \\ -x_2 \\ 1 \end{bmatrix} \\
&\triangleq \begin{bmatrix} \mathcal{R}_q(t_0,t_0,v_1) & \mathcal{R}_q(t_0,t_0,v_2) & \mathcal{N}_q(t_0,t_0,v_1,u) - \mathcal{N}_q(t_0,t_0,v_2,u) \end{bmatrix} \begin{bmatrix} x_1 \\ -x_2 \\ 1 \end{bmatrix} \qquad (4.42)
\end{aligned}
$$

Then the space $\mathcal{X} \setminus L$ is path connected and L has codimension at least 2.

Proof. The proof follows Theorem 4.7 replacing $\mathcal{O}_q(v)$ with $\mathcal{R}_q(t_0, t_0, v)$ and replacing $\Gamma_{q-1}(v)\mathcal{U}(t_0)$ with $\mathcal{N}_q(t_0, t_0, v, u)$. □

4.2.6 EMHO Convergence

Guaranteeing SMS convergence of the EMHO over an interval (t_0, t_f) requires that the input $u(\cdot)$ distinguishes all modes and each mode is observable over this interval. If this is satisfied, the EMHO can choose the best state and mode estimate matching the measured output over the interval (t_0, t_f). If the mode has an embedded value, one can project and search near the projected mode value. In this way, the EMHO can search through the embedded search space and return the correct state and mode sequence (provided the mode distinguishing input $u(t)$ and observability of each mode). This approach solves the entire interval (t_0, t_f).

As an alternative, we propose solving a smaller problem over subintervals $(t_i - T, t_i) \subset (t_0, t_f)$. Moreover, we reduce the complexity by only requiring that a portion of the problem is solved at each step. By improving each estimate at by a fixed rate, we will guarantee a time $T_{reach} > t_0$ after which the estimate will be correct and the state estimate will converge exponentially pointwise. However, to guarantee convergence in this manner requires additional properties on the distinguishability of the input $u(\cdot)$.

Definition 4.4. *Consider a two mode SLTV system in* (1.2) *and a fixed* $\epsilon > 0$*. An input* $u(\cdot)$ *is an* ϵ*-mode distinguishing input over* $[t_0, t_f]$ *if for all* $\{x_0, v\}, \{\bar{x}_0, \bar{v}\} \in \mathbb{R}^n \times \{0, 1\}$ *with* $v \neq \bar{v}$

$$\int_{t_0}^{t_f} \|y(t) - \bar{y}(t)\|^2 dt \geq \epsilon > 0. \tag{4.43}$$

Remark 4.4. *If the conditions in Theorem 2.21 are satisfied, then almost every input causes mode distinguishability, i.e. for almost every input* $u(\cdot)$ *and each time interval* $[t_0, t_f]$ *there exists an* $\epsilon(t_0, t_f, u(\cdot))$ *such that* (4.43) *is satisfied. Definition 4.4 specifies the degree* (ϵ) *of distinguishability between the modes.*

For the EMHO, there are a sequence of starting points $\{t_i\}$ for the smaller optimization problems. For the convergence guarantee in the following theorem, we require that the input $u(\cdot)$ has a fixed $\epsilon > 0$ such that $u(\cdot)$ is an ϵ-mode distinguishing input over *every* interval $[t_i, t_{t+1}]$. We now introduce the EMHO observer algorithm.

Algorithm 3 EMHO

```
init: Set $t_0 = 0$.
observer: For $i = 0, 1, 2, \ldots$
At $t_i$, $t_{i+1}$ is the next time in the sequence $\{t_0, t_1, \cdots\}$ which contains all the switching times.
At time $t_{i+1}$, $\hat{x}_{i+1}$ and $\hat{v}_{k+1}$ are calculated to satisfy

    $$V_E\left(\hat{x}_{i+1}, \hat{v}_{i+1}; t_i, t_{i+1}\right) \leq \beta V_E(\hat{x}_i, \hat{v}_i; t_{i-1}, t_i)) \tag{4.44}$$

    where $V_E$ is defined previously, and $\beta \in (0, 1)$. The EMHO may search in the embedded space $[0, 1]$ but the final estimate $\hat{v}_{k+1}$ must be in $\{0, 1\}$.
At any time $t \in [t_i, t_{i+1})$ the estimates of the state and mode are $\hat{x}(t) = x(t; \hat{x}_i, t_i)$ and $\hat{v}(t) = v_i$.
```

Theorem 4.9. *Consider a SLTV system where over each interval* $[t_i, t_{i+1}]$ *each subsystem is observable and the input* $u(t)$ *is an* ϵ*-mode distinguishing input, and the* *set* $\{t_i\}$ *contains all switching times. Then there exists a time* $T_{reach}(\hat{x}_0, \hat{v}_0, \beta) \geq t_0$ *after which* $\hat{v}(t) = v(t)$ *and the error* $e(t_i) = x(t_i) - \hat{x}(t_i)$ *converges asymptotically. Moreover, if it is an SLTI system satisfying the above conditions then for* $t_i \geq T_{reach}$ *the error* $e(t_i)$ *converges exponentially pointwise, i.e. there exists* $M \in (0, \infty)$ *such that*

$$\|x(t_i) - \hat{x}(t_i)\| \leq M e^{\zeta i} \|x_0 - \hat{x}_0\|$$

where $\zeta = -0.5 \ln(\beta)$.

Proof. First note that $V_E(t_1) \triangleq V_E(\hat{x}_1, \hat{v}_1; t_0, t_1)$ is finite. Thus (4.44) is a contraction implying that there exists a time $T_{reach} > 0$ such that for all $t_i \geq T_{reach}$, $V_E(t_i) \triangleq V_E(\hat{x}_i, \hat{v}_i; t_{i-1}, t_i) \leq \epsilon$. Since $u(\cdot)$ is an ϵ-mode distinguishing input over $[t_i, t_{i+1}]$ for each i, $\hat{v}_i = v(t_i)$ for all $t_i \geq T_{reach}$. The mode estimate is then correct for all time $t \geq T_{reach}$ since switching times are contained in the set $\{t_i\}$.

Over each interval $[t_i, t_{i+1}]$ where $t_i \geq T_{reach}$ the mode estimate is correct and the active mode is observable. Because the active mode is observable, $V_E(\hat{x}_i, \hat{v}_i; t_{i-1}, t_i) = 0$ only if $\hat{x}_i = x(t_i)$. Thus step 4 implies $e(t_i)$ converges to zero asymptotically. If it is an SLTI system, the proof that the error $e(t_i)$ converges exponentially pointwise follows from the proof of Theorem 4.1. □

5. FAULT DETECTION IN SPMSM WITH APPLICATIONS TO HEAVY HYBRID VEHICLES

5.1 Introduction and Motivation

5.1.1 Faults in a Permanent Magnet Synchronous Machine

The widespread need for conservation of diminishing fossil fuels, the economic benefits of more efficient fuel usage, and reduced environmental impact has motivated the development of heavy hybrid and heavy electric vehicles such as the Deere 644k Hybrid Wheel Loader and the Caterpillar D7E Dozer. An electric motor often utilized in these vehicles is the Permanent Magnet Synchronous Machine (PMSM). PMSMs are popular in such vehicles because of their higher torque density compared to induction and switched reluctance electric motors [51]. There are two types of the PMSM, interior mount and surface mount. The surface mount PMSM, denoted SPMSM herein, has permanent magnets attached to the surface of the rotor. Typically, these magnets are made of rare-earth materials such as neodymium iron boron (NeFeB) which produce a relatively high maximum energy product BH for a given size and weight. Only the SPMSM is considered in this chapter.

The stator of a SPMSM contains windings associated with each phase of a 3-phase machine. See Figures 5.1 and 5.2. These windings are spaced according to a particular geometric design. The windings associated with the same electrical phase can be in close proximity within winding bundles on the stator. Due to high temperature heating from I^2R losses in the windings, vibrations, and materials aging, the stator coils are prone to shorts. According to SKF Electric Motor Condition Monitoring Company, 30% of motor failures are due to stator winding failures [52]. The aforementioned bundles are common places for shorts, and are termed inter-turn short-circuit (ITSC) faults. A General Electric study, cited in [52], reports that 80% of motor failures begin as turn-to-turn insulation failures, i.e. ITSC faults. This is partly because machine vibrations can cause the bundles to rub against a sharp edge of the stator often causing an insulation failure in two of the bundle wires resulting in an ITSC fault. A "tooth" of the stator (around which a coil is wound) is another possible location for an ITSC fault. Here, two insulation failures on wires on the same tooth can lead to the an ITSC fault using the metal in the tooth to complete the short circuit.

When an ITSC fault occurs in the stator windings, a closed loop of wire is effectively created within the windings of the phase containing the fault. This closed loop of wire is coupled magnetically to the changing magnetic fields created by the remaining healthy phase windings and the rotating magnets. The magnetic flux through the closed loop of wire creates an eddy current which circulates within the wire. If left undetected, the ITSC fault can lead to further insulation failures risking a short to ground and potentially a fire. A short-to-ground event can cause damage to the electic machine and other electrical equipment.

5.1.2 Chapter Objectives

This chapter investigates the fault-modeling and fault-detection of a three-phase SPMSM using an observer strategy. The (ITSC fault) observer must detect an ITSC fault before such can cause unsafe operating conditions. According to the recent survey paper [53], diverse researchers have considered several methods for detecting ITSC faults in a PMSM. One such technique, termed motor current signature analysis (MCSA), detects changes in the frequency content of the current and voltage waveforms using filtering techniques based upon Fast Fourier Transform and Discrete Wavelet Transform algorithms [52–54]. Other proposed techniques for fault detection include finite element models and artificial intelligence algorithms. However, these techniques require considerable machine-specific tuning and analysis [53].

In order to avoid considerable machine-specific tuning and analysis, the observer structure utilized herein builds on an analytical model (having known parameters) of the stator windings as a function of the degree of fault. As with all observers, sensor measurements of the system inputs and outputs drive an algorithm (dependent on the analytic model) that produces state estimates, fault level estimates, and associated output estimates over some interval of time. The error between the estimated outputs and the actual sensor driven outputs determines, according to some metric, whether or not an ITSC fault has occurred as well as its severity. Finally, in order to determine safe or unsafe continued motor operation due to thermal heating maximums, the observer herein additionally estimates the eddy loop current denoted i_{fs} whose magnitude can cause excessive heating. Of course, stator winding faults are not restricted to ITSC faults and include shorts to ground and open circuit faults. Although these faults do occur in practice, the focus of this chapter is ITSC faults which cause the majority of motor failures [52].

Building around the moving horizon observer (MHO) of [22], we re-pose the observer problem as a dynamic model-based optimization problem. Conditions for the observer to converge are given therein. Further details are given in Section 5.4.

Another objective of this research is to develop a fault mitigation controller framework that allows the hybrid vehicle (of which the SPMSM is an integral part) to continue to function albeit at a substantially reduced operational level. In the case of a large earth mover, this might allow the vehicle to limp back to its truck hauler for delivery to the service center. In the case of a small hybrid vehicle like a Toyota Prius, the vehicle could drive slowly to a service center or other destination.

A so-called supervisory level controller along the lines set forth in [4], [55], and [56] coordinates vehicle control by determining optimized power flows to the individual subsystems. For example, for a diverse set of situations, the supervisory level controller would determine how best to utilize the electric motor vs. the internal combustion engine (ICE) or recover energy with regenerative braking. For efficient and feasible optimization strategies, the supervisory level models are power flow based and utilize efficiency maps pertinent to the individual subsystems. In the case of the SPMSM, such an efficiency map depends on whether or not the motor has a fault as well as on the degree of fault.

When faults in the windings exceed a level of 10-20% or more, safety may dictate a shut down of the vehicle. The permanent magnets of the traction PMSM (one of two PMSM in the powertrain) are attached to the powertrain output shaft, i.e. the output shaft is the PMSM rotor; thus as long as the shaft turns, the permanent magnets will cause an eddy current to flow in the shorted stator coils. As will be seen, such eddy currents can be extremely large causing high temperatures in the motor coils that exceed the maximum allowable operating temperature and thus unsafe operation. For fault levels at 10-20% or below, it may be possible to limp the motor and vehicle along.

In summary, our fault tolerant controller at the supervisory level uses the MHO ITSC fault observer as a component of the SPMSM which determines the "mode" or fault level of its operation. The supervisory controller can then determine a possible fault tolerant or fault mitigating power flow control strategy. In addition, the observer estimates the eddy loop current i_{fs} in order to determine approximate thermal losses so as to determine safe or unsafe operation when a fault has occurred.

5.1.3 Recasting the Observer Problem in a Switched System Observability Setting

It is convenient at the supervisory level to consider a finite set of possible fault levels between 0 (non-fault case) and 10-20%. In the case of the Prius, we consider a maximum fault level of 10% based on experimental evidence for reasonable vehicle operation. Each different fault level induces a different linear state model of the SPMSM. As such, each of the fault levels can be viewed as a mode associated with a specific linear dynamical state model. The ability to distinguish and identify the modes and mode switching times then reduces to the so-called switched observability problem discussed in the subsection below. The details of the SPMSM stator model with and without fault are developed in Sections 5.2 and 5.3. However, in general, for each degree of fault σ the state model has the form

$$\begin{aligned} E(\sigma)\dot{x} &= A(\sigma)x + B(\sigma)u \\ y &= C(\sigma)x + D(\sigma)u, \end{aligned} \tag{5.1}$$

where $x \in \mathbb{R}^n$ will represent the stator currents and eddy current, $u \in \mathbb{R}^m$ represents the voltage inputs and back electromotive forces, $y \in \mathbb{R}^p$ represents the current and voltage measurements, $E(\sigma) \in \mathbb{R}^{n\times n}$ is an inductance matrix, and $A(\sigma)$, $B(\sigma)$, $C(\sigma)$, and $D(\sigma)$ are real matrices of appropriate dimension. Equation 5.1 is valid for every degree of fault $\sigma \in [0, 1]$, i.e. the matrices change as a function of σ. We remark again that for each such fault level, mode, there is an associated efficiency map that must be used by the supervisory level controller to determine reasonable operation of the vehicle and how best to limp the vehicle along if the fault level is sufficiently low.

Determining feasibility of reconstructing the degree of fault σ requires proving distinguishability of each LTI system associated with the degrees of fault $\sigma_1 \neq \sigma_2 \in [0, 1]$. However, we shall see that distinguishability between one pair of degrees of fault (σ_1, σ_2) will imply that almost all degrees of fault are distinguishable. This allows for the application of the switched linear system observability results ([6, 20]) to the ITSC fault detection problem. We now review the relevant switched system observability results.

5.1.4 Review of Switched System Observability Results

The results surveyed in this section use a mode signal v to represent the set of finite modes of operation so as to distinguish it from the fault severity level σ. A switched linear system has the form

$$\begin{aligned} \dot{x} &= A_v x + B_v u \\ y &= C_v x + D_v u, \end{aligned} \tag{5.2}$$

where $v \in \{1, 2, \cdots, n_{modes}\}$ is the unknown switching sequence, $A_i \in \mathbb{R}^{n \times n}$, $B_i \in \mathbb{R}^{n \times m}$, $C_i \in \mathbb{R}^{p \times n}$, $D_i \in \mathbb{R}^{p \times m}$ for $i = 1, 2, \cdots, n_{modes}$, and u is the measurable control input. Given a piecewise continuous mode sequence v, piecewise continuous input u, and initial condition $x_0 \in \mathbb{R}^n$, the differential equation (5.2) has a unique solution $x(t)$. Consequently, the output sequence corresponding to the state sequence $x(t)$ is unique. Given that the input u and output y are measured, the switched system observability problem is to determine the initial state x_0 and mode sequence $v(t)$ from the given measurements. Conditions for solvability are first addressed.

In the case of no input, $u \equiv 0$, it is proven in [6] that the switching sequence $v(t)$ and initial state x_0 is observable given output measurements y if and only if for each pair of modes $i \neq j \in \{1, 2, \cdots, n_{modes}\}$ the extended linear system

$$\begin{aligned} \tilde{x} &= A_{i,j}\tilde{x} \\ \tilde{y} &= C_{i,j}\tilde{x} \end{aligned}$$

with system matrices

$$A_{i,j} = \begin{bmatrix} A_i & 0 \\ 0 & A_j \end{bmatrix}, \quad C_{i,j} = \begin{bmatrix} C_i & C_j \end{bmatrix},$$

is observable (in the classical sense). The addition of a smooth input u is considered in [18]. Therein, it is proven that the switching sequence $v(t)$ and initial state x_0 is reconstructable given input and output measurements for almost every smooth input if each pair (A_i, C_i), $i \in \{1, 2, \cdots, n_{modes}\}$, is observable (in the classical sense) and there is a nonzero difference in the Toeplitz matrices, $\Gamma_{2n}(A_i, B_i, C_i, D_i) - \Gamma_{2n}(A_j, B_j, C_j, D_j) \neq 0$, for each $i \neq j \in \{1, 2, \cdots, n_{modes}\}$, where

$$\Gamma_{2n}(A, B, C, D) = \begin{bmatrix} D & 0 & 0 & \cdots & 0 & 0 \\ CB & D & 0 & \cdots & 0 & 0 \\ CAB & CB & D & \cdots & 0 & 0 \\ CA^2B & CAB & CB & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ CA^{2n-1}B & CA^{2n-2}B & CA^{2n-3}B & \cdots & CB & D \end{bmatrix}. \tag{5.3}$$

These observability results are extended in [8] for nonsmooth inputs, but this is beyond the scope of this review.

5.1.5 Application to ITSC Fault Observability

Let $\sigma_1 \neq \sigma_2 \in [0,1]$ be two degrees of fault. Are these two degrees of fault distinguishable? To verify this, one can construct LTI systems for each degree of fault. Using the notation in (5.2), define $A_{\sigma_i} = E^{\dagger}(\sigma_i)A(\sigma_i)$, $B_{\sigma_i} = E^{\dagger}(\sigma_i)B(\sigma_i)$, $C_{\sigma_i} = C(\sigma_i)$, and $D_{\sigma_i} = C(\sigma_i)$ for $i = 1, 2$. LTI systems $(A_{\sigma_1}, B_{\sigma_1}, C_{\sigma_1}, D_{\sigma_1})$ and $(A_{\sigma_2}, B_{\sigma_2}, C_{\sigma_2}, D_{\sigma_2})$ are distinguishable for almost all inputs if

$$\|\Gamma_{2n}(A_{\sigma_1}, B_{\sigma_1}, C_{\sigma_1}, D_{\sigma_1}) - \Gamma_{2n}(A_{\sigma_2}, B_{\sigma_2}, C_{\sigma_2}, D_{\sigma_2})\|_F^2 \neq 0. \tag{5.4}$$

Treating σ_1 and σ_2 as variables, the norm defined in (5.4) is a polynomial in σ_1 and σ_2. If (5.4) is nonzero for some pair (σ_1, σ_2), then the set of indistinguishable degrees of fault is an algebraic variety of lower dimension intersected with the interval $[0,1]$, i.e., almost all degrees of fault are distinguishable.

In summary, the ITSC fault detection problem can be viewed as a switched system with unknown switching sequence $\sigma(t)$. The objective is to estimate the switching sequence $\sigma(t)$ and fault current i_{fs} using a modified form of the MHO introduced in [22]. In Section 5.4, if certain nonlinear observability conditions are satisfied (highly difficult to verify) the modified MHO observer can be proven to converge. Alternatively, the switched system observability conditions in (5.4) are easily verified and sufficient to guarantee that distinguishability between almost all degrees of ITSC fault, provided there exists a pair (σ_1, σ_2) which are distinguishable. When σ_1 and σ_2 are sufficiently close, there is, of course, a level of distinguishability based on how close (5.4) is to zero. Practically speaking, this is inconsequential for the MHO since the degree of fault is approximated with a nonlinear optimization rather than "distinguishing" between two adjacent levels of fault.

Section 5.2 introduces a model for the SPMSM without fault. Section 5.3 introduces the ITSC fault model for SPMSM. Sections 5.4 and 5.5 develop the ITSC fault

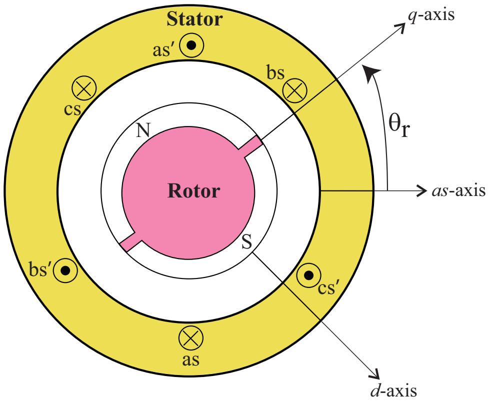

Fig. 5.1. This figure is a cross-sectional illustration of the SPMSM. The SPMSM has permanent magnets on the surface of the rotor and coils wound into the stator. Typically, SPMSM have more than two permanent magnets fixed to the rotor surface, unlike the two shown for illustrative purposes.

detection observer. The developed observer is simulated in Section 5.6. Application to fault-tolerant supervisory vehicle control in heavy hybrid vehicles is explored in Section 5.7.

5.2 Surface PMSM without Fault

Figure 5.1 illustrates the positioning of the permanent magnets on the rotor. The permanent magnets are positioned on the surface of the rotor to provide the largest magnetic flux variation in the stator windings for a given magnet strength. Nearly all of the rotor surface is magnetically hard, i.e. the rotor surface is covered by permanent magnets which maintain polarity under normal operation [51]. Motor torque is produced through the interaction of the magnetic fields produced by the rotor and those of the stator windings. The SPMSM is powered by a DC-AC inverter as illustrated in Figure 5.2. The wye configuration of the SPMSM stator is common in electric machines [51] and is the only configuration considered in this work.

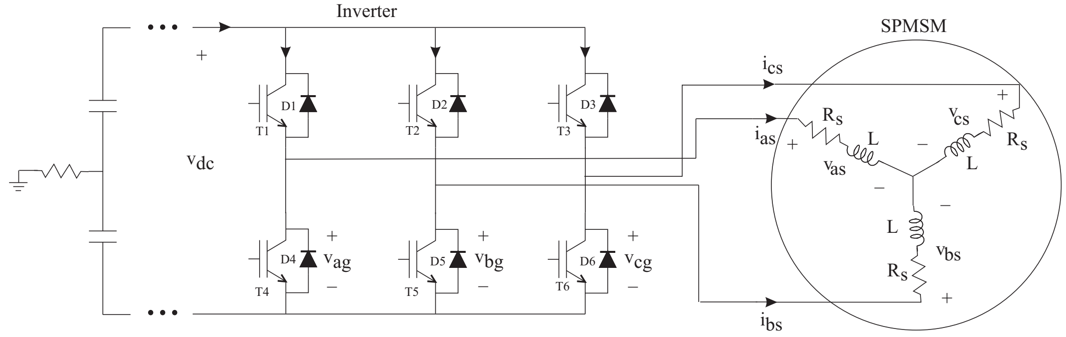

Fig. 5.2. The SPMSM stator connected to the DC-AC inverter. The wye configuration of the SPMSM stator winding is wound with a neutral point as shown on the right. As illustrated on the far left, the negative rail may not be connected to ground directly.

For the unfaulted case, the voltages of the three-phase SPMSM using the phase specific voltages and currents is given by ([51] and [57])

$$\begin{bmatrix} v_{as} \\ v_{bs} \\ v_{cs} \end{bmatrix} = \begin{bmatrix} R_s & 0 & 0 \\ 0 & R_s & 0 \\ 0 & 0 & R_s \end{bmatrix} \begin{bmatrix} i_{as} \\ i_{bs} \\ i_{cs} \end{bmatrix} + \begin{bmatrix} L & M & M \\ M & L & M \\ M & M & L \end{bmatrix} \frac{d}{dt} \begin{bmatrix} i_{as} \\ i_{bs} \\ i_{cs} \end{bmatrix} + \begin{bmatrix} e_a \\ e_b \\ e_c \end{bmatrix}, \tag{5.5}$$

where $v_{\zeta s}$ and $i_{\zeta s}$ denote the stator voltage and current in phase $\zeta = a, b, c$, respectively; R_s is the stator coil resistance in each phase; L and M denote the self and mutual inductance, respectively; and e_ζ is the back electromotive force (emf) in phase $\zeta = a, b, c$. Note, that Kirchoff's current law imposes the constraint $i_{as} + i_{bs} + i_{cs} = 0$, can be used to construct a reduced-order state model. Prior to an ITSC fault, the back emf is

$$\begin{bmatrix} e_a \\ e_b \\ e_c \end{bmatrix} = \omega_r \lambda_m \begin{bmatrix} \cos(\theta_r) \\ \cos(\theta_r - 2\pi/3) \\ \cos(\theta_r + 2\pi/3) \end{bmatrix}, \tag{5.6}$$

where ω_r and θ_r are the electrical rotor speed and position, respectively, and λ_m is the flux linkage. For almost all nonzero values of L and M, the coefficient matrix of the derivative of the phase currents is nonsingular. Hence (5.5) can be converted to a time-varying affine state model due to the time-varying back electromotive force voltage vector of (5.6).

The electromagnetic torque couples the electrical and mechanical components of the SPMSM. Without fault, the electromagnetic torque T_e and mechanical load torque T_L are related by a conservation of power equation

$$T_e \omega_m = e_a i_{as} + e_b i_{bs} + e_c i_{cs} = J \omega_m \dot{\omega}_m + B \omega_m^2 + T_L \omega_m, \tag{5.7}$$

with mechanical angular speed $\omega_m = \frac{d\theta_m}{dt} = \omega_r / n_p$ where the rotor has $n_p/2$ magnetic pole pairs, moment of inertia J, and viscous friction coefficient B, as illustrated in Figure 5.3.

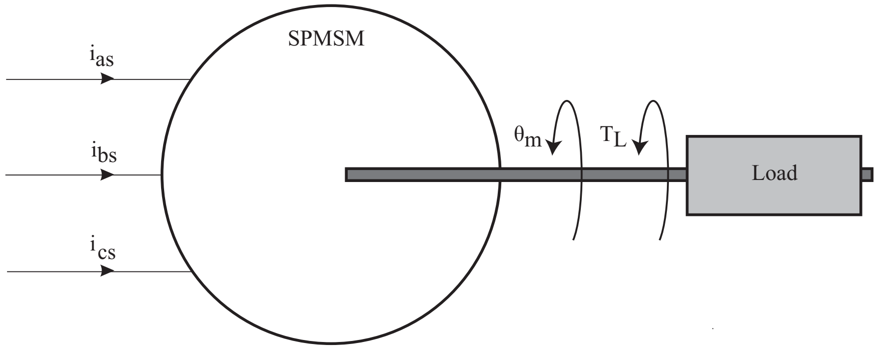

Fig. 5.3. SPMSM rotor connected to mechanical load. The rotor position is denoted θ_m and load torque T_L.

5.2.1 Extensions to Supervisory Powerflow Modeling

For supervisory level control, each component of the powertrain is minimally modeled as a power transfer device. To develop a power flow model for the SPMSM, we relate the power transferred from the inverter in each phase $\zeta = a, b, c$, denoted $P_{inv,\zeta} = v_{\zeta s} i_{\zeta s}$, to the rotor via electromagnetic power contributed in each phase $\zeta = a, b, c$, denoted $P_\zeta = e_\zeta i_{\zeta s}$. The relationship between the inverter-supplied power and electromagnetic power can be expressed in matrix form by premultiplying both sides of (5.5) by $\mathrm{diag}(i_{as}, i_{bs}, i_{cs})$ to obtain

$$\begin{bmatrix} P_{inv,a} \\ P_{inv,b} \\ P_{inv,c} \end{bmatrix} = \begin{bmatrix} R_s & 0 & 0 \\ 0 & R_s & 0 \\ 0 & 0 & R_s \end{bmatrix} \begin{bmatrix} i_{as}^2 \\ i_{bs}^2 \\ i_{cs}^2 \end{bmatrix} + \begin{bmatrix} i_{as} & 0 & 0 \\ 0 & i_{bs} & 0 \\ 0 & 0 & i_{cs} \end{bmatrix} \begin{bmatrix} L & M & M \\ M & L & M \\ M & M & L \end{bmatrix} \frac{d}{dt} \begin{bmatrix} i_{as} \\ i_{bs} \\ i_{cs} \end{bmatrix} + \begin{bmatrix} P_a \\ P_b \\ P_c \end{bmatrix}. \tag{5.8}$$

The total power supplied by the inverter is $P_{inv} \triangleq P_{inv,a} + P_{inv,b} + P_{inv,c}$. Hence by conservation of power,

$$P_{inv} = R_s i_{as}^2 + R_s i_{bs}^2 + R_s i_{cs}^2 + \frac{d}{dt}\Upsilon + P_a + P_b + P_c, \tag{5.9}$$

where the quantity

$$\Upsilon = \frac{1}{2} \begin{bmatrix} i_{as} & i_{bs} & i_{cs} \end{bmatrix} \begin{bmatrix} L & M & M \\ M & L & M \\ M & M & L \end{bmatrix} \begin{bmatrix} i_{as} \\ i_{bs} \\ i_{cs} \end{bmatrix} \tag{5.10}$$

is a Lyapunov-like energy function.

By using the quantity Υ it is possible to avoid certain kinds of singularities when optimizing the powerflow equations. Clearly, the terms $R_s i_{as}^2$, $R_s i_{bs}^2$, and $R_s i_{cs}^2$ represent winding losses while P_a, P_b, and P_c are back electro-motive powers. Hence, the analog of (5.7) in the supervisory power flow context is

$$P_a + P_b + P_c = J\omega_m \dot{\omega}_m + B\omega_m^2 + P_L, \tag{5.11}$$

where P_L is the power delivered to the load. These equations are ultimately used to develop efficiency maps that relate the input and output powers as functions of the mechanical rotor speed ω_m and desired output power P_L. Note, the winding losses are a function of the commanded current signals i_{as}, i_{bs}, and i_{cs}. The efficiency maps will be constructed by computing an optimal current control, which satisfies the physical operating constraints of the motor.

5.3 Extended Matrix Equations: Modeling ITSC Fault in Surface PMSM

In this section we extend the model for the SPMSM developed in the previous section to include a single ITSC fault. The fault model will include a degree or level of fault via the parameter $\sigma \in [0, 1]$. In the special no-fault-case when $\sigma = 0$, the fault model reduces to the model in (5.5)-(5.11).

5.3.1 ITSC Fault Equation Description

As discussed in [57], an ITSC fault causes imbalance or loss of symmetry between the variables of the three phases of the stator windings. This imbalance makes the conventional dq0-model [51] much less convenient for analysis of the SPMSM. Consequently, we construct the ITSC fault model using phase variables. For notation, let i_{fs} denote the shorted coil's eddy current induced by the nearby time-varying magnetic fields. Let $\sigma = N_f/N_T$ denote the fraction of faulted turns N_f among the total N_T turns in the faulted phase. Based on [57], this shorted coil has resistance σR_s, flux linkage $\sigma\lambda_m$, self inductance $\sigma^2 L$, and mutual inductance σM between the remaining healthy phases. The phase containing the ITSC fault has $(N_T - N_f)$ unfaulted turns reducing the resistance to $(1 - \sigma)R_s$, flux linkage to $(1 - \sigma)\lambda_m$, self inductance to $(1 - \sigma)^2 L$, and mutual inductance between the other healthy phases to $(1 - \sigma)M$. The shorted coil and the phase containing the ITSC fault are also inductively coupled. Since the shorted coil is wound on the same stator tooth as the remaining healthy turns in that phase, the shorted coil and loop containing the shorted coil have a mutual inductance $\sigma(1-\sigma)L$.[1] For simplicity, we will assume that the fault occurs in phase-a. It is a straightforward extension to model the fault in phases-b or-c. If there are faults in two phases simultaneously, two eddy currents will be present as per the models developed in the appendix.

The stator voltage equations with a single ITSC fault in phase-a, suitably modified from those appearing in [57], are given by

$$\begin{bmatrix} v_{as} \\ v_{bs} \\ v_{cs} \\ 0 \end{bmatrix} = \begin{bmatrix} (1-\sigma)R_s & 0 & 0 & 0 \\ 0 & R_s & 0 & 0 \\ 0 & 0 & R_s & 0 \\ 0 & 0 & 0 & \sigma R_s \end{bmatrix} i_{abcf} + L_f(\sigma)\frac{d}{dt} i_{abcf} + \begin{bmatrix} e_a \\ e_b \\ e_c \\ e_f \end{bmatrix}, \tag{5.12}$$

where $i_{abcf} \triangleq [i_{as}, i_{bs}, i_{cs}, i_{fs}]^\top$ and

$$L_f(\sigma) = \begin{bmatrix} (1-\sigma)^2 L & (1-\sigma)M & (1-\sigma)M & \sigma(1-\sigma)L \\ (1-\sigma)M & L & M & \sigma M \\ (1-\sigma)M & M & L & \sigma M \\ \sigma(1-\sigma)L & \sigma M & \sigma M & \sigma^2 L \end{bmatrix}. \tag{5.13}$$

The back emf terms are given by

$$\begin{bmatrix} e_a \\ e_b \\ e_c \\ e_f \end{bmatrix} = \omega_r \lambda_m \begin{bmatrix} (1-\sigma)\cos(\theta_r) \\ \cos(\theta_r - 2\pi/3) \\ \cos(\theta_r + 2\pi/3) \\ \sigma\cos(\theta_r) \end{bmatrix}. \tag{5.14}$$

Note that the fault loop has back emf e_f which has the same phase angle as the back emf in phase-a where the fault occurs. We can also observe that when there are no faults (i.e. $\sigma = 0$) equations (5.12)-(5.14) reduce to the unfaulted model in (5.8)-(5.11).

[1] This equation differs from those in [57] to ensure that the mutual inductances are physically realizable.

5.3.2 Extensions to Supervisory Powerflow Modeling: Fault Case

The above fault-dependent equation descriptions can be extended to explore the power relationship between the inverter, stator, and rotor post ITSC fault. The electromechanical power couples the electrical and mechanical components of the SPMSM as per the following conservation of power equation

$$T_e\omega_m = P_a + P_b + P_c + P_f = J\omega_m\dot{\omega}_m + B\omega_m^2 + T_L\omega_m, \tag{5.15}$$

where $P_\zeta = e_\zeta i_{\zeta s}$ for $\zeta = a, b, c, f$ Equation (5.15) which is the analog of (5.7). Note that P_f may appear to increase the total electromagnetic power in (5.15), but according to Lenz's Law the power P_f will always oppose the changing magnetic field. When the inverter-supplied power P_{inv} is zero, then P_f will oppose rotor movement similar to a frictional loss. When P_{inv} is nonzero, then P_f will reduce the combined change in magnetic field due to the mutual inductance from the remaining healthy coils and the rotor movement.

By pre-multiplying (5.12) by the vector of phase and fault currents, the power flows between the inverter and stator (analogous to (5.8)) are related by

$$\begin{bmatrix} P_{inv,a} \\ P_{inv,b} \\ P_{inv,c} \\ 0 \end{bmatrix} = \begin{bmatrix} (1-\sigma)R_s & 0 & 0 & 0 \\ 0 & R_s & 0 & 0 \\ 0 & 0 & R_s & 0 \\ 0 & 0 & 0 & \sigma R_s \end{bmatrix} \begin{bmatrix} i_{as}^2 \\ i_{bs}^2 \\ i_{cs}^2 \\ i_{fs}^2 \end{bmatrix} + \begin{bmatrix} i_{as} & 0 & 0 & 0 \\ 0 & i_{bs} & 0 & 0 \\ 0 & 0 & i_{cs} & 0 \\ 0 & 0 & 0 & i_{fs} \end{bmatrix} L_f(\sigma)\tfrac{d}{dt} \begin{bmatrix} i_{as} \\ i_{bs} \\ i_{cs} \\ i_{fs} \end{bmatrix} + \begin{bmatrix} P_a \\ P_b \\ P_c \\ P_f \end{bmatrix}. \tag{5.16}$$

Finally, the total inverter power for the faulted case, $P_{inv} = P_{inv,a} + P_{inv,b} + P_{inv,c}$, satisfies the conservation of power equation

$$P_{inv} = (1-\sigma)R_s i_{as}^2 + R_s i_{bs}^2 + R_s i_{cs}^2 + \sigma R_s i_{fs}^2 + \frac{d}{dt}\Upsilon_f(\sigma) + P_a + P_b + P_c + P_f, \tag{5.17}$$

where the new Lyapunov-like energy function Υ_f is

$$\Upsilon_f(\sigma) = \begin{bmatrix} i_{as} & i_{bs} & i_{cs} & i_{fs} \end{bmatrix} L_f(\sigma) \begin{bmatrix} i_{as} \\ i_{bs} \\ i_{cs} \\ i_{fs} \end{bmatrix}.$$

As expected, when $\sigma = 0$, equations (5.15) and (5.16) reduce to the equivalent unfaulted equations given in (5.7) and (5.8), respectively.

5.3.3 Fault Current Simulation

In Section 5.6, an SPMSM is simulated at a constant rotor speed of $\omega_m = 700$ rpm with controlled currents given in (5.49) for parameter values given in Table 5.1. To develop some qualitative understanding and to demonstrate how an ITSC fault affects the motor, we simulate the fault model (5.12) subject to an ITSC fault in phase-a occurring at 0.5s. Given the controlled currents as in (5.49), after the fault occurs the eddy current, i_{fs}, is excited, as illustrated in Figure 5.4. To demonstrate how the fault severity affects the fault current, i_{fs} is simulated for four fault severity levels, $\sigma_f = 1\%, 2\%, 5\%, 10\%$, again shown in Figure 5.4. When the ITSC fault occurs, the fault current, i_{fs}, is excited to roughly ten times the magnitude of 50A for the controlled current specified in (5.49). As long as the rotor is turning, the permanent magnets mounted thereon, will induce a large eddy current in the faulted coil. The eddy current generates heat that can become a safety hazard by causing further electrical insulation failures.

To maintain the desired stator current waveforms in (5.49), the commanded stator voltages v_{as}, v_{bs}, and v_{cs} will also change based on the degree of fault, as shown in Figure 5.5. Note, the simulation illustrated in Figures 5.4 and 5.5 presumes that the controlled voltages maintain the desired stator currents “instantaneously”. This is why the stator voltages jump at 0.5s. Usually current control is implemented via a closed loop current controller. In practice, the current control loop is less

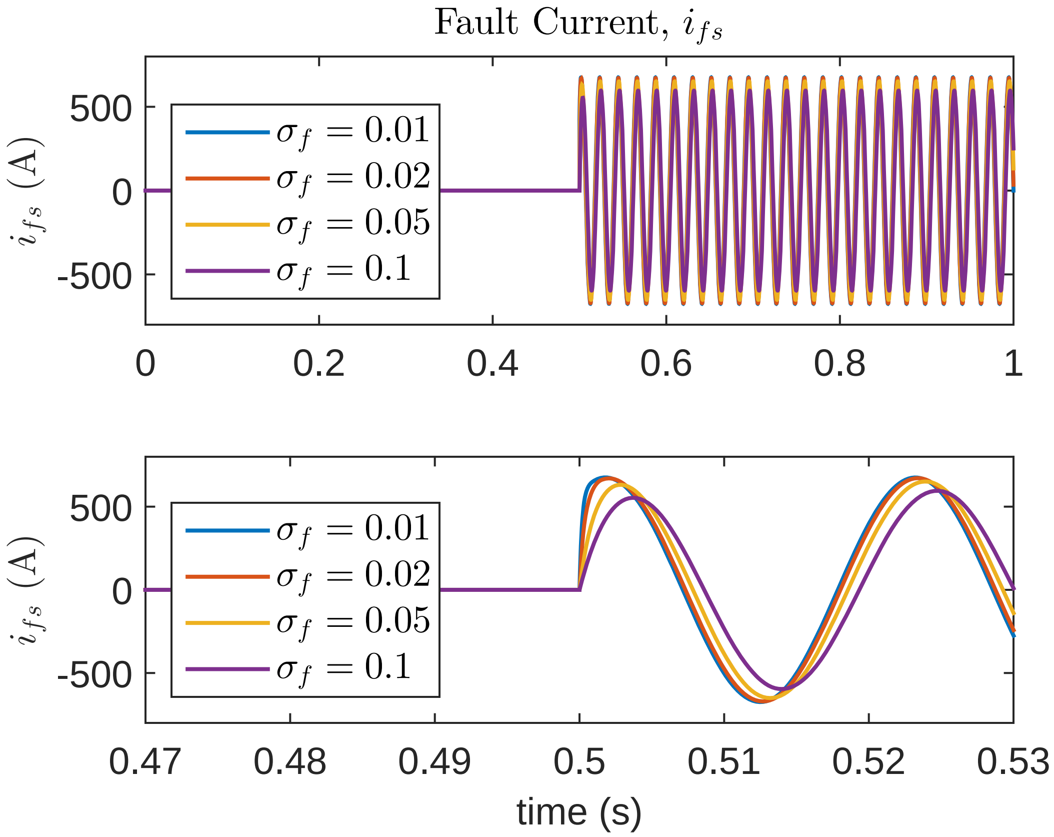

Fig. 5.4. Fault current for various degrees of fault. The fault of severity σ_f occurs at 0.5s. The lower figure zooms in on the interval surrounding 0.5s to see the difference between each level of fault severity.

responsive, but will have a reasonably fast time constant. One observes that the transient behavior in this simulation quickly dies away (about 5ms). The simulation in Section 5.6 is concerned primarily with the steady-state behavior so the simplifying assumption that the current loop is more or less instantaneous will have little effect.

Although the fault current i_{fs} is excited to over ten times the magnitude of the controlled stator currents, the amount of energy dissipated via heat in the shorted coil depends on the faulted coil resistance $\sigma_f R_s$. Figure 5.6 plots the instantaneous inverter-supplied power P_{inv} and electro-motive power P_{abcf}. When the ITSC fault occurs at 0.5s, both the inverter-supplied power and the electro-motive power exhibit oscillatory behavior due to the imbalance between the power transfer of the three phases. To show how the magnitude of the power flows are affected by the ITSC fault, Figure 5.7 plots the average inverter-supplied power $\bar{P}_{inv}$ and the average electro-motive power $\bar{P}_{abcf}$ for each degree of fault, $\sigma_f = 1\%, 2\%, 5\%, 10\%$. The averages $\bar{P}_{inv}$ and $\bar{P}_{abcf}$ are computed at time t by averaging the instantaneous power over the window $[t - T_{period}, t]$ where $T_{period} = \frac{2\pi}{\omega_r}$ is the electrical period. As Figure 5.7 illustrates, the electromagnetic output power $\bar{P}_{abcf}$ drops as the degree of fault increases. It is also interesting to note that the inverter supplied power $\bar{P}_{inv}$ also changes slightly as a function of the degree of fault. At 10% fault, the efficiency $100 \times \bar{P}_{abcf}/\bar{P}_{inv}$ drops to about 50%. Since this "lost" energy is converted to heat within the shorted loop, it is safety-critical that the fault is detected quickly.

Is the ITSC fault detectable? From Figure 5.5, the stator voltages required to maintain the desired stator differ before and after the ITSC fault at 0.5s. However, for a 1% fault, the steady-state voltage signals are only minimally affected. Fortunately, the inverter-supplied power, P_{inv}, provides a far more measurable difference when the fault occurs. As seen in Figure 5.6, the inverter-supplied power P_{inv} oscillates after the fault occurs. This oscillation is caused by an power contribution imbalance between the faulted and the two unfaulted phase windings. For given commanded currents, the average inverter-supplied power is also affected by the fault as shown in Figure 5.7. The electromagnetic power P_{abcf} is also plagued by the same oscillatory

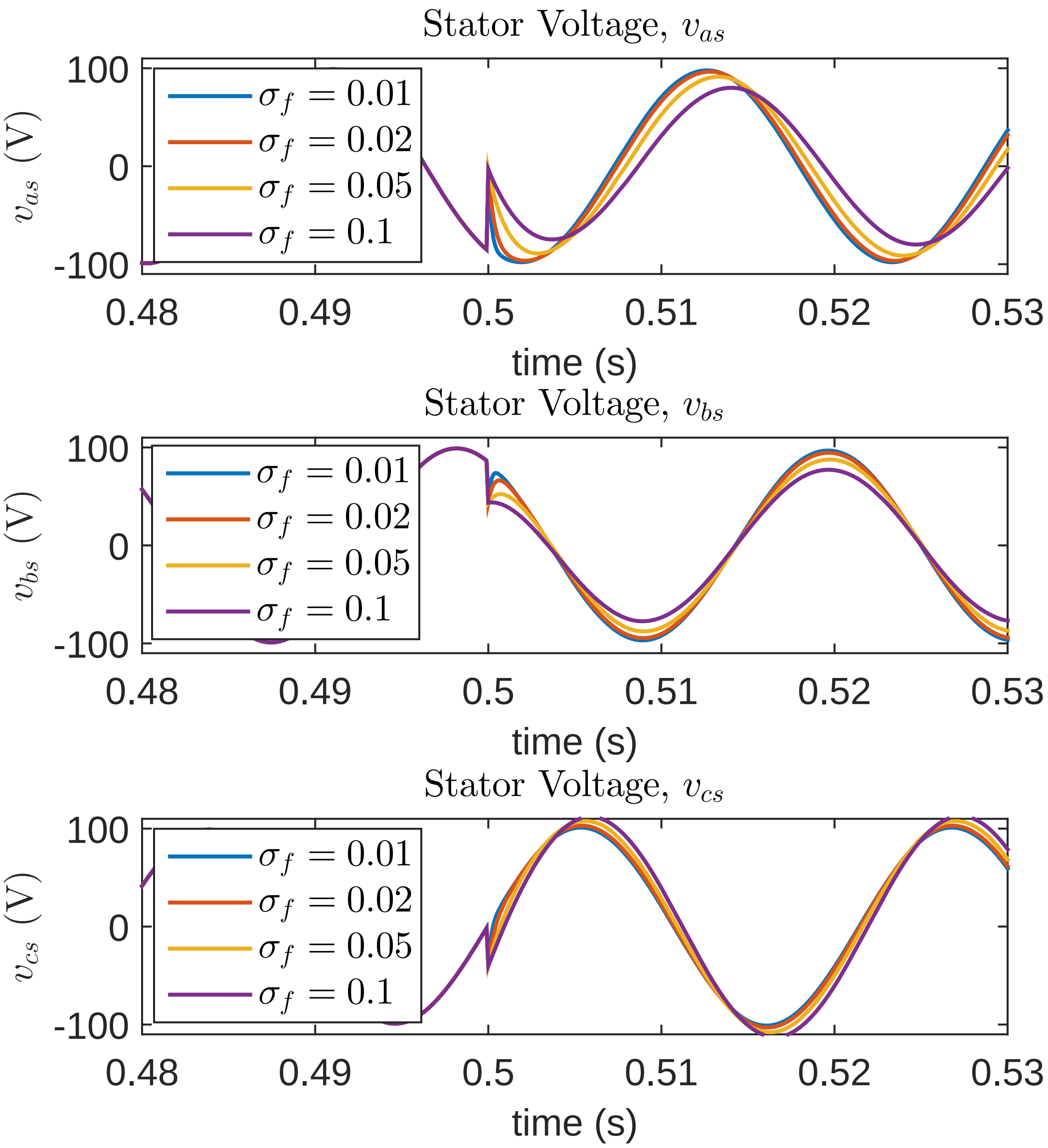

Fig. 5.5. Stator voltages for various degrees of fault. The fault of severity σ_f occurs at 0.5s. The stator voltage is assumed to be chosen to maintain stator currents given in (5.49).

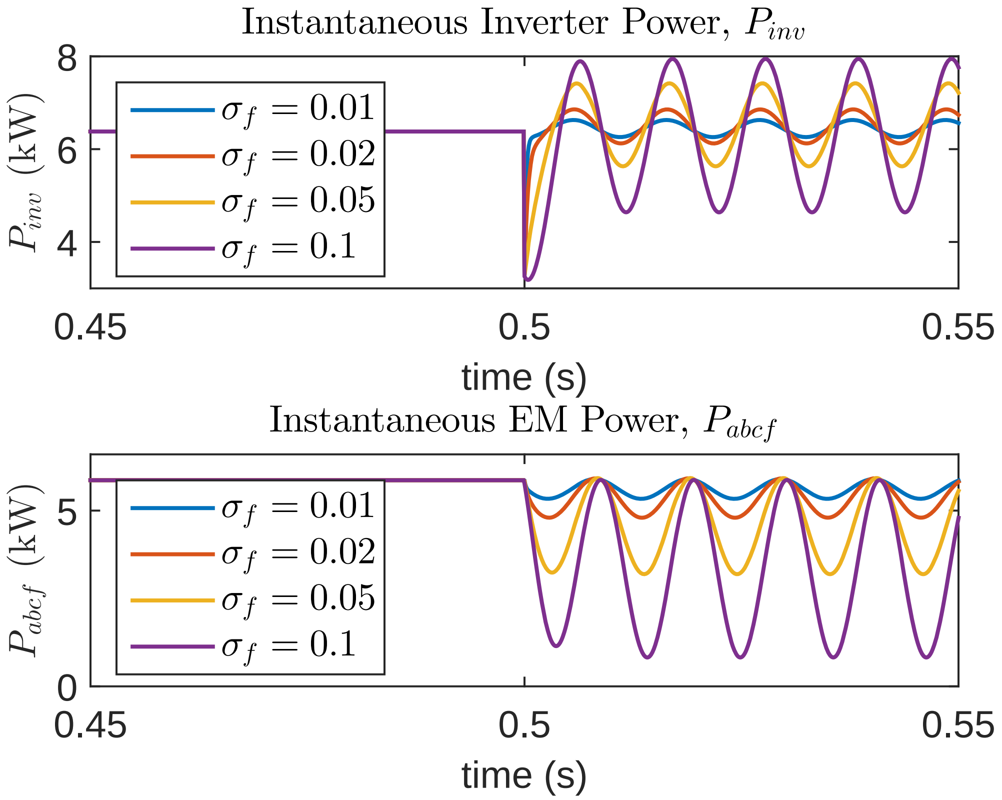

Fig. 5.6. (top) Plot of the inverter supplied power P_{inv} for various degrees of fault. The fault of severity σ_f occurs at 0.5s. (bottom) Plot of the electromagnetic power P_{abcf} for various degrees of fault. The average electromagnetic power is also computed as an average instantaneous power over the window $[t - T_{period}, t]$.

and average power effects although the electromagnetic power is usually unavailable for direct measurement, see Figures 5.6 and 5.7. The measurable differences caused by the ITSC fault demonstrates feasibility of the ITSC fault detection problem. The development of the fault detection method proposed in this chapter begins in the following section.

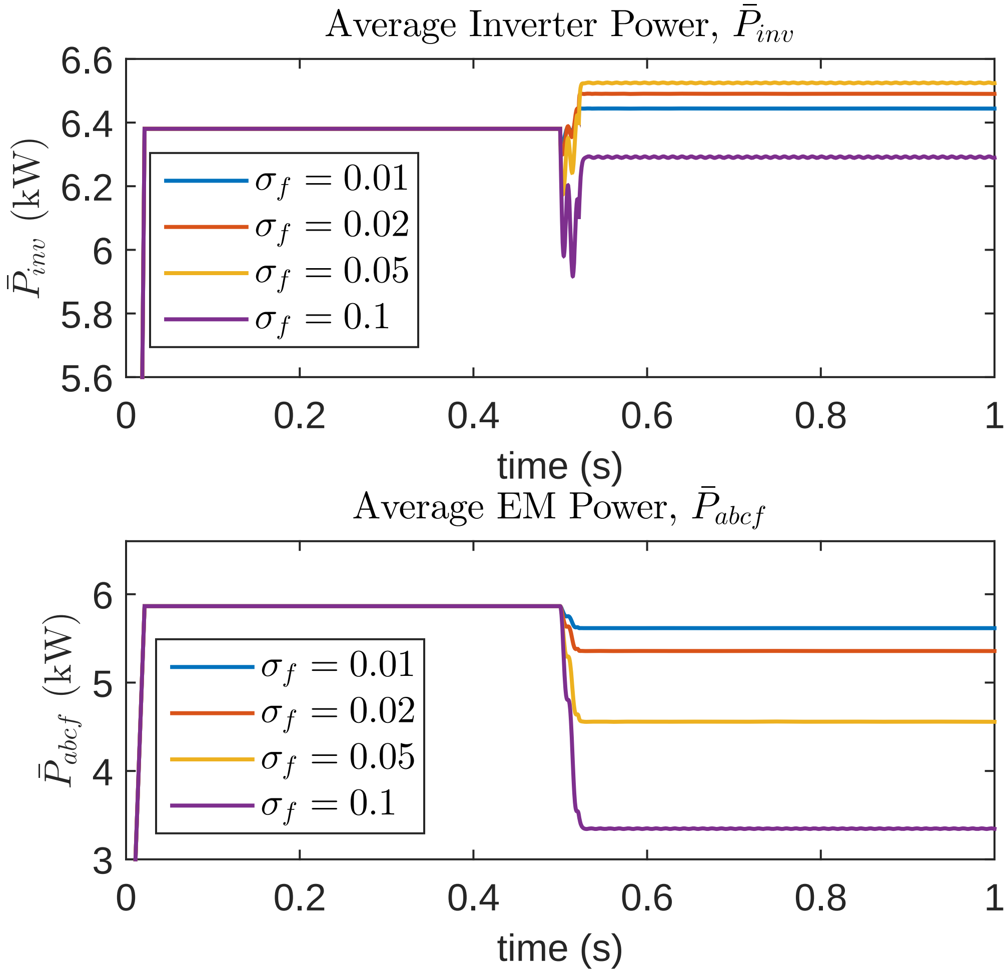

Fig. 5.7. (top) Plot of the average inverter supplied power $\bar{P}_{inv}$ for various degrees of fault. The average power is computed as the average instantaneous power over a window $[t - T_{period}, t]$ for each time t, where $T_{period} = \frac{2\pi}{\omega_r}$ is the period. The fault of severity σ_f occurs at 0.5s. (bottom) Plot of the average electromagnetic power $\bar{P}_{abcf}$ for various degrees of fault. The average electromagnetic power is also computed as an average instantaneous power over the window $[t - T_{period}, t]$.

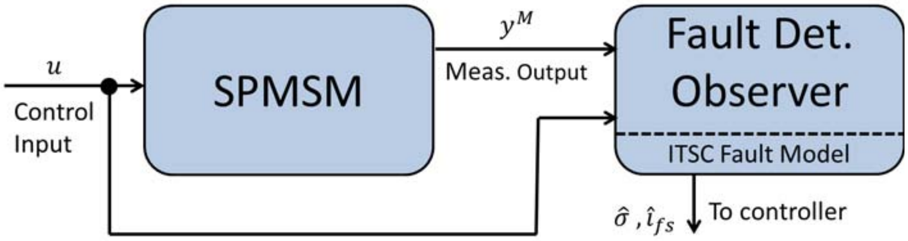

Fig. 5.8. The fault detection observer uses known control input u, measured output y^M, and the ITSC fault model to produce an estimate for the degree of fault $\hat{\sigma}$ and the fault current $\hat{\imath}_{fs}$.

5.4 Nominal Fault Detection Observer

5.4.1 ITSC Observer Problem Statement

How can the ITSC fault be detected? In our context, the fault detection observer estimates the degree of fault and fault current consistent with input and output measurements. The fault detection observer is illustrated in Figure 5.8. The objective of this section is to formalize the ITSC fault detection observer problem. This begins by defining the measured inputs and outputs.

The currents in the stator of the SPMSM are controlled by the inverter through voltages applied to the stator winding leads relative to the negative rail, denoted v_{ag}, v_{bg}, and v_{cg}. These measurable terminal voltages v_{ag}, v_{bg}, and v_{cg} determine the stator voltages relative to neutral, v_{as}, v_{bs}, and v_{cs}, which in turn drive the stator currents as per (5.12). Ideally, we would directly measure the stator to neutral voltages $v_{\zeta s}$, $\zeta = a, b, c$. However, electric machine manufacturers rarely provide direct access to neutral making the stator voltages directly unmeasurable or expensive to measure in terms of sensor placement in practice. Sensors for the line to line voltages are more readily available, i.e. measurements of $v_{\zeta w} \triangleq v_{\zeta s} - v_{ws}$ for $\zeta \neq w \in \{a, b, c\}$. The line to line voltages are measurable from the controlled voltages v_{ag}, v_{bg}, and v_{cg} as per

$$\begin{aligned} v_{ab}^{M} &= v_{ag}^{M} - v_{bg}^{M} \\ v_{bc}^{M} &= v_{bg}^{M} - v_{cg}^{M} \\ v_{ac}^{M} &= v_{ag}^{M} - v_{cg}^{M}, \end{aligned} \tag{5.18}$$

where the superscript M denotes measured signals. We also consider the electrical position θ_r and speed ω_r of the rotor to be measured signals. Using these measurements, the back emf e_{abcf} can be expressed as

$$\begin{bmatrix} e_a \\ e_b \\ e_c \\ e_f \end{bmatrix} = \begin{bmatrix} 1-\sigma & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \sigma & 0 & 0 \end{bmatrix} \begin{bmatrix} \omega_r^M \lambda_m \cos(\theta_r^M) \\ \omega_r^M \lambda_m \cos(\theta_r^M - 2\pi/3) \\ \omega_r^M \lambda_m \cos(\theta_r^M + 2\pi/3) \end{bmatrix}. \tag{5.19}$$

Note that the only unknown in (5.19) is σ, which is estimated. Thus the rightmost matrix in (5.19) becomes another measured input.

Since many commercial electric drive systems utilize stator current control, sensors are often available for the stator currents $i_{\zeta s}, \zeta = a, b, c$. We assume that each of the stator currents is available for measurement. In practice, we can reduce the number of sensors since the stator currents satisfy Kirchoff's current law, i.e., $i_{as} + i_{bs} + i_{cs} = 0$. One may be able to use a reduced number of sensors, but this reduction is not explored in this chapter.

When an ITSC fault occurs, the same voltage potential on the phase terminals produces different stator current responses. Essentially, the ITSC fault detection observer matches the given voltage signals to the resulting current measurements to determine the degree of fault σ, the fault current i_{fs}, and the stator currents $i_{\zeta s}$, $\zeta = a, b, c$. We can now pose the ITSC fault observer problem.

ITSC Observer Problem: Estimate the fault severity σ, fault current i_{fs}, and stator currents $i_{\zeta s}$, $\zeta = a, b, c$, given the ITSC fault model (5.12), measured signal

$$y^M = \begin{bmatrix} v_{ab}^M & v_{bc}^M & v_{ca}^M & i_{as}^M & i_{bs}^M & i_{cs}^M \end{bmatrix}^\top, \tag{5.20}$$

and known electrical rotor speed ω_r^M and position θ_r^M where the superscript M denotes measured variables.

ITSC fault detection is a nonlinear observer problem. For each fixed degree of fault σ, the dynamics in (5.12) are linear with respect to i_{abcf}, but an unknown degree of fault σ introduces a nontrivial nonlinearity. One approach to solving nonlinear observability problems is to use linear observers, such as the classical Luenberger dynamical observer [58]. Linear observers are numerically simple and well understood, but in general perform poorly on highly nonlinear systems. As an alternative, we propose the optimization-based approach developed in [22], known as a moving horizon estimator or moving horizon observer (MHO).

5.4.2 Moving Horizon Observer

As mentioned in the introduction, the MHO re-poses the estimation problem as an optimization problem. Consider the following nonlinear system

$$\begin{aligned} \dot{x} &= f(x, u^M) \\ y^M &= g(x, u^M), \end{aligned} \tag{5.21}$$

where $x \in \mathbb{R}^n$ is the state, $y^M \in \mathbb{R}^p$ is the measured output, $u^M : \mathbb{R} \to \mathbb{R}^m$ is the bounded measurable input, and $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ and $g : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^p$ are known, locally Lipschitz functions with respect to both x and u^M. Recall that for $\mathcal{A}$, $\mathcal{B}$ metric spaces, $h : \mathcal{A} \to \mathcal{B}$ is a locally Lipschitz function if for all $a \in \mathcal{A}$ there exists a neighborhood U_a of a and a constant K such that for all $a_1, a_2 \in U_a$, $\|h(a_1) - h(a_2)\|_A \leq K\|a_1 - a_2\|_B$, where $\|\cdot\|_A$ and $\|\cdot\|_B$ denote the metric in $\mathcal{A}$

and $\mathcal{B}$, respectively. The MHO developed in [22] is based on solving the optimization problem

$$\min_{\hat{x}_0 \in \mathbb{R}^n} \int_{t-T}^{t} \|y^M(t) - \hat{y}(t)\|^2 dt \tag{5.22}$$

subject to:

$$\dot{\hat{x}}(t) = f(\hat{x}(t), u^M(t)),\ \hat{x}(t-T) = \hat{x}_0 \tag{5.23}$$

$$\hat{y}(t) = g(\hat{x}(t), u^M(t)). \tag{5.24}$$

where T is the finite horizon and $\hat{y}(t)$ is the estimated output driven by the state trajectory $\hat{x}(t)$ which satisfies the underlying differential equation with the estimated initial condition $\hat{x}_0$. The specific approach in [22] is not to solve (5.22) at each time t but rather to sequentially solve the optimization over successive horizon windows $[t_k - T, t_k]$ where $t_1 < t_2 < t_3 < \cdots$. However, our approach is not to achieve the absolute minimum over $[t_k - T, t_k]$, but rather to impose a cost reduction by a factor of $\beta \in (0, 1)$ from one window to the next. So if at time t_k, the norm in (5.22) is equal to K_k, then over the next horizon $[t_{k+1} - T, t_{k+1}]$ the minimization in (5.22) is iterated until the norm is less than $K_{k+1} = \beta K_k$. This would continue until the norm in (5.22) is in a sufficiently small neighborhood of zero, in which case $y^M(t) - \hat{y}(t; \hat{x}_0) \approx 0$. Given the presence of modeling errors, sensor noise, and numerical round-off, reaching the "perfect minimum" of zero is unlikely. The benefit of this approach is that the observer/estimate convergence improves incrementally over successive horizons in contrast to the larger computational effort needed to achieve the minimum of (5.22) over each horizon.

To guarantee solvability of the observer problem, it is assumed that for each initial condition $x_0, x_0' \in \mathbb{R}^n$, the corresponding output trajectories $y(x(t), u(t))$ and $y(x'(t), u(t))$ satisfy

$$\int_{t-T}^{t} \|y(x(t), u(t)) - y(x'(t), u(t))\|^2 dt \geq \gamma \|x_0 - x_0'\|^2, \tag{5.25}$$

for some fixed $\gamma > 0$. This uniform observability condition reduces to the classical observability Gramian in the case of time-varying linear systems and time-invariant linear systems as shown in Appendix B. The uniform observability condition in (5.25) is difficult to verify for nonlinear systems. As mentioned earlier, for the ITSC fault detection problem we will presume that the unfaulted state model is observable, which is easily verified for the parameter values of a typical SPMSM and available sensor measurements. Further as asserted earlier, the faulted model is observable for almost all fault levels $\sigma \in [0,1]$ if observable for at least one fault level σ_1. Hence, the structure of the SPMSM model allows us to assert generic observability of the system without having to verify the condition of (5.25).

In general, the MHO is a versatile observer often used to solve nonlinear observability problems [22, 59, 60]. Thus it is well suited for the ITSC fault detection problem. For linear state models, the MHO can be seen as a dual problem to the linear quadratic regulator (LQR) problem and thus enjoys a similar historical success [48, 61].

5.4.3 ITSC Observability

Recall that for the ITSC fault detection problem, the variables to be estimated are the stator currents i_{ws}, $w = a, b, c, f$, and the degree of fault σ. First we validate that the observability problem is feasible, i.e., different fault levels are distinguishable and the stator currents i_{ws} are observable.

To analyze the distinguishability of two degrees of fault $\sigma_1 \neq \sigma_2 \in [0,1]$, we first need to construct a switched linear time-invariant (SLTI) model that incorporates the measured signals in (5.20) and then verify distinguishability with (5.4). Unfortunately, only the line-to-line voltages v_{ab}, v_{bc}, and v_{ca} are measurable whereas the stator voltages v_{as}, v_{bs}, and v_{cs}, that appear in the state dynamics of (5.12) are not. Another problem with (5.12) is that Kirchoff's current law (KCL) disallows arbitrary initial conditions, because in the wye configuration $i_{as} + i_{bs} + i_{cs} = 0$. This means that (5.12) contains redundant information and a lower dimensional state model can capture all the relevant dynamical information.

To construct the lower dimensional state model (4^{th} order to 3^{rd} order) that utilizes the measured signals in (5.20), we do the following:

1. using KCL, substitute $i_{cs} = -i_{as} - i_{bs}$ in (5.12), i.e., for $i_{abf} \triangleq [i_{as}, i_{bs}, i_{fs}]^\top$

$$i_{abcf} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix} i_{abf} \triangleq M_{abf} i_{abf} \tag{5.26}$$

2. premultiply both sides of (5.12) by

$$M_v \triangleq \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{5.27}$$

   to obtain differential equations as functions of (i) $v_{ab}^M = v_{as} - v_{bs}$ and (ii) $v_{bc}^M = v_{bs} - v_{cs}$.

This results in the reduced-order equivalent state and output model:

$$\widetilde{L}_f(\sigma) \frac{d}{dt} i_{abf} = -\widetilde{R}_f(\sigma) i_{abf} + \widetilde{Q}(\sigma) u^M, \tag{5.28}$$

$$\widetilde{y}^M = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} i_{abf} \triangleq \widetilde{C}(\sigma) i_{abf}, \tag{5.29}$$

where $i_{abf} \triangleq [i_{as}, i_{bs}, i_{fs}]^\top$ is the reduced state vector. The measured input vector is

$$u^M = \begin{bmatrix} \lambda_m \omega_r^M \cos(\theta_r^M) & \lambda_m \omega_r^M \cos(\theta_r^M - 2\pi/3) & \lambda_m \omega_r^M \cos(\theta_r^M + 2\pi/3) & v_{ab}^M & v_{bc}^M \end{bmatrix}^\top, \tag{5.30}$$

and the measured output is $\widetilde{y}^M = [i_{as}, i_{bs}]^\top$. The new linear system matrices in (5.28) are

$$\begin{aligned}
\widetilde{L}_f(\sigma) &= M_v L_f(\sigma) M_{abf}, \\
\widetilde{R}_f(\sigma) &= M_v \begin{bmatrix} (1-\sigma)R_s & 0 & 0 & 0 \\ 0 & R_s & 0 & 0 \\ 0 & 0 & R_s & 0 \\ 0 & 0 & 0 & \sigma R_s \end{bmatrix} M_{abf}, \\
\widetilde{Q}(\sigma) &= \begin{bmatrix} -(1-\sigma) & 1 & 0 & 1 & 0 \\ 0 & -1 & 1 & 0 & 1 \\ -\sigma & 0 & 0 & 0 & 0 \end{bmatrix},
\end{aligned}$$

where M_{abf} and M_v are defined in (5.26) and (5.27), respectively.

To verify the fault distinguishability conditions in (5.4), a standard LTI system is constructed from (5.28) for each degree of fault σ_1 via the tuple of linear system matrices

$$(A_{\sigma_1}, B_{\sigma_1}, C_{\sigma_1}, D_{\sigma_1}) = \left(\widetilde{L}_f^\dagger(\sigma_1)\widetilde{R}_f(\sigma_1), \widetilde{L}_f^\dagger(\sigma_1)\widetilde{Q}(\sigma_1), \widetilde{C}(\sigma_1), 0\right). \tag{5.31}$$

If there exists two degrees of fault $\sigma_1 \neq \sigma_2$, for which (5.4) is satisfied, then (for generic inputs) almost all degrees of fault are distinguishable. For the SPMSM parameterized in Table 5.1 with $\sigma_1 = 0$ and $\sigma_2 = 1$, we compute

$$\|\Gamma_{2n}(A_{\sigma_1}, B_{\sigma_1}, C_{\sigma_1}, D_{\sigma_1}) - \Gamma_{2n}(A_{\sigma_2}, B_{\sigma_2}, C_{\sigma_2}, D_{\sigma_2})\|_F^2 = 3.16 \times 10^{20} \neq 0. \tag{5.32}$$

Thus $\sigma_1 = 0$ and $\sigma_2 = 1$ are distinguishable for almost all inputs, as per [18]. To show that almost all degrees of fault are distinguishable for almost all inputs, we consider the nontrivial polynomial (nontrivial by (5.32)) in σ_1 and σ_2 defined by (5.33),

$$\|\Gamma_{2n}(A_{\sigma_1}, B_{\sigma_1}, C_{\sigma_1}, D_{\sigma_1}) - \Gamma_{2n}(A_{\sigma_2}, B_{\sigma_2}, C_{\sigma_2}, D_{\sigma_2})\|_F^2. \tag{5.33}$$

Hence, the set of pairs (σ_1, σ_2) such that (5.33) is equal to zero is an algebraic variety of lower dimension, i.e., at worst unions of lines in $\mathbb{R}^2$. In addition, this algebraic variety must intersect the square $[0,1] \times [0,1]$ for two degrees of fault to be indistinguishable. Thus it is possible that the algebraic variety does not intersect $[0,1] \times [0,1]$ for pairs (σ_1, σ_2) with $\sigma_1 \neq \sigma_2$, i.e., that all degrees of fault are distinguishable. Hence for generic inputs it follows that almost all degrees of fault are, in fact, distinguishable.

The next question is whether the state i_{abf} is observable once the correct degree of fault is identified. This is verified using classical observability tests on the pair $(A_{\sigma_i}, C_{\sigma_i})$, such as the rank of the observability matrix. For the SPMSM parametrized in Table 5.1 with $\sigma_1 = 0$ and $\sigma_2 = 1$, we obtain

$$\begin{aligned} &\text{rank}[\mathcal{O}_3(A_{\sigma_1}, C_{\sigma_1})] = 2 \\ &\text{rank}[\mathcal{O}_3(A_{\sigma_2}, C_{\sigma_2})] = 3, \end{aligned}$$

where $\mathcal{O}_3(A, C)$ is the observability matrix for the pair (A, C), i.e.,

$$\mathcal{O}_i(A, C) = \begin{bmatrix} C^\top & (CA)^\top & \cdots & (CA^{i-1})^\top \end{bmatrix}^\top.$$

The result that $\text{rank}[\mathcal{O}_3(A_{\sigma_1}, C_{\sigma_1})] = 2$ implies that the state i_{abf} is not completely observable. This is understandable since $\sigma_1 = 0$ represents the unfaulted SPMSM and the fault current i_{fs} is unobservable because it is zero prior to an ITSC fault. On the other hand, since $\text{rank}[\mathcal{O}_3(A_{\sigma_2}, C_{\sigma_2})] = 3$, the entire state i_{abf} is observable for $\sigma_2 = 1$. Using the same arguments as in Section 55.1.5, this implies that i_{abf} is observable for almost all degrees of fault. Mathematically, the set of degrees of fault σ for which i_{abf} is unobservable is among a finite set of roots to a polynomial in σ. Any root, say $\sigma_* \notin [0,1]$ is not a physically realizable degree of fault. Hence, it is again possible that the current i_{abf} is observable for all degrees of fault and in the worst case i_{abf} is unobservable for a finite number of degrees of fault. Thus the ITSC observer problem is feasible for almost all degrees of fault.

5.4.4 Nominal ITSC Observer

Although it is possible to build a MHO for the reduced order model of the previous section, from a modeling perspective as well as a more direct utilization of the full order model developed earlier, we simply add the KCL equation as a constraint. There are also numerical advantages due to the sparseness of the larger set of equations.

Since the degree of fault is unknown but takes values in the interval $[0,1]$, we denote the observer below to be the nominal embedded moving horizon observer (EMHO).[2] As mentioned earlier, we assume that the ITSC fault occurs in phase-a. Relaxing this assumption is a straightforward extension, but the additional notation is not included for clarity.

In the EMHO framework, the ITSC fault detection problem has mode $\sigma \in [0,1]$ and state $i_{abcf} \triangleq [i_{as}, i_{bs}, i_{cs}, i_{fs}]^\top$. As described in Section 5.4.2, we consider a discretized set of final times given by $t_1, t_2, \cdots, t_k, \cdots$. For simplicity, we consider evenly spaced final times, i.e., $t_{k+1} - t_k = T_{shift}$.

So for a given horizon $[t_k - T, t_k]$ and $0 \leq h \leq T$, the nominal ITSC fault EMHO problem with fault in phase-a is given by

$$\min_{\substack{\hat{i}_{abcf}(t_k - h) \in \mathbb{R}^4 \\ \hat{\sigma}:[t_k - T, t_k] \to [0,1]}} \int_{t_k - T}^{t_k} \|y^M(t) - \hat{y}(t)\|^2 dt \tag{5.34}$$

subject to:

$$\hat{v}_{abcf} = R_f(\hat{\sigma})\hat{i}_{abcf} + L_f(\hat{\sigma})\frac{d}{dt}\hat{i}_{abcf} + e_{abcf}(\hat{\sigma}) \tag{5.35}$$

$$\begin{aligned} \hat{y} &= [\hat{v}_{as} - \hat{v}_{bs}, \hat{v}_{bs} - \hat{v}_{cs}, \hat{v}_{cs} - \hat{v}_{as}, \hat{i}_{as}, \hat{i}_{bs}, \hat{i}_{cs}]^\top \qquad (5.36) \\ &= [\hat{v}_{ab}, \hat{v}_{bc}, \hat{v}_{ca}, \hat{i}_{as}, \hat{i}_{bs}, \hat{i}_{cs}]^\top \end{aligned}$$

$$0 = \hat{i}_{as} + \hat{i}_{bs} + \hat{i}_{cs}, \tag{5.37}$$

[3] where (5.37) is a result of KCL,

$$\hat{i}_{abcf} = [\hat{i}_{as}, \hat{i}_{bs}, \hat{i}_{cs}, \hat{i}_{fs}]^\top, \tag{5.38}$$

$$\hat{v}_{abcf} = [\hat{v}_{as}, \hat{v}_{bs}, \hat{v}_{cs}, 0]^\top, \tag{5.39}$$

[2] This formulation is the dual to the embedded hybrid optimal control problem in that σ can vary continuously in $[0,1]$ (see [12, 61]).

[3] Kirchoff's current law takes the form of (5.37) only for ITSC faults, i.e., (5.37) only applies for shorts between phases and not shorts to ground. Modeling shorts to ground are not considered in this chapter.

$$R_f(\hat{\sigma}) = \begin{bmatrix} (1-\hat{\sigma})R_s & 0 & 0 & 0 \\ 0 & R_s & 0 & 0 \\ 0 & 0 & R_s & 0 \\ 0 & 0 & 0 & \sigma R_s \end{bmatrix} \tag{5.40}$$

$$L_f(\hat{\sigma}) = \begin{bmatrix} (1-\hat{\sigma})^2 L & (1-\hat{\sigma})M & (1-\hat{\sigma})M & \hat{\sigma}(1-\hat{\sigma})L \\ (1-\hat{\sigma})M & L & M & \hat{\sigma}M \\ (1-\hat{\sigma})M & M & L & \hat{\sigma}M \\ \hat{\sigma}(1-\hat{\sigma})L & \hat{\sigma}M & \hat{\sigma}M & \hat{\sigma}^2 L \end{bmatrix}, \tag{5.41}$$

and

$$e_{abcf}(\hat{\sigma}) = \omega_r \lambda_m \begin{bmatrix} (1-\hat{\sigma})\cos(\theta_r) \\ \cos(\theta_r - 2\pi/3) \\ \cos(\theta_r + 2\pi/3) \\ \hat{\sigma}\cos(\theta_r) \end{bmatrix}. \tag{5.42}$$

Of course, this problem is solved sequentially for each interval $[t_k - T, t_k]$ for $k = 1, 2, \cdots$. It is not necessary that these intervals be disjoint. As we will see in the forthcoming development, there are numerical advantages to having these intervals overlap.

Several aspects of the ITSC EMHO warrant explanation and elaboration. First, the variable h allows for the estimated state $\hat{i}_{abcf}(t_k - h)$ to be anywhere within the interval $[t_k - T, t_k]$. For example, when $h = T$ the EMHO observer reduces to the MHO observer described in Section 5.4.2, in that one is estimating the initial condition $\hat{i}_{abcf}(t_k - T)$ for the interval $[t_k - T, t_k]$. Another way of saying this is that the state estimate at the beginning of the interval, $\hat{i}_{abcf}(t_k - T)$, is either a delayed estimate of the current state $i_{abcf}(t_k)$ or must be integrated using (5.12). This value could be sensitive to errors in the estimated initial condition. Clearly, then the choice of h has an effect on the numerical implementation of the EMHO.

Moving the state estimate to the beginning of the interval, h small, has a smaller delay and less integration required to obtain the current estimate. Thus, small h naturally emphasizes the most recent measurements and adapts more quickly to changes in the measured output. However, if $h < T_{shift}$, where $T_{shift} = t_{k+1} - t_k$ for each k, then $t_{k+1} - h$ is not contained in the previous interval $[t_k - T, t_k]$ as illustrated in Figure 5.9. The practical consequence of selecting $h < T_{shift}$ occurs when integrating the previous estimate $\hat{i}_{abcf}(t_k - h)$ from $t_k - h$ to $t_{k+1} - h$ to hot-start the next estimate $\hat{i}_{abcf}(t_{k+1} - h)$. Namely, the issue is that when computing $\hat{i}_{abcf}(t_k - h)$ and $\hat{\sigma}([t_k - T, t_k])$, no measurements from the interval $[t_k, t_{k+1} - h]$ were utilized. Consequently, one either makes assumptions about the interval $[t_k, t_{k+1} - h]$ to allow for the integration (such as assuming the degree of fault σ does not change) or uses another suboptimal initial guess (such as using $\hat{i}_{abcf}(t_k)$ to hot-start $\hat{i}_{abcf}(t_{k+1} - h)$). As passing the previous estimate forward to the next interval is critical for fast algorithm convergence, we further restrict h to be greater than T_{shift}, i.e., $T_{shift} \leq h \leq T$.

A second point to be made is that if there is a short to ground, then (5.37) is not valid because a short to ground allows some of the current to circumvent the neutral node in the stator windings. Thus we disallow shorts to ground in this discussion.

Thirdly, the minimization over $\hat{\sigma} : [t_k - T, t_k] \to [0, 1]$ denotes searching for all functions $\hat{\sigma}$ with domain $[t_k - T, t_k]$ and range in $[0, 1]$. The nominal ITSC EMHO problem requires an optimization of $\hat{i}_{abcf}(t_k - h) \in \mathbb{R}^4$ and $\hat{\sigma}$ over functions with range in $[0, 1]$. What has not been utilized in (5.34)-(5.37), is the steady state behavior inherent in the ITSC observer problem described in Section 5.4.1. The exploitation of the steady state behavior significantly reduces computation as discussed in the following section.

Finally, if the estimates $\hat{i}_{abcf}(t_k - h) = i_{abcf}(t_k - h)$ and $\hat{\sigma}$ are exact, then the cost function in (5.34) is zero since both the estimates and actual stator currents would be solutions to the same differential equations and have the same output function. Since (5.34) is nonnegative, the correct estimates are a minimizing solution to the cost function. If the only solution to (5.34) is the correct stator current and degree of fault, the observer problem is feasible. Feasibility has been discussed theoretically in Section 5.4.3 and demonstrated through simulation to follow.

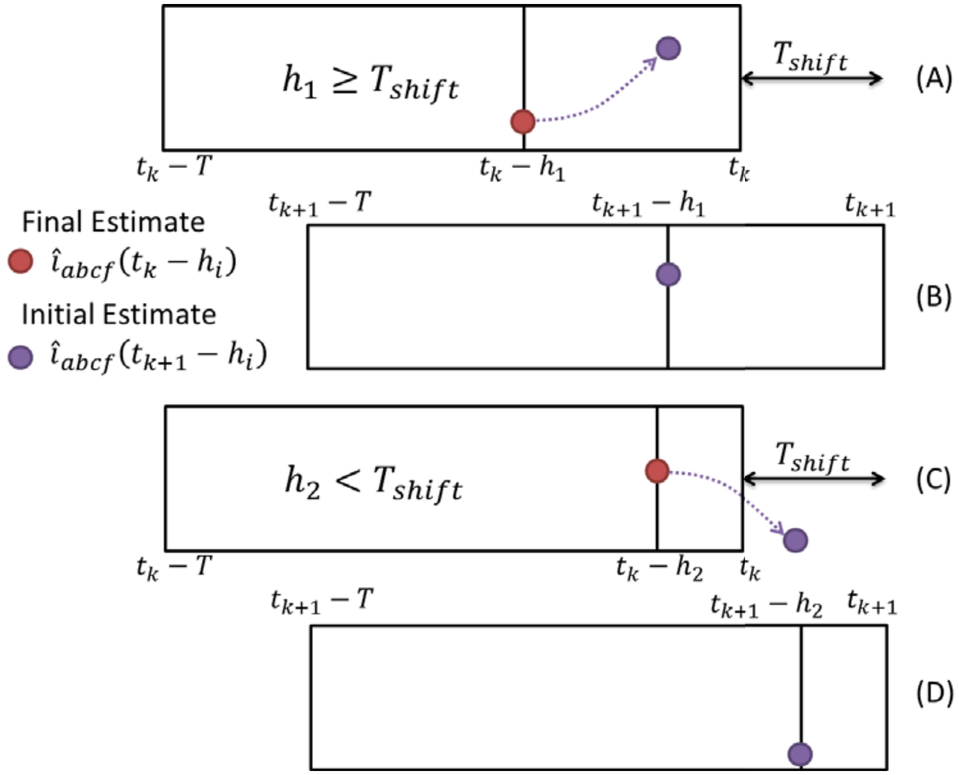

Fig. 5.9. (A) and (B) illustrate how the previous horizon estimate $\hat{i}_{abcf}(t_k - h_1)$ is integrated forward to hot-start $\hat{i}_{abcf}(t_{k+1} - h_1)$ when $h_1 \geq T_{shift}$. Notice, that the integration is within the interval $[t_k - T, t_k]$. (C) and (D) illustrate when $h_2 < T_{shift}$. Note, that the integration is not contained in $[t_k - T, t_k]$.

5.5 Practical Observer Implementation

The time constants associated with the stator currents in the SPMSM are much faster than (i) changes in the mechanical load and (ii) changes in the voltage or power commands. As a result, our analysis presupposes that the stator currents and voltages are in steady-state. Specifically, the steady-state stator currents and voltages are assumed to exhibit periodic sinusoidal behavior with frequency ω_r due to the sinusoidal back emf e_{abcf}. Note, this sinusoidal steady-state behavior occurs pre and post ITSC fault since in both cases the back emf e_{abcf} is sinusoidal.

How can we exploit the steady-state periodic sinusoidal behavior of the pre and post fault SPMSM to simplify the optimization problem in (5.34)? The approach is to explicitly impose the structure that $\hat{i}_{\zeta s}$, $\zeta = a, b, c, f$, are sinusoids with constant magnitudes and phase over subintervals of length t_{part}. The estimation of $\hat{i}_{\zeta s}$ can then be re-posed as estimating gains $\hat{I}_{q\zeta}$ and $\hat{I}_{d\zeta}$, $\zeta = a, b, c, f$, as per the following equations:

$$\hat{i}_{as} = \hat{I}_{qa}\cos(\theta_r) + \hat{I}_{da}\sin(\theta_r) \tag{5.43a}$$

$$\hat{i}_{bs} = \hat{I}_{qb}\cos(\theta_r - 2\pi/3) + \hat{I}_{db}\sin(\theta_r - 2\pi/3) \tag{5.43b}$$

$$\hat{i}_{cs} = \hat{I}_{qc}\cos(\theta_r + 2\pi/3) + \hat{I}_{dc}\sin(\theta_r + 2\pi/3) \tag{5.43c}$$

$$\hat{i}_{fs} = \hat{I}_{qf}\cos(\theta_r) + \hat{I}_{df}\sin(\theta_r). \tag{5.43d}$$

How does (5.43) simplify the optimization problem in (5.34)? The primary simplification is when solving the differential equation in (5.35). With stator and fault current estimates with the form of (5.43), the derivatives $\frac{d}{dt}\hat{i}_{\zeta s}$, $\zeta = a, b, c, f$, have the analytic form

$$\frac{d}{dt}\hat{i}_{as} = -\hat{I}_{qa}\omega_r\sin(\theta_r) + \hat{I}_{da}\omega_r\cos(\theta_r) \tag{5.44a}$$

$$\frac{d}{dt}\hat{i}_{bs} = -\hat{I}_{qb}\omega_r\sin(\theta_r - 2\pi/3) + \hat{I}_{db}\omega_r\cos(\theta_r - 2\pi/3) \tag{5.44b}$$

$$\frac{d}{dt}\hat{i}_{cs} = -\hat{I}_{qc}\omega_r\sin(\theta_r + 2\pi/3) + \hat{I}_{da}\omega_r\cos(\theta_r + 2\pi/3) \tag{5.44c}$$

$$\frac{d}{dt}\hat{i}_{fs} = -\hat{I}_{qf}\omega_r\sin(\theta_r) + \hat{I}_{df}\omega_r\cos(\theta_r). \tag{5.44d}$$

Hence, the differential equation in (5.35) can be replaced with an algebraic equation (with respect to estimated variables $\hat{I}_{q\zeta}$ and $\hat{I}_{d\zeta}$, $\zeta = a, b, c, f$). This greatly reduces the complexity and computational time required to compute $\hat{y}$ in the cost function.

To apply the assumption that the stator currents are fixed sinusoids over intervals of length t_{part}, we subdivide each horizon $[t_k - T, t_k]$ into n_{part} partitions of width t_{part}. We assume here that the horizon length T is a scalar multiple of t_{part}. With these partitions, the modified version of the ITSC EMHO estimates gains $\hat{I}_{q\zeta}^{(i)}$ and $\hat{I}_{d\zeta}^{(i)}$, $\zeta = a, b, c, f$, for each partition $i = 1, 2, \cdots, n_{part}$ of $[t_k - T, t_k]$.

The partitioning of the interval $[t_k - T, t_k]$ is also used to simplify estimating the degree of fault $\sigma(t)$. From a physical prospective, the ITSC faults occur when there is a electrical short between two locations within a stator winding. This electrical insulation failure happens at specific points and tends to have a binary behavior, i.e., short or no short. Consequently, the degree of fault $\sigma(t)$ is expected to be piecewise constant. This is exploited by considering the estimate $\hat{\sigma}$ to be constant over each partition of the interval $[t_k - T, t_k]$.

Over each partition of $[t_k - T, t_k]$, the last row of (5.35) becomes an algebraic equality constraint on the fault current estimate with respect to the gains $\hat{I}_{qf}$ and $\hat{I}_{df}$. This equality constraint is implemented in the simulation using a penalty function approach, i.e., adding a penalty function of the form

$$\int_{t_k - T}^{t_k} w_p \|\hat{v}_{fs}\|^2 dt, \tag{5.45}$$

to the cost function of (5.34). Here $w_p \in \mathbb{R}^+$ is a large weight and $\hat{v}_{fs}$ is the last row of (5.35), i.e.

$$\hat{v}_{fs} = R_s \hat{i}_{fs} + \sigma^2 L \frac{d}{dt}\hat{i}_{fs} + \sigma M \frac{d}{dt}\hat{i}_{bs} + \sigma M \frac{d}{dt}\hat{i}_{cs} + \sigma(1-\sigma) L \frac{d}{dt}\hat{i}_{as} + e_f, \tag{5.46}$$

with derivatives given in (5.44). Note that a feasible estimate for $\hat{i}_{fs}$ will satisfy $\hat{v}_{fs} \equiv 0$. Any nonzero value $\hat{v}_{fs}$ is penalized by the term in (5.45).

Another adaptation of the cost function in (5.34) is to add a positive definite weight matrix $Q \in \mathbb{R}^{6\times 6}$ to weight the output tracking error $y^M - \hat{y}$. With Q, the observer can be tuned to place the largest weight on a set of outputs which most directly affects the observability of the degree of fault σ. The modified cost function then has the form

$$\int_{t_k-T}^{t_k} \left(\left(y^M - \hat{y}\right)^\top Q \left(y^M - \hat{y}\right) + w_p \|\hat{v}_{fs}\|^2 \right) dt. \tag{5.47}$$

Incorporating the above ideas into the cost function over each horizon $[t_k - T, t_k]$, the practical version of the ITSC EMHO is

$$\begin{aligned} &\min_{\substack{\hat{I}^{(i)}_{q\zeta},\hat{I}^{(i)}_{d\zeta},\hat{\sigma}^{(i)} \\ \zeta=a,b,c,f \\ i=1,\cdots,n_{part}}} \int_{t_k-T}^{t_k} \left(\left(y^M - \hat{y}\right)^\top Q \left(y^M - \hat{y}\right) + w_p \|\hat{v}_{fs}\|^2 \right) dt \\ &\text{subject to: (5.35)–(5.44), (5.46).} \end{aligned} \tag{5.48}$$

The superscript (i) denotes the i^{th} partition of $[t_k - T, t_k]$. The constraints (5.35)–(5.44), and (5.46) are understood to apply to each partition.

Finally, to simplify the transition from one optimization problem to the next, the horizon is always uniformly shifted forward in time by t_{part}, i.e., $t_{k+1} = t_k + t_{part}$. This allows one last important modification to the ITSC EMHO concerning how estimates in preceding horizons are used to initialize or "hot-start" subsequent optimization problems. The scheme is illustrated in Figure 5.10. The method of partitioning each optimization horizon evenly has the advantage that estimates in some partitions of a previous horizon coincide with estimates of the current horizon. The partition $[t_{k+1} - t_{part}, t_{k+1}]$ does not coincide with the previous partition estimates. Thus, the estimate for $[t_k - t_{part}, t_k]$ is used to initialize the partition $[t_{k+1} - t_{part}, t_{k+1}]$ as shown in Figure 5.10.

5.6 Simulation Results

This section demonstrates the effectiveness of the ITSC EMHO. The three-phase SPMSM considered in this simulation has parameters given in Table 5.1. The SPMSM is simulated over $[0, 1]$ according to the following scenario: i) the rotor speed is a

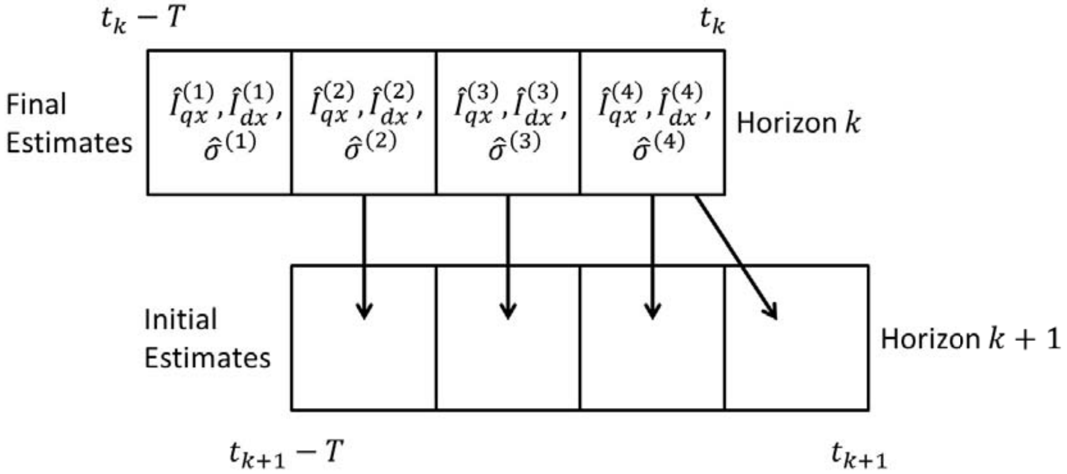
$t_k - T$
t_k
Final Estimates
$\hat{I}_{qx}^{(1)}, \hat{I}_{dx}^{(1)}, \hat{\sigma}^{(1)}$
$\hat{I}_{qx}^{(2)}, \hat{I}_{dx}^{(2)}, \hat{\sigma}^{(2)}$
$\hat{I}_{qx}^{(3)}, \hat{I}_{dx}^{(3)}, \hat{\sigma}^{(3)}$
$\hat{I}_{qx}^{(4)}, \hat{I}_{dx}^{(4)}, \hat{\sigma}^{(4)}$
Horizon k
Initial Estimates
Horizon $k+1$
$t_{k+1} - T$
t_{k+1}

Fig. 5.10. This figure shows how the final estimates for partitions in the horizon $[t_k - T, t_k]$ are used as initial estimates for the horizon $[t_{k+1} - T, t_{k+1}]$.

constant $\omega_m = 700$ rpm, ii) using current control the stator current (before and after fault) over $[0, 1]$ satisfies (current in Amperes)

$$\begin{aligned} i_{as} &= 50\cos(\theta_r) \\ i_{bs} &= 50\cos(\theta_r - 2\pi/3) \\ i_{cs} &= 50\cos(\theta_r + 2\pi/3), \end{aligned} \tag{5.49}$$

and iii) a fault of severity σ_f occurs at $t_{fault} = 0.5s$, i.e. $\sigma(t) = 0$ for $t \in [0, 0.5)$ and $\sigma(t) = \sigma_f$ for $t \in [0.5, 1]$. The scenario is simulated in MATLAB R2014b to construct the outputs

$$y^M = \begin{bmatrix} v_{ab}^M & v_{bc}^M & v_{ca}^M & i_{as}^M & i_{bs}^M & i_{cs}^M \end{bmatrix}^\top,$$

where the line to line voltages $v_{ab} = v_{as} - v_{bs}$, $v_{bc} = v_{bs} - v_{cs}$, and $v_{ca} = v_{cs} - v_{as}$ are computed using (5.12) given that the stator currents satisfy (5.49). To simulate the fault current i_{fs}, the differential equation in the last line of (5.12) is integrated using the *ode23t* function in MATLAB with the default integration settings. For EMHO implementation, the output y^M is sampled at a rate of $dt = 0.1$ms.

Table 5.1.
Simulation and SPMSM Parameters

Variable	Symbol	Value
Self Inductance	L	2.31 mH
Mutual Inductance	M	-1.15 mH
Magnet Strength	λ_m	0.267 Wb
Stator Resistance	R_s	137 mΩ
Poles	n_p	8
Bus Voltage	V_{bus}	500 V
Rotor Speed	ω_m	700rpm
Fault Time	t_{fault}	0.5s
Simulation Step Size	dt	0.1ms

The simulated ITSC EMHO has a horizon $T = 50$ms and two partitions of equal width, i.e. $t_{part} = 25$ms. The ITSC EMHO parameters are summarized in Table 5.2. To emphasize tracking the line to line voltage equations over stator current tracking, a weighting matrix $Q \in \mathbb{R}^6$ is added to the cost function, i.e. the cost function is given by

$$\int_{t_1-T}^{t_1} \left(y^M(t) - \hat{y}(t)\right)^\top Q \left(y^M(t) - \hat{y}(t)\right) dt, \tag{5.50}$$

where $Q = \text{diag}(10, 10, 10, 1, 1, 1)$. In addition, to enforce the constraint that $\hat{i}_{fs}$ satisfies the last row of (5.12), we add to the cost function (5.50) a penalty function of the form

$$\int_{t_1-T}^{t_1} w_p \|\hat{v}_{fs}\|^2 dt,$$

where $w_p = 1000$ and $\hat{v}_{fs}$ represents the last row on the right-hand side of (5.12), i.e.

$$\hat{v}_{fs} \triangleq \hat{\sigma} R_s \hat{i}_{fs} + \hat{\sigma}(1 - \hat{\sigma}) L \frac{d}{dt}\hat{i}_{as} + \hat{\sigma} M \frac{d}{dt}\hat{i}_{bs} + \hat{\sigma} M \frac{d}{dt}\hat{i}_{cs} + \hat{\sigma}^2 L \frac{d}{dt}\hat{i}_{fs} + e_f(\hat{\sigma}).$$

If $\hat{i}_{abcf}$ and $\hat{\sigma}$ are consistent with (5.12), $\hat{v}_{fs} \equiv 0$. As described in Section 5.5, the penalty function is used as an alternative method for enforcing this equality constraint.

Table 5.2.

ITSC EMHO Parameters

Variable	Symbol	Value
Number of Partitions	n_{part}	2
Horizon Width	T	50ms
Partition Width	t_{part}	25ms

The estimation error for reconstructing the fault current, i.e. $|i_{fs} - \hat{i}_{fs}|$ is shown in Figure 5.11 for four different degrees of fault $\hat{\sigma}_f = 0.01, 0.02, 0.05, 0.1$. The error $|i_{fs} - \hat{i}_{fs}|$ is scaled by $\max(i_{fs})$ which represents the amplitude of the steady state fault current i_{fs} for each degree of fault $\hat{\sigma}_f$. The estimation error for reconstructing

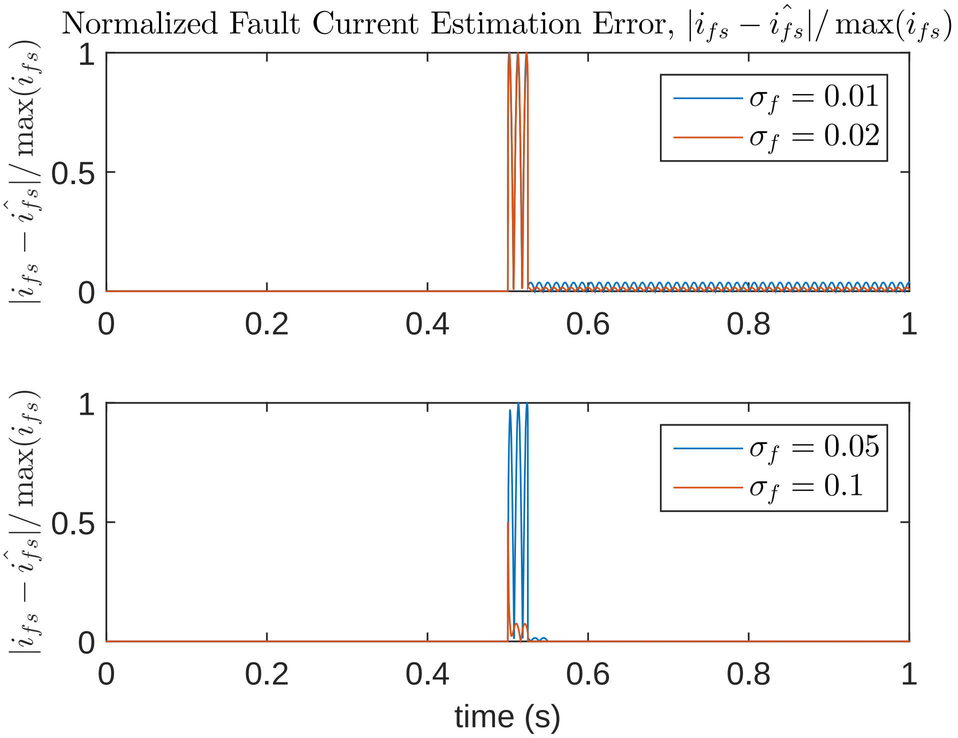

Fig. 5.11. The fault current reconstruction error $|i_{fs} - \hat{i}_{fs}|$ is simulated for four levels of fault, $\sigma_f = 0.01, 0.02, 0.05$, and 0.1. The figure is normalized by $\max(i_{fs})$ which represents the magnitude of the steady state fault current i_{fs} for each degree of fault. In each simulation, the fault occurs at 0.5s.

the degree of fault is shown in Figure 5.12 for each of the four different degrees of fault $\hat{\sigma}_f$. The estimation error for i_{as}, i_{bs}, and i_{cs} are not included since these are also measured variables and hence the estimation error is on the order of 10^{-6} (tolerance of the optimization).

It is clear from Figure 5.11, that after one partition of 25ms, the fault current estimate $\hat{i}_{fs}$ is within 5% of the actual fault current i_{fs}. Similarly, the degree of fault estimation error is within 0.001 after one partition of 25ms as shown in Figure 5.12. This "bump" in the estimates right after the fault occurs is caused by an initial guess which is far from the new level of fault. However, the next optimization window improves the estimate of the degree of fault and fault current and converges quickly.

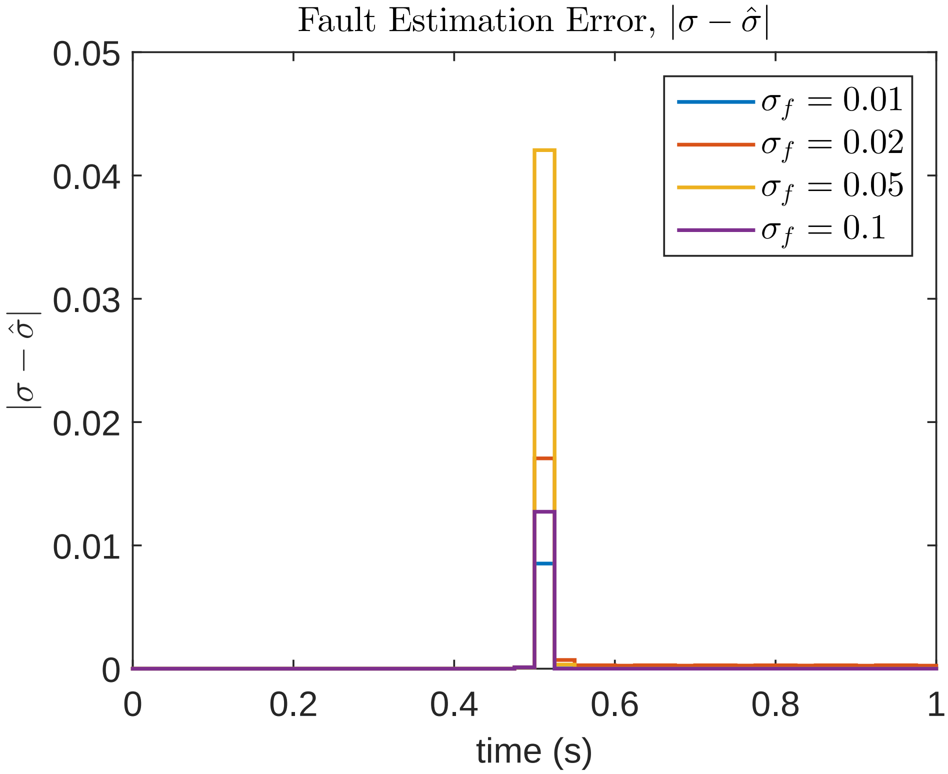

Fig. 5.12. The degree of fault reconstruction error $|\sigma - \hat{\sigma}|$ is simulated for four levels of fault, $\sigma_f = 0.01, 0.02, 0.05$, and 0.1. In each simulation, the fault occurs at 0.5s.

The ability to improve on the previous estimates is a consequence of the manner in which estimates from previous partitions are used to "hot-start" subsequent partitions. The reader can recall that the initial states are passed from one partition to hot-start the next as illustrated in Figure 5.10.

The ITSC EMHO has additional applications beyond fault detection. One such application is fault-tolerant control schemes where the estimate for the degree of fault can be used to determine "safe" operating conditions after a fault has occurred. The next section explores a fault-tolerant power flow control application for a hybrid electric vehicle, such as the Toyota Prius. This fault detection scheme also has applications for both fault detection and fault mitigating control in heavy hybrid vehicles. The application to heavy hybrid vehicles is discussed in Section 5.7.

5.7 Application: Heavy Hybrid Vehicles

According to Harrington and Krupnick at Resources for the Future, the National Highway Traffic Safety Administration mandated the first-ever federal requirements for improving fuel economy in heavy-duty commercial vehicles in 2011 [65]. The focus on reducing fuel consumption in heavy vehicles on the highway has also had an impact in the off-road heavy vehicle industry. Leading companies of off-road vehicles, such as Caterpillar and John Deere, have released hybrid versions of off-road construction and forestry equipment. Although fuel prices have dropped in the past few years, the environmental, economic, and regulatory influences on heavy vehicle design promise continued growth in the area of heavy hybrid technology.

Electric machines are a common component in heavy hybrid vehicles, such as the Caterpillar D7E Dozer [66] and the John Deere 644k Hybrid Wheel Loader [67]. The Deere 644k Hybrid Wheel Loader uses two permanent-magnet synchronous machines (PMSM), one primarily as a generator and the other as a transmission drive. Due to the tough working conditions of these vehicles, the areas of safety, robust performance, and reduced repair costs are key marketable features. In the event of a fault within the electric machine, fault detection, and fault-tolerant control in the heavy hybrid vehicles can improve each of these marketable features. The detection of an interturn short circuit (ITSC) fault in the stator windings of the PMSM is critical to maintaining the safe operation of these vehicles. In this section we outline the impact of this work on ITSC fault detection in PMSM to the industry of heavy hybrid vehicles.

5.7.1 Increased Scale

The simulation in Section 5.6 demonstrates the effective use of the ITSC fault detection scheme using an embedded moving horizon observer (EMHO). The surface PMSM (SPMSM) explored in Section 5.6 has a maximum power of about 30kW. Heavy hybrid drivetrains require motors on the scale of hundreds of kilowatts. Fortunately, the size of the motors does not effect the structure of the mathematical model for SPMSM or the structure of the EMHO used to detect ITSC faults. As such, the same techniques developed for ITSC fault detection for SPMSM can be applied directly to SPMSM in heavy hybrid vehicles.

5.7.2 Interior PMSM

Many heavy hybrid vehicle manufacturers prefer interior PMSM (IPMSM) over the surface mounted counterparts. Although the control of SPMSM is simpler, the IPMSM has manufacturing advantages as well as some additional control techniques. The magnets in the IPMSM are embedded in the rotor laminations. This allows for permanent magnets which are rectangular and easier to produce in addition to avoiding the problem of attaching magnets to the surface of the rotor. Another key advantage to the IPMSM, is that the iron in the rotor can be magnetized between the magnetic poles and provide the so-called reluctance torque. The reluctance torque is especially useful at producing power at high speeds when the bus voltage limits the output power. Despite the advantages of the IPMSM, stators in IPMSM and SPMSM are similar and can suffer from the same ITSC winding faults. In this subsection, we will introduce a stator voltage model from the IPMSM and discuss the applications of the SPMSM fault detection work.

The unfaulted interior PMSM (neglecting leakage inductance) can be modeled by [51]

$$v_{abc} = R_s i_{abc} + \frac{d}{dt}[L_{AB}(\theta_r) i_{abc}] + e_{abc} \tag{5.51}$$

where $v_{abc} = [v_{as}, v_{bs}, v_{cs}]^\top$, $i_{abc} = [i_{as}, i_{bs}, i_{cs}]^\top$, R_s denotes the stator resistance in each coil, θ_r and ω_r are the electrical position and speed of the rotor, the back emf e_{abc} satisfies

$$e_{abc} = \begin{bmatrix} e_a \\ e_b \\ e_c \end{bmatrix} = \lambda_m \omega_r \begin{bmatrix} \cos(\theta_r) \\ \cos(\theta_r - 2\pi/3) \\ \cos(\theta_r + 2\pi/3) \end{bmatrix}, \tag{5.52}$$

and the inductance matrix $L_{AB}(\theta_r)$ has the form

$$L_{AB} = \begin{bmatrix} L_A + L_B \cos 2\theta_r & -\frac{1}{2}L_A + L_B \cos 2\left(\theta_r - \frac{\pi}{3}\right) & -\frac{1}{2}L_A + L_B \cos 2\left(\theta_r + \frac{\pi}{3}\right) \\ -\frac{1}{2}L_A + L_B \cos 2\left(\theta_r - \frac{\pi}{3}\right) & L_A + L_B \cos 2\left(\theta_r - \frac{2\pi}{3}\right) & -\frac{1}{2}L_A + L_B \cos 2(\theta_r + \pi) \\ -\frac{1}{2}L_A + L_B \cos 2\left(\theta_r + \frac{\pi}{3}\right) & -\frac{1}{2}L_A + L_B \cos 2(\theta_r + \pi) & L_A + L_B \cos 2\left(\theta_r + \frac{2\pi}{3}\right) \end{bmatrix}. \tag{5.53}$$

In the case of the SPMSM, the sinusoidal inductance terms $L_B \cos(\cdot)$ is zero.

Modeling an IPMSM with ITSC faults is an area of future research. From the developments in Section 5.3, we expect that the back emf e_{abc} and the inductance matrix $L_{AB}(\theta_r)$ will become functions of the degree of fault $\sigma \in [0, 1]$. The key difference is modeling how L_A and L_B change after a fault has occured. Despite the current lack of an ITSC fault model for the IPMSM, the fault detection framework and observer structure can be extended to the IPMSM pending the model for the ITSC faults. The structure for the IPMSM ITSC fault detection problem is shown in Figure 5.13.

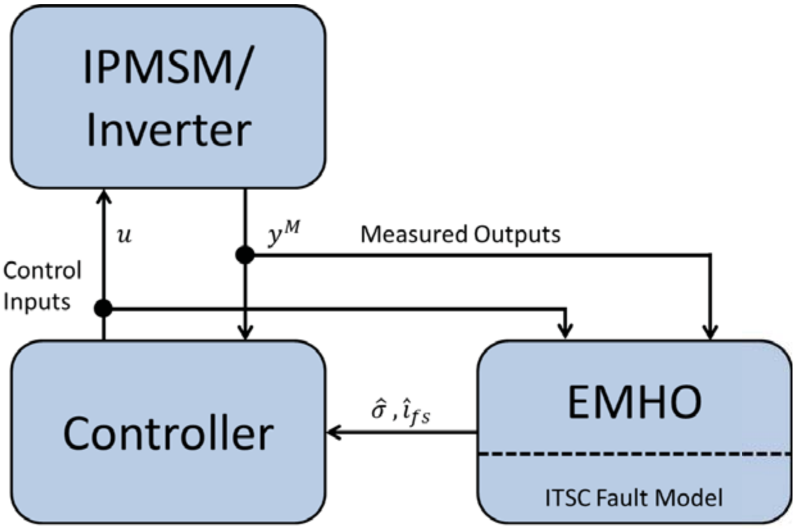

Fig. 5.13. Fault detection scheme for IPMSM with estimated degree of fault $\hat{\sigma}$ and estimated fault current $\hat{i}_{fs}$.

5.7.3 Fault-Tolerant Control

After an ITSC fault has occurred, the eddy loop acts as an induction heater within the stator windings. For heavy vehicles, oil-cooled stator windings improve the ability to cool the stator windings after an ITSC fault and may allow for a reduced operating condition for short periods of time. This reduced operating condition, or “limp-home” mode, can allow vehicles in remote work sites to reach a safe location for repairs. Since off-road heavy vehicles can spend considerable time in remote locales, the ability to “limp home” provides a significant advantage.

Similar to the fault-tolerant scheme for the Prius, we propose using the ITSC fault model of the PMSM (whether surface or interior magnets) to generate fault-tolerant controls, operating limits, and efficiency curves at various degrees of fault σ. The method for constructing these efficiency curves and fault-tolerant controls are discussed in Section 5.3 and [62]. The basic structure for the fault tolerant control with a high-level power flow controller is shown in Figure 5.14.

5.8 Future Work

In this chapter, we have developed a moving horizon observer to detect ITSC faults in surface permanent magnet synchronous machines. A simplified version of the observer is validated through simulation. Application to supervisory control in heavy hybrid vehicles is also developed.

The development of an ITSC fault model for interior permanent magnet synchronous machines is an area of future research. With this model, a moving horizon observer can be developed to detect ITSC faults in much the same manner as presented in this paper. Another area of future research is validating the fault models and fault detection scheme in physical devices. The model validation of the fault model for surface permanent magnet machines was started in [57], but verification of the interior permanent magnet machine fault model is still incomplete.

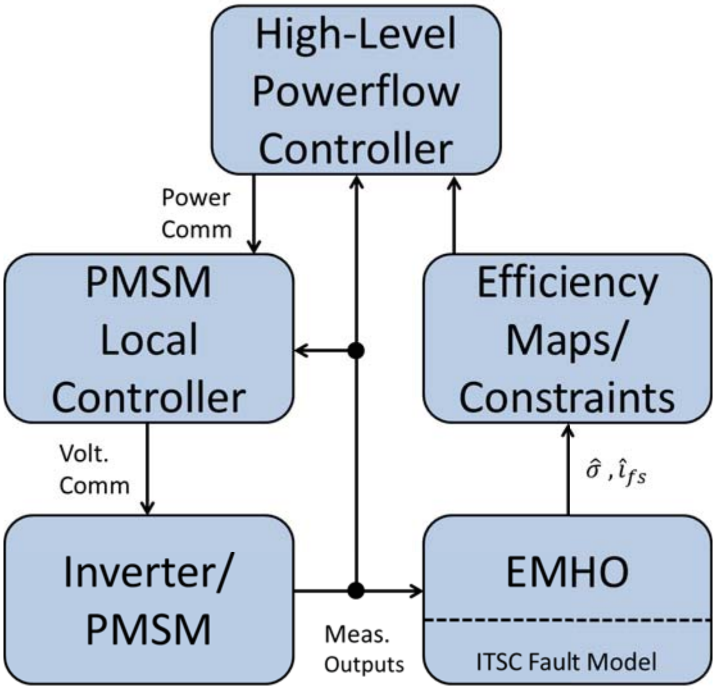
High-Level Powerflow Controller
Power Comm
PMSM Local Controller
Efficiency Maps/ Constraints
Volt. Comm
$\hat{\sigma}$, $\hat{\imath}_{fs}$
Inverter/ PMSM
EMHO
Meas. Outputs
ITSC Fault Model

Fig. 5.14. Fault-tolerant control scheme with estimated degree of fault $\hat{\sigma}$ and estimated fault current $\hat{i}_{fs}$.

Optimizing the computational time for the moving horizon observer is also an area of future work. In part, this requires optimizing the number of horizons, horizon width, and the search algorithm. This is a dual formulation to the problem in model predictive control of determining optimal horizon parameters. As computational power in vehicles continues to increase and processor prices decrease, we expect that using moving horizon observers for fault detection will become an increasingly attractive solution to improving electric machine safety, reliability, and repair costs.

Appendix A

When an ITSC fault occurs in two phases simultaneously, say phase-a and phase-b, there exists fault currents i_{fs}^a and i_{fs}^b within each of the two fault loops. The degree of fault in each phase is denoted σ_a and σ_b. For ease of notation we define $\tau_a = 1 - \sigma_a$ and $\tau_b = 1 - \sigma_b$. The stator voltage model is given by

$$v_{abcf} = R_f(\sigma_a, \sigma_b) i_{abcf} + L_f(\sigma_a, \sigma_b) \frac{d}{dt} i_{abcf} + e_{abcf}(\sigma_a, \sigma_b),$$

where

$$v_{abcf} = \begin{bmatrix} v_{as} & v_{bs} & v_{cs} & 0 & 0 \end{bmatrix}^\top, \quad i_{abcf} = \begin{bmatrix} i_{as} & i_{bs} & i_{cs} & i_{fs}^a & i_{fs}^b \end{bmatrix}^\top,$$

$$R_f(\sigma_a, \sigma_b) = \begin{bmatrix} \tau_a R_s & 0 & 0 & 0 & 0 \\ 0 & \tau_b R_s & 0 & 0 & 0 \\ 0 & 0 & R_s & 0 & 0 \\ 0 & 0 & 0 & \sigma_a R_s & 0 \\ 0 & 0 & 0 & 0 & \sigma_b R_s \end{bmatrix},$$

$$L_f(\sigma_a, \sigma_b) = \begin{bmatrix} \tau_a^2 L & \tau_a \tau_b M & \tau_a M & \tau_a \sigma_a L & \tau_a \sigma_b M \\ \tau_a \tau_b M & \tau_b^2 L & \tau_b M & \tau_b \sigma_a M & \tau_b \sigma_b L \\ \tau_a M & \tau_b M & L & \sigma_a M & \sigma_b M \\ \tau_a \sigma_a L & \tau_b \sigma_a M & \sigma_a M & \sigma_a^2 L & \sigma_a \sigma_b M \\ \tau_a \sigma_b M & \tau_b \sigma_b L & \sigma_b M & \sigma_a \sigma_b M & \sigma_b^2 L \end{bmatrix},$$

and

$$e_{abcf}(\sigma_a,\sigma_b) = \lambda_m\omega_r \begin{bmatrix} \tau_a\cos(\theta_r) \\ \tau_b\cos(\theta_r - 2\pi/3) \\ \cos(\theta_r + 2\pi/3) \\ \sigma_a\cos(\theta_r) \\ \sigma_b\cos(\theta_r - 2\pi/3) \end{bmatrix}$$

An ITSC fault occurs in all three phases, there is an additional fault current i^c_{fs} and degree of fault σ_c. The stator voltage model extends is an extension of the two-phase stator voltage model. The electromechanical power couples the electrical and mechanical components of the SPMSM as per the following equation

$$T_e\omega_m = P_a + P_b + P_c + P^a_f + P^b_f = J\omega_m\dot{\omega}_m + B\omega_m^2 + T_L\omega_m,$$

where $P^a_f = \sigma_a\lambda_m\omega_r i^a_{fs}\cos(\theta_r)$, $P^b_f = \sigma_b\lambda_m\omega_r i^b_{fs}\cos(\theta_r - 2\pi/3)$, and $P_\zeta = e_\zeta i_{\zeta s}$ for $\zeta = a, b, c$. The total inverter power, $P_{inv} = P_{inv,a} + P_{inv,b} + P_{inv,c}$, is given by

$$\begin{aligned} P_{inv} &= \tau_a R_s i_{as}^2 + \tau_b R_s i_{bs}^2 + R_s i_{cs}^2 + \sigma_a R_s (i^a_{fs})^2 + \sigma_b R_s (i^b_{fs})^2 \\ &\quad + \frac{d}{dt}\Upsilon_f(\sigma_a,\sigma_b) + P_a + P_b + P_c + P^a_f + P^b_f, \end{aligned}$$

where $\Upsilon_f(\sigma_a,\sigma_b) = i_{abcf}^\top L_f(\sigma_a,\sigma_b) i_{abcf}$.

Appendix B

For the LTV system

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) \tag{5.54}$$

$$y(t) = C(t)x(t) + D(t)u(t), \tag{5.55}$$

the output $y(t)$ can be expressed as a function of the initial state x_0 and input $u(t)$ as per

$$y(t) = C(t)\Phi(t,t_0)x_0 + C(t)\int_{t_0}^{t}\Phi(t,q)B(q)u(q)dq + D(t)u(t), \tag{5.56}$$

where $\Phi(t, t_0)$ is the state transition matrix [13]. Using (5.56), the left-hand side of the strong observability condition in (5.25) can be expressed as

$$\begin{aligned}
&\int_{t-T}^{t} \|y(x(t), u(t)) - y(x'(t), u(t))\|^2 dt \\
&\qquad\qquad = \int_{t-T}^{t} \|C(q)\Phi(q, t-T)x_0 - C(q)\Phi(q, t-T)x_0'\|^2 dq \\
&\qquad\qquad = (x_0 - x_0')^\top W_O(t, t-T)(x_0 - x_0') \\
&\qquad\qquad \geq \lambda_{min}(W_O(t, t-T))\|x_0 - x_0'\|_2^2
\end{aligned}$$

where $W_O(t, t-T)$ is the observability Grammian for (5.54). The LTV system (5.54) is observable over $[t-T, t]$ if and only if the observability Grammian $W_O(t, t-T)$ is positive definite, i.e., if and only if $\lambda_{min}(W_O(t, t-T)) > 0$ [13]. Setting $\gamma = \lambda_{min}(W_O(t, t-T))$, the strong observability condition in (5.25) is thus equivalent to observability for LTV systems.

6. FUTURE WORK

This thesis develops the feasibility conditions for switched systems state and mode sequence reconstruction, a robust observability metric and algorithm, and the embedded moving horizon observer (EMHO). This chapter outlines future research topics in the area of switched system observability and observer design. This chapter is divided into future work for robust observability and EMHO design.

6.1 Robust Observability Extensions

The $\mathcal{P}$–robustness algorithm in Chapter 3, specifically Algorithm 1, requires several assumptions that can be relaxed in the future. First, the Assumption 3.1 requires surjectivity of $L_{uV\mathcal{S}}$. For real perturbations, $\mathcal{S} \subset \mathbb{R}^{n\times m}$, surjectivity of $L_{uV\mathcal{S}}$ is believed to be unnecessarily restrictive. As described in [38], surjectivity of $L_{uV\mathcal{S}}$ is sufficient for satisfying a certain regularity condition required to guarantee algorithm convergence. However, surjectivity of $L_{uV\mathcal{S}}$ may not be necessary for all $\mathcal{P}$–robustness algorithms. For example, consider when all system matrices, property matrices, and perturbation matrices are real, i.e., $M \in \mathbb{R}^{n\times m}$ and $\mathcal{P}, \mathcal{S} \subset \mathbb{R}^{n\times m}$. In this case, the singular vectors u and V of $M - R - \delta M$ are real and $L_{uV\mathcal{S}} : \mathcal{S} \to \mathbb{C}^{1\times(m-n+1)}$ is clearly not surjective. However, experimentally it appears that Algorithm 1 converges in this case.

One possibility for relaxing the surjectivity assumption is requiring that $L_{uV\mathcal{S}}(\mathcal{S})$ contains $L_{uV}(M - \mathcal{P} - \mathcal{S})$. In this case, for each $R \in \mathcal{P}$ and $\delta M \in \mathcal{S}$, there exists another perturbation $\delta M' \in \mathcal{S}$ such that

$$L_{uV}(M - R - \delta M - \delta M') = 0.$$

This is exactly the property used in steps 5 and 6 of Algorithm 1. To guarantee convergence using the condition $L_{uV}(M - \mathcal{P} - \mathcal{S}) \subset L_{uV\mathcal{S}}(\mathcal{S})$ also requires modifying the proof of the necessary conditions in Theorem 3.2, which also utilize the surjectivity assumption.

The second assumption that can be relaxed is Assumption 3.2 which requires each property matrix $R \in \mathcal{P}$ to be full rank, i.e., $\operatorname{rank} R = n$. This assumption guarantees that Algorithm 1 converges to a finite property matrix R_*. This condition is sufficient but not necessary for the existence of finite optimal property matrices. One approach to removing this restriction is to instead compute the P_{max}-bounded $\mathcal{P}$–robustness problem which restricts each property matrix R to satisfy $\|R\|_F \leq P_{max}$. This would allow modifying Algorithm 1 to again guarantee that R_* is bounded and converges.

Another key area of future work is proving the rate of convergence for Algorithm 1. Based on the work of [28] and [68], it is expected that the convergence rate will be locally quadratic, given appropriate assumptions, since Algorithm 1 is related to Newton's Method. One apparent challenge with proving the rate of convergence is connecting the difference of the Lyapunov-like energy functions $P_{k+1} - P_k$ to convergence of δM_k and R_k in the presence of the adaptive weight g_k.

In addition, it is believed that Algorithm 1 may be modified to compute the smallest rank reducing perturbation δM_* with respect to the spectral norm, i.e., replacing $\|\delta M\|_F$ with $\sigma_1(\delta M)$ in (3.4) from Definition 3.1. Extending Algorithm 1 to include the spectral norm in addition to the Frobenius norm metric will unify the robustness property literature. One approach may be to modify Steps 5 and 6 of Algorithm 1 to find $\delta\widetilde{M}_k$ and $\delta\overline{M}_k$ to be the smallest matrices (with respect to the spectral norm) such that

$$L_{uV\mathcal{S}}(\delta\widetilde{M}_k) = L_{uV}(\delta M_k - \Delta\widetilde{R}_k) \tag{6.1}$$

$$L_{uV\mathcal{S}}(\delta\overline{M}_k) = L_{uV}(M - R_k - \Delta\overline{R}_k - \delta M_k). \tag{6.2}$$

A similar modification to $\Delta\widetilde{R}_k$ to reduce $\sigma_1(\delta\widetilde{M}_k)$ will also be required. One idea for computing the spectral norm as opposed to the Frobenius norm may be found in replacing the linear operator L_{uV} with $L_{u_1V_1}$ where u_1 is the first lsv of $M - \delta M_k - R_k$ and V_1 contains the first rsv and $n+1$ through m of $M - \delta M_k - R_k$. The connection between the linear operator $L_{u_1V_1\mathcal{S}}$ and $\sigma_1(\delta\widetilde{M})$ and $\sigma_1(\delta\overline{M})$ is not fully explored, but may allow for simple and elegant generalization.

The final area for extending the $\mathcal{P}$–robustness framework and algorithm is to connect the distance to the nearest SMS unobservable switched system to convergence properties for the EMHO. The distance to the nearest SMS unobservable switched system, is clearly related to the degree of distinguishability and the class of ϵ-mode distinguishing inputs (Definition 4.4). The key is to relate the extended observability Gramian in (2.39) to the distance to the nearest SMS unobservable switched system. The closer the switched system is to unobservable, the closer the observability Gramian in (2.39) is to singularity which is intimately linked to EMHO convergence.

6.2 EMHO Extensions

The EMHO is a new an open topic for future research. This thesis introduced the basic framework and a few basic convergence properties. Three main topics appear "ripe" for future work: 1) proving convergence of the EMHO algorithm at each step, 2) extending EMHO to switched nonlinear systems, and 3) developing a robust EMHO in the presence of disturbances and sensor noise.

Section 4.2.6 proves convergence given that at each iteration a state and mode estimate tuple $(\hat{x}_{i+1}, \hat{v}_{k+1})$ satisfies

$$V_E(\hat{x}_{i+1}, \hat{v}_{i+1}; t_i, t_{i+1}) \leq \beta V_E(\hat{x}_i, \hat{v}_i; t_{i-1}, t_i)) \tag{6.3}$$

for a fixed $\beta \in (0, 1)$. Although, SMS observability guarantees that such a tuple $(\hat{x}_{i+1}, \hat{v}_{k+1})$ exists, the algorithm for computing this tuple is not yet specified. Experimentally, optimization programs such as sequential quadratic programs and interior-point algorithms have been used to solve for the tuple $(\hat{x}_{i+1}, \hat{v}_{k+1})$. It is expected that given certain structural properties on the switched system (say SLTI for example) it can be proven that the optimization problem to solve for the tuple $(\hat{x}_{i+1}, \hat{v}_{k+1})$ is locally convex. Moreover, if it is locally convex then one may be able to explicitly bound the number of iterations required to satisfy (6.3).

Another extension of the EMHO convergence is the application to switched nonlinear systems. Proving convergence for the EMHO applied to switched nonlinear systems is nontrivial. Details of this extension are beyond the scope of this section, but an interested reader should explore the uniform reconstructability condition in [22, Equation 3.4].

Possibly the most practical extension to the EMHO is to consider the effect of disturbances and sensor noise on the EMHO. In the presence of disturbances and sensor noise, either the disturbance and sensor noise must be estimated (likely to be intractable) or convergence to perfect state and mode estimates should be relaxed. One such relaxation is to compute a minimum L_2 output tracking error that is solvable, i.e., a lower bound $V_{min} \leq V_E(\hat{x}_i, \hat{v}_i; t_{i-1}, t_i)$. This lower bound V_{min} represents the level of uncertainty for state and mode reconstruction. Given the lower bound V_{min}, one can construct a bound on the mode estimation error $v(t_i) - \hat{v}(t_i)$ and state tracking error $e(t_i)$ as $i \to \infty$. Some results for switched MHO convergence in the presence of noise can be found in [48], but the connection to the EMHO is still an area of active research.

REFERENCES

REFERENCES

[1] S. Wei, M. Zefran, K. Uthaichana, and R. A. DeCarlo, "Hybrid Model Predictive Control for Stabilization of Wheeled Mobile Robots Subject to Wheel Slippage," in *Proc. 2007 IEEE Int. Conf. Robot. Autom.*, no. April. IEEE, apr 2007, pp. 2373–2378. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4209438

[2] S. Wei, K. Uthaichana, M. Zefran, and R. Decarlo, "Hybrid model predictive control for the stabilization of wheeled mobile robots subject to wheel slippage," *IEEE Trans. Control Syst. Technol.*, vol. 21, no. 6, pp. 2181–2193, 2013. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs{_}all.jsp?arnumber=6579667

[3] S. Liu, X. Feng, D. Kundur, T. Zourntos, and K. L. Butler-Purry, "Switched system models for coordinated cyber-physical attack construction and simulation," in *2011 IEEE 1st Int. Work. Smart Grid Model. Simulation, SGMS 2011*, 2011, pp. 49–54.

[4] R. Meyer, R. DeCarlo, P. Meckl, C. Doktorcik, and S. Pekarek, "Hybrid model predictive power flow control of a fuel cell-battery vehicle," in *Am. Control Conf.*, 2011, pp. 2725–2731[1]. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp={&}arnumber=5991428{&}isnumber=5989965

[5] J. Neely, S. Pekarek, R. DeCarlo, and N. Vaks, "Real-time hybrid model predictive control of a boost converter with constant power load," in *2010 Twenty-Fifth Annu. IEEE Appl. Power Electron. Conf. Expo.* IEEE, feb 2010, pp. 480–490. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5433628

[6] R. Vidal, A. Chiuso, S. Soatto, and S. Sastry, "Observability of Linear Hybrid Systems," in *Hybrid Syst. Comput. Control.* Springer Berlin/Heidelberg, 2003, pp. 526–539.

[7] D. Gomez-Gutierrez, A. Ramirez-Trevino, J. Ruiz-Leon, and S. Di Gennaro, "Observability of Switched Linear Systems: A geometric approach," in *49th IEEE Conf. Decis. Control.* IEEE, dec 2010, pp. 5636–5642. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5717936

[8] ——, "On the Observability of Continuous-Time Switched Linear Systems Under Partially Unknown Inputs," *IEEE Trans. Automat. Contr.*, vol. 57, no. 3, pp. 732–738, mar 2012. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6007052

[9] D. Gomez-Gutierrez, S. Celikovsky, A. Ramirez-Trevino, J. Ruiz-Leon, and S. Di Gennaro, "Sliding mode observer for Switched Linear Systems," in *2011 IEEE Int. Conf. Autom. Sci. Eng.* IEEE, aug 2011, pp. 725–730. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6042494

[10] D. Gómez-Gutiérrez, S. Čelikovský, A. Ramírez-Treviño, and B. Castillo-Toledo, "On the Observer Design Problem for Continuous-Time Switched Linear Systems with Unknown Switchings," *J. Franklin Inst.*, 2015. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0016003215000642

[11] A. Balluchi, L. Benvenuti, M. D. D. Benedetto, and A. L. Sangiovanni-Vincentelli, "Design of Observers for Hybrid Systems," in *Hybrid Syst. Comput. Control*, ser. Lecture Notes in Computer Science, C. Tomlin and M. Greenstreet, Eds. Springer Berlin/Heidelberg, 2002, vol. 2289, pp. 76–89.

[12] S. C. Bengea and R. a. DeCarlo, "Optimal control of switching systems," *Automatica*, vol. 41, no. 1, pp. 11–27, jan 2005. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0005109804002237

[13] R. A. DeCarlo, *Linear Systems: A State Variable Approach with Numerical Implementation*. Englewood Cliffs, New Jersey: Prentice Hall, 1989.

[14] W. M. Wonham, *Linear Multivariable Control: A Geometric Approach*, 2nd ed. Springer-Verlag New York, 1979.

[15] G. Basile and G. Marro, "Controlled and conditioned invariant subspaces in linear system theory," *J. Optim. Theory Appl.*, vol. 3, no. 5, pp. 306–315, 1969.

[16] ——, "On the observability of linear, time-invariant systems with unknown inputs," *J. Optim. Theory Appl.*, vol. 3, no. 6, pp. 410–415, 1969.

[17] ——, *Controlled and conditioned invariant subspaces in linear system theory.* Prentice Hall Englewood Cliffs, 1992.

[18] M. Babaali and G. J. Pappas, "Observability of switched linear systems in continuous time," *Hybrid Syst. Comput. Control*, vol. 3414, no. March, pp. 103–117, 2005. [Online]. Available: http://www.springerlink.com/index/150UC4MJQW7AJAE9.pdf

[19] W. Rudin, *Real and complex analysis*, 3rd ed. New York: McGraw-Hill Book Co., 1987.

[20] S. C. Johnson, R. A. DeCarlo, and M. Zefran, "Set-transition observability of switched linear systems," in *2014 Am. Control Conf.* IEEE, jun 2014, pp. 3267–3272. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6858960

[21] E. Kreindler and P. Sarachik, "On the concepts of controllability and observability of linear systems," *Autom. Control. IEEE ...*, 1964. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs{_}all.jsp?arnumber=1105665

[22] H. Michalska and D. Mayne, "Moving horizon observers and observer-based control," *IEEE Trans. Automat. Contr.*, vol. 40, no. 6, pp. 995–1006, jun 1995. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=388677

[23] C.-T. Chen, *Linear System Theory and Design*, 3rd ed. New York, NY, USA: Oxford University Press, Inc., 1998.

[24] R. Eising, “Between controllable and uncontrollable,” *Syst. Control Lett.*, vol. 4, no. 5, pp. 263–264, jul 1984. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0167691184800353

[25] M. Wicks and R. DeCarlo, “Computing the distance to an uncontrollable system,” *IEEE Trans. Automat. Contr.*, vol. 36, no. 1, pp. 39–49, 1991. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=62266

[26] C. Kenney and A. J. Laub, “Controllability and stability radii for companion form systems,” *Math. Control. Signals, Syst.*, vol. 1, no. 3, pp. 239–256, oct 1988. [Online]. Available: http://link.springer.com/10.1007/BF02551286

[27] M. Wicks, “Matrix rank-robustness problems in systems and control: theory and computation,” Ph.D. dissertation, Purdue University, 1992.

[28] M. A. Wicks and R. A. DeCarlo, “Rank Robustness of Complex Matrices with Respect to Real Perturbations,” *SIAM J. Matrix Anal. Appl.*, vol. 15, no. 4, pp. 1182–1207, oct 1994. [Online]. Available: http://epubs.siam.org/doi/abs/10.1137/S0895479891194297

[29] S. R. Khare, H. K. Pillai, and M. N. Belur, “Computing the radius of controllability for state space systems,” *Syst. Control Lett.*, vol. 61, no. 2, pp. 327–333, feb 2012. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0167691111002957

[30] S. C. Johnson and R. A. DeCarlo, “Bounding the distance to the nearest unobservable switched linear time-invariant system,” in *2015 Am. Control Conf.* Chicago, IL: IEEE, 2015, pp. 1083 – 1088.

[31] J. Clotet and M. D. Magret, “Upper Bounds for the Distance between a Controllable Switched Linear System and the Set of Uncontrollable Ones,” *Math. Probl. Eng.*, vol. 2013, no. 4, pp. 1–9, 2013. [Online]. Available: http://www.hindawi.com/journals/mpe/2013/948147/

[32] G. Hu and E. Davison, “Real Controllability/Stabilizability Radius of LTI Systems,” *IEEE Trans. Automat. Contr.*, vol. 49, no. 2, pp. 254–257, feb 2004. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1266782

[33] N. K. Son and D. D. Thuan, “The structured controllability radii of higher order systems,” *Linear Algebra Appl.*, vol. 438, no. 6, pp. 2701–2716, mar 2013. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0024379512008026

[34] S. Lam and E. J. Davison, “Computation of the Real Controllability Radius and Minimum-Norm Perturbations of Higher-Order, Descriptor, and Time-Delay LTI Systems,” *IEEE Trans. Automat. Contr.*, vol. 59, no. 8, pp. 2189–2195, aug 2014. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6705617

[35] J. V. Burke, A. S. Lewis, and M. L. Overton, “Pseudospectral Components and the Distance to Uncontrollability,” *SIAM J. Matrix Anal. Appl.*, vol. 26, no. 2, pp. 350–361, jan 2004. [Online]. Available: http://epubs.siam.org/doi/abs/10.1137/S0895479803433313

[36] M. Gu, E. Mengi, M. L. Overton, J. Xia, and J. Zhu, "Fast Methods for Estimating the Distance to Uncontrollability," *SIAM J. Matrix Anal. Appl.*, vol. 28, no. 2, pp. 477–502, jan 2006. [Online]. Available: http://epubs.siam.org/doi/abs/10.1137/05063060X

[37] C. He, "Estimating the distance to uncontrollability: A fast method and a slow one," *Syst. Control Lett.*, vol. 26, no. 4, pp. 275–281, nov 1995. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/0167691195000233

[38] M. Wicks and R. DeCarlo, "Computing robustness of system properties with respect to structured real matrix perturbations," in *[1992] Proc. 31st IEEE Conf. Decis. Control.* IEEE, 1992, pp. 1909–1914. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=371098

[39] B. De Moor and S. Boyd, "Analytic Properties of Singular Values and Vectors," *ESAT-SISTA Rep.*, vol. 28, p. 17, 1989.

[40] D. G. Luenberger, *Optimization by Vector Space Methods.* John Wiley & Sons, Inc., 1969.

[41] J. Brewer, "Kronecker products and matrix calculus in system theory," *IEEE Trans. Circuits Syst.*, vol. 25, no. 9, pp. 772–781, sep 1978. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1084534

[42] D. S. Bernstein, *Matrix Mathematics: Theory, Facts, and Formulas with Application to Linear System Theory.* Princeton University Press, 2005.

[43] A. J. Laub, *Kronecker Products.* Society for Industrial and Applied Mathematics, Philadelphia, PA, 2005.

[44] T. Kato, *A short introduction to perturbation theory for linear operators.* Springer New York, 2012.

[45] A. Alessandri and P. Coletta, "Design of Luenberger Observers for a Class of Hybrid Linear Systems," in *HSCC '01 Proc. 4th Int. Work. Hybrid Syst. Comput. Control*, 2001, pp. 7–18.

[46] S. Pettersson, "Designing Switched Observers for Switched Systems Using Multiple Lyapunov Functions and Dwell-Time Switching," in *Anal. Des. Hybrid Syst. 2006.* Elsevier, 2006, pp. 18–23. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/B9780080446134500070

[47] A. Tanwani, H. Shim, and D. Liberzon, "On observability for switched linear systems: characterization and observer design," *IEEE Trans. Automat. Contr.*, vol. 58, no. 4, 2013.

[48] G. Ferrari-Trecate, D. Mignone, and M. Morari, "Moving horizon estimation for hybrid systems," *IEEE Trans. Automat. Contr.*, vol. 47, pp. 1663–1676, 2002.

[49] J. R. Munkres, *Topology*, 2nd ed., ser. Featured Titles for Topology Series. Prentice Hall, Incorporated, 2000.

[50] M. Babaali and M. Egerstedt, "Observability of Switched Linear Systems," in *Hybrid Syst. Comput. Control. LNCS.* Springer Verlag, 2004, pp. 48—-63.

[51] P. Krause, O. Wasynczuk, S. Sudhoff, and P. Pekarek, *Analysis of Electric Machinery and Drive Systems*, 3rd ed. Piscataway, NJ: Wiley-IEEE Press, 2013.

[52] C. Lai, A. Balamurali, V. Bousaba, K. L. V. Iyer, and N. C. Kar, "Analysis of stator winding inter-turn short-circuit fault in interior and surface mounted permanent magnet traction machines," in *2014 IEEE Transp. Electrif. Conf. Expo.* IEEE, jun 2014, pp. 1–6. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6861775

[53] A. Gandhi, T. Corrigan, and L. Parsa, "Recent advances in modeling and online detection of stator interturn faults in electrical motors," *IEEE Trans. Ind. Electron.*, vol. 58, no. 5, pp. 1564–1575, 2011.

[54] L. Romeral, J. C. Urresty, J. R. Riba Ruiz, and A. Garcia Espinosa, "Modeling of surface-mounted permanent magnet synchronous motors with stator winding interturn faults," *IEEE Trans. Ind. Electron.*, vol. 58, no. 5, pp. 1576–1585, 2011.

[55] R. T. Meyer, R. a. DeCarlo, and S. Pekarek, "Hybrid Model Predictive Power Management of a Battery-Supercapacitor Electric Vehicle," *Asian J. Control*, vol. 18, no. 1, pp. 150–165, jan 2016. [Online]. Available: http://doi.wiley.com/10.1002/asjc.553http://doi.wiley.com/10.1002/asjc.1259

[56] K. Uthaichana, R. DeCarlo, S. Bengea, S. Pekarek, and M. Zefran, "Hybrid optimal theory and predicitive control for power management in hybrid electric vehicle," *J. Nonlinear Syst. Appl*, vol. 2, no. 1-2, pp. 96—-110, 2011.

[57] K.-H. Kim, D.-U. Choi, B.-G. Gu, and I.-S. Jung, "Fault model and performance evaluation of an inverter-fed permanent magnet synchronous motor under winding shorted turn and inverter switch open," *IET Electr. Power Appl.*, vol. 4, no. 4, p. 214, 2010.

[58] D. Luenberger, *Introduction to dynamic systems: theory, models, and applications.* Wiley, 1979.

[59] A. Alessandri, M. Baglietto, G. Battistelli, and V. Zavala, "Advances in moving horizon estimation for nonlinear systems," *49th IEEE Conf. Decis. Control*, pp. 5681–5688, dec 2010. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5718126

[60] Y. Guo and B. Huang, "Moving horizon estimation for switching nonlinear systems," *Automatica*, vol. 49, pp. 3270–3281, 2013.

[61] S. Bengea, K. Uthaichana, M. Žefran, and R. DeCarlo, "Optimal Control of Switching Systems via Embedding into Continuous Optimal Control Problem," in *Control Syst. Handbook, Second Ed.*, W. S. Levine, Ed. CRC Press 2011, 2010, ch. 31, pp. 31–1–31–23.

[62] R. T. Meyer, S. C. Johnson, R. A. DeCarlo, and S. Pekarek, "Supervisory powerflow control of Toyota Prius subject to stator winding faults," *In Preperation*, 2016.

[63] R. A. DeCarlo and R. Saeks, *Interconnected Dynamical Systems*, ser. Electrical and Computer Engineering. New York: CRC Press, 1981.

[64] R. T. Meyer, M. Zefran, and R. A. DeCarlo, “A Comparison of the Embedding Method With Multiparametric Programming, Mixed-Integer Programming, Gradient-Descent, and Hybrid Minimum Principle-Based Methods,” *IEEE Trans. Control Syst. Technol.*, vol. 22, no. 5, pp. 1784–1800, sep 2014. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6733440

[65] W. Harrington and A. Krupnick, “Improving Fuel Economy in Heavy Duty Vehicles,” Issue Brief 12-01. Resources for the Future, Tech. Rep., 2012.

[66] Caterpillar, “D7E.” [Online]. Available: http://www.cat.com/en{_}US/products/rental/equipment/dozers/medium-dozers/18429156.html

[67] Deere & Company, “644K Hybrid Wheel Loader.” [Online]. Available: https://www.deere.com/en{_}US/products/equipment/wheel{_}loaders/644k{_}hybrid/644k{_}hybrid.page

[68] M. A. Wicks and R. A. DeCarlo, “Computing Most Nearly Rank-Reducing Structured Matrix Perturbations,” *SIAM J. Matrix Anal. Appl.*, vol. 16, no. 1, pp. 123–137, jan 1995. [Online]. Available: http://epubs.siam.org/doi/abs/10.1137/S089547989222758X

VITA

VITA

Scott C. Johnson was born in Fort Wayne, Indiana in 1989. Mr. Johnson attended grade school and high school in Caldwell, Idaho. He received a B.S. degree in Mathematics/Physics from The College of Idaho, Caldwell, ID in 2007. He also received a M.S. degree in Electrical Engineering from Purdue University, West Lafayette, Indiana. Currently he is a Ph.D. candidate at Purdue University and will receive his degree in August 2016.

Mr. Johnson's primary research interests include switched system observer design, observability, and robust observers.