# A Robust Algorithm for Online Switched System Identification *

Zhe Du , Necmiye Ozay , and Laura Balzano

Electrical and Computer Engineering, University of Michigan
{*zhedu,necmiye,girasole*}*@umich.edu*

**Abstract**

In this paper, we consider the problem of online identification of Switched AutoRegressive eXogenous (SARX) systems, where the goal is to estimate the parameters of each subsystem and identify the switching sequence as data are obtained in a streaming fashion. Previous works in this area are sensitive to initialization and lack theoretical guarantees. We overcome these drawbacks with our two-step algorithm: (i) every time we receive new data, we first assign this data to one candidate subsystem based on a novel robust criterion that incorporates both the residual error and an upper bound of subsystem estimation error, and (ii) we use a randomized algorithm to update the parameter estimate of chosen candidate. We provide a theoretical guarantee on the local convergence of our algorithm. Though our theory only guarantees convergence with a good initialization, simulation results show that even with random initialization, our algorithm still has excellent performance. Finally, we show, through simulations, that our algorithm outperforms existing methods and exhibits robust performance.

***Keywords:*** System identification, Online identification algorithm, Convergence analysis

## 1 Introduction

A SARX system is a special type of hybrid system composed of multiple subsystems/modes each with different parameters. At each time step only one subsystem is dominating and the dominant subsystem may switch over time. Given system inputs and outputs at each time step, our goal is to identify the switching sequence (discrete states) as well as to estimate the parameters of the subsystems every time we receive new data. This is a problem involving both clustering and estimation.

In additional to applications in adaptive control, SARX system identification has been applied to video and texture segmentation [13, 10, 11]. Due to the autoregressive nature of SARX model, it can also be applied to earthquake record analysis [7], brain electrical activity mapping [9], meteorological objects identification [3], and financial time series analysis [4].

### 1.1 Prior Work

There have been many studies on the switched system identification problem in the offline/batch setting. A type of algebraic method was proposed in [14], which uses Veronese embedding to decouple the task of estimating the system parameters and switching sequence, and an exact solution is provided when the process and data are noise-free. Furthermore, the case when system orders are not necessarily

*This work is supported by DARPA grant N66001-14-1-4045, DARPA grant 16-43-D3M-FP-037 and NSF Grant ECCS-1508943.

equal or known is discussed in [8]. For systems with noise and measurements corrupted by outliers, [10] extends the algebraic method by converting it to a rank minimization problem that is relaxed to a semi-definite program. Methods utilizing sparsity are proposed in [1, 11].

As opposed to the offline/batch setting, where we have access to all the data at once, there are many problems in which the data appears in a streaming (online) fashion. That is, at each time step, we receive data with which we need to identify current dominant subsystem as well as give the latest estimate of the system parameters. Note that naively employing an offline algorithm by using all the data in the past at each step would be computationally intractable. The majority of online algorithms use a two-step approach that alternates between determining the switching sequence and updating the parameter estimates. The work in [13] is one of the first to study online identification of switched systems using an extension of the offline algebraic method [14]. In the algorithms proposed in [2], [5], candidate estimates are built for each of the subsystems first. Then, every time a new data point arrives, the discrete state is determined by assigning the data to one of the candidates according to some criterion, and then the estimate of chosen candidate is updated with the new data. The algorithm in [2] first identifies the discrete states based on prior or posterior residual error, and then updates the estimate using recursive least squares. The algorithm in [5] identifies the discrete states by minimizing prior residual error similarly and then update the estimates with a modified Outer Bounding Ellipsoid (OBE) algorithm.

### 1.2 Contributions and Outline

We observe that in two-step algorithms, choosing a candidate based on minimum residual error can be sensitive to candidate initializations, since when a new subsystem dominates, it might "take-over" a partially convergent candidate estimate if there is no candidate yet closer to its true parameters.

The main contribution of our paper is a more robust two-step algorithm that can effectively overcome this issue. We initialize candidate estimates for each of the subsystems. Every time we receive new data, we determine the discrete state by assigning this data to one of candidates based on a robust criterion that incorporates both residual error and an upper bound of estimation error. After we assign the data to a candidate, we update the selected candidate using a variant of the randomized Kaczmarz algorithm proposed in [12] or normalized least mean squares (NLMS) [6]. We provide partial and local convergence results for our algorithm. In our partial convergence analysis we assume that we can always make correct assignments, i.e. identify the discrete state correctly, thus the parameter estimation updated for the candidates can be treated as if we are using data from a single subsystem. In local convergence analysis, we assume all candidate estimates have "good enough" initializations, and show that with some probability no misassignment will ever be made and prove the convergence of parameter estimates. Our numerical simulations verify the convergence result and show obvious improvements of our algorithm over state of the art.

The paper is organized as follows: in Section 2, we present the problem formulation of online SARX system identification; Section 3 briefly discusses the drawbacks of existing algorithms; Section 4 introduces our algorithm; Section 4 gives the theoretical analyses of our algorithm; some discussions and extensions are provided in Section 6; simulation evaluation are given in Section 7.

## 2 Problem Formulation

### 2.1 SARX System

A SARX system is defined by the following expression:

$$y_t = \sum_{j=1}^{n_a} a_j(z_t) y_{t-j} + \sum_{k=1}^{n_c} c_k(z_t) u_{t-k} + n_t \tag{1}$$

where $u_t \in \mathbb{R}$ and $y_t \in \mathbb{R}$ are the input and output of the system, and $n_t \in \mathbb{R}$ is an additive noise term. The discrete state $z_t \in \{1, \dots, m\} \equiv [m]$ indexes the dominant/active subsystem at time $t$, and $\{z_t\}_t$ denotes the switching sequence. Coefficients $\{a_j(z_t)\}_{j=1}^{n_a}$ and $\{c_j(z_t)\}_{j=1}^{n_c}$ are the parameters of subsystem $z_t$. Let $\boldsymbol{\phi}_{y,t} = [y_{t-1}, \dots, y_{t-n_a}]^\intercal, \boldsymbol{\phi}_{u,t} = [u_{t-1}, \dots, u_{t-n_c}]^\intercal, \boldsymbol{\phi}_t = [\boldsymbol{\phi}_{y,t}^\intercal, \boldsymbol{\phi}_{u,t}^\intercal]^\intercal$, and furthermore, let $\mathbf{w}_{z_t} = [a_1(z_t), \dots, a_{n_a}(z_t), c_1(z_t), \dots, c_{n_c}(z_t)]^\intercal$. With this notation, the SARX system dynamics (1) can be written in vector form:

$$y_t = \mathbf{w}_{z_t}^\intercal \boldsymbol{\phi}_t + n_t \ . \tag{2}$$

Let $n = n_a + n_c$ be the system order, which can also be viewed as the ambient dimension of our problem.

### 2.2 Assumptions

In this work, we make the following assumptions, where Assumption 1 and the noise upper bound in Assumption 2 are needed for the algorithm to work. (Case where noise is unbounded is discussed in Section 6.2.) Assumption 2 to Assumption 5 are mainly for analysis purposes.

**Assumption 1.** *The model orders $n_a$, $n_c$ on RHS of (1), and the number of subsystems, $m$, are known.*

**Assumption 2.** *The noise $n_t$ is random with $\mathbb{E}[n_t] = 0$ and $\mathbb{E}[n_t^2] = \sigma_n^2$. $|n_t| \le n_{\max}$ and $n_{\max}$ is known. $n_t$ is independent of input $u_t$.*

**Assumption 3.** *For all $t$, $\|\boldsymbol{\phi}_t\| \le \phi_{\max}$. We also assume a lower bound on the SNR: for all $t$, $\frac{\|\boldsymbol{\phi}_t\|}{|n_t|} \ge S_{\min}$.*

**Assumption 4.** *There exists $s_{\max} \ge s_{\min} > 0$ such that $\forall\ S \subset \mathbb{N}^+$ with cardinality $N_R$ (defined in Section 4),*

$$s_{\min}^2 I_n \preceq \sum\nolimits_{t \in S} \boldsymbol{\phi}_t \boldsymbol{\phi}_t^\intercal \preceq s_{\max}^2 I_n \ . \tag{3}$$

**Assumption 5.** *If subsystem $i$ generates data pair $\{\boldsymbol{\phi}_t, y_t\}$, i.e. $y_t = \mathbf{w}_i^\intercal\ \boldsymbol{\phi}_t + n_t$, then $\forall j \ne i$, $|\mathbf{w}_j^\intercal \boldsymbol{\phi}_t - \mathbf{w}_i^\intercal \boldsymbol{\phi}_t| \ge \psi$.*

Assumption 4 is similar to persistent excitation conditions in the literature, and it plays a critical role in the convergence rate. Assumption 5 guarantees there is no ambiguous data, since if data pair $\{\boldsymbol{\phi}_t, y_t\}$ satisfies both $y_t = \mathbf{w}_1^\intercal \boldsymbol{\phi}_t$ and $y_t = \mathbf{w}_2^\intercal \boldsymbol{\phi}_t$, then even with the true parameters $\mathbf{w}_1$ and $\mathbf{w}_2$, we cannot tell which system generates $y_t$.

### 2.3 Goal

The goal of online system identification is as follows. After we collect the data pair $\{\boldsymbol{\phi}_t, y_t\}$ at each time step, we want to identify discrete state $z_t$ and estimate parameters of the subsystem that generates $y_t$.

## 3 Drawbacks of Existing Algorithms

Existing algorithms e.g. [2], [5], commonly have a two-step structure after candidate estimate for each of the subsystem is initialized: (1) every time new data is available, it is assigned to the candidate with minimum prior/posterior residual error; (2) the parameter estimate of the chosen candidate is updated with this data. We will show that using only residual error as the criterion to assign data can be unreliable.

Fig. 1 shows a toy example of what could go wrong with the above mentioned algorithms. There are 3 subsystems, and red circles illustrate their true parameter vectors in the ambient space; the

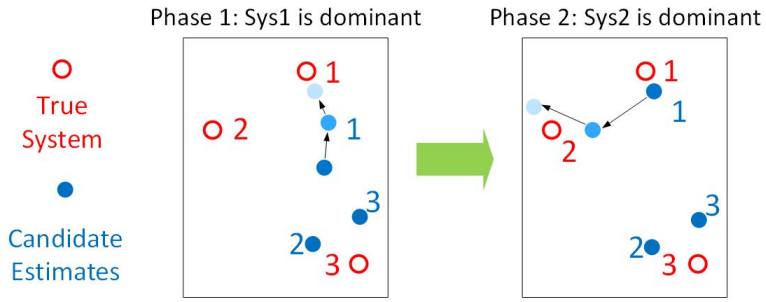


Figure 1: Demonstration of potential drawback of existing algorithms.

3 candidate estimates are initialized at the three dark blue points in the left box. Assume from $t = 1$ to 10, subsystem 1 is dominant (left box), and from $t = 11$ to 20, subsystem 2 is dominant (right box). Considering the positions and true and estimated parameters, it's *likely* that from $t = 1$ to 10 data generated by subsystem 1 will be assigned to candidate 1 since it's the closest candidate. When $t = 10$, candidate 1 is an improved estimate of system 1 parameters, given by the light blue point in the left box. At time $t = 11$, subsystem 2 becomes dominant. Considering the current positions of all candidates, candidate 1 is still closest to subsystem 2, so it's *likely* that data generated by subsystem 2 will also be assigned to candidate 1, and we could expect candidate 1 will start to drift from subsystem 1 parameter values toward subsystem 2, given by the trajectory in the right box. In this sense, all previous efforts used to let candidate 1 learn subsystem 1 will be wasted.

In our algorithm, the basic idea to solve this drawback is to take the accuracy of the candidate estimates into account and be more cautious when assigning data to candidates with higher accuracy. The details will be discussed in Section 4.1.

# 4 Our Algorithm

In this paper we propose Algorithm 1 for online identification of SARX models. This is also a two-step algorithm, but with an improved data assignment to consider not only the residual but also system estimation accuracy. This section gives an overview of the algorithm steps.

Lines 1 to 4 show initialization. $\widehat{\mathbf{w}}_{i,0}$ is the initial estimate for candidate $i$, and $c_i$ is number of assignments to candidate $i$. $\mathbf{\Phi}^R_{i,t} \in \mathbb{R}^{n \times N_R}$, $\mathbf{y}^R_{i,t} \in \mathbb{R}^{N_R}$, $\mathbf{\Phi}^C_{i,t} \in \mathbb{R}^{n \times N_C}$, $\widehat{\mathbf{W}}^C_{i,t} \in \mathbb{R}^{n \times N_C}$, $\mathbf{h}^C_{i,t} \in \mathbb{R}^{N_C}$, $\epsilon^u_{i,t}$ are the corresponding window variables for candidate $i$ at time $t$, which will be explained in details later in this section. $N_R$ and $N_C$ are the number of columns of $\mathbf{\Phi}^R_{i,t}$ and $\mathbf{\Phi}^C_{i,t}$ respectively, which are also the window lengths for the randomized Kaczmarz algorithm and error upper bound estimation respectively.

At each time step, via Lines 6 to 13, we assign the data to one of the candidates using a new criterion to determine the discrete state. Then, we update the chosen candidate estimate using an idea similar to the randomized Kaczmarz algorithm in [12] in Lines 14 to Line 24.

## 4.1 Making Assignment/Identifying the Discrete State

With data pair $\{\boldsymbol{\phi}_t, y_t\}$, we compute the normalized residual error $r_i$ for each candidate in Line 8, where $\widehat{\mathbf{w}}_{i,t-1}$ is the estimate of candidate $i$ at time $t-1$. We then compute the potential new estimate $\widetilde{\mathbf{w}}_{i,t}$ for each candidate if we were to use $\{\boldsymbol{\phi}_t, y_t\}$ to update $\widehat{\mathbf{w}}_{i,t-1}$.

The assignment criterion is given in Line 11. The criterion has two components: the first term is the normalized residual error $r_i$ and the second term measures whether $\widetilde{\mathbf{w}}_{i,t}$ has a larger estimation error than $\widehat{\mathbf{w}}_{i,t-1}$. The variables $\alpha$, $\beta$ and $\nu$ are tuning parameters. Variable $\epsilon^u_{i,t-1}$ is an estimate of upper bound on the magnitude of candidate $i$'s estimation error $\boldsymbol{\epsilon}_{i,t-1} \equiv \mathbf{w} - \widehat{\mathbf{w}}_{i,t-1}$ with respect to

**Algorithm 1:** Our Main Algorithm

Initialize $N_R, N_C (N_R \geq n, N_C \geq N_R^2), \alpha, \beta, \nu$

**for** $i = 1, \ldots, m$ **do**

  $\widehat{\mathbf{w}}_{i,0} = \mathbf{0}_{n\times 1}, c_i = 0,\ \mathbf{\Phi}^R_{i,t} = \mathbf{0}_{n\times N_R},\ \mathbf{y}^R_{i,t} = \mathbf{0}_{N_R\times 1},$

  $\mathbf{\Phi}^C_{i,t} = \mathbf{0}_{n\times N_C},\ \widehat{\mathbf{W}}^C_{i,t} = \mathbf{0}_{n\times N_C},\ \mathbf{h}^C_{i,t} = \mathbf{0}_{N_C\times 1},\ \epsilon^u_{i,0} = \infty$

**for** $t = 1, 2, \ldots$ **do**

  Receive $\{\boldsymbol{\phi}_t, y_t\}$.

  Compute normzlized residual errors and potential new estimates for all candidates:

    $r_i = |y_t - \widehat{\mathbf{w}}^\intercal_{i,t-1}\boldsymbol{\phi}_t| \cdot \|\boldsymbol{\phi}_t\|^{-1} \quad \forall i \in [m]$

    $\widetilde{\mathbf{w}}_{i,t} = \widehat{\mathbf{w}}_{i,t-1} - \|\boldsymbol{\phi}_t\|^{-2}\boldsymbol{\phi}_t(\widehat{\mathbf{w}}^\intercal_{i,t-1}\boldsymbol{\phi}_t - y_t) \quad \forall i \in [m]$

  Choose a candidate to assign data:

    $\widehat{z}_t = \arg\min_i\ r_i \cdot \max\left(1, \alpha\frac{\|\widetilde{\mathbf{w}}_{i,t} - \widehat{\mathbf{w}}_{i,t-1}\|}{2(\epsilon^u_{i,t-1}+\nu)}\right)^\beta$

  Update counter and window varaibles:

    $c_{\widehat{z}_t} = c_{\widehat{z}_t} + 1$

    $\mathbf{\Phi}^R_{\widehat{z}_t,t} = [\mathbf{\Phi}^R_{\widehat{z}_t,t-1}[:, 2{:}\text{end}], \boldsymbol{\phi}_t],\ \mathbf{y}^R_{\widehat{z}_t,t} = [\mathbf{y}^R_{\widehat{z}_t,t-1}[2{:}\text{end}]; y_t]$

  Update estimate of chosen candidate:

    **if** $c_{\widehat{z}_t} < N_R$ **then**

      $\boldsymbol{\phi}^*_t = \boldsymbol{\phi}_t, y^*_t = y_t, \eta^*_t = \|\boldsymbol{\phi}_t\|^{-2}$

    **else**

      Sample $l_t \in [N_R]$ w.p. $\|\mathbf{\Phi}^R_{\widehat{z}_t,t}[:, l_t]\|^2 / \|\mathbf{\Phi}^R_{\widehat{z}_t,t}\|_F^2$

      $\boldsymbol{\phi}^*_t = \mathbf{\Phi}^R_{\widehat{z}_t,t}[:, l_t],\ y^*_t = \mathbf{y}^R_{\widehat{z}_t,t}[l_t],\ \eta^*_t = \|\boldsymbol{\phi}^*_t\|^{-2}$

    $\widehat{\mathbf{w}}_{\widehat{z}_t,t} = \widehat{\mathbf{w}}_{\widehat{z}_t,t-1} - \eta^*_t\boldsymbol{\phi}^*_t(\widehat{\mathbf{w}}^\intercal_{\widehat{z}_t,t-1}\boldsymbol{\phi}^*_t - y^*_t)$

  Update error upper bound and window variables:

    $\mathbf{\Phi}^C_{\widehat{z}_t,t}, \widehat{\mathbf{W}}^C_{\widehat{z}_t,t}, \mathbf{h}^C_{\widehat{z}_t,t}, \epsilon^u_{\widehat{z}_t,t}$=**UpdateUpperBound**

    $\forall i \neq \widehat{z}_t, \{\widehat{\mathbf{w}}_{i,t}, \mathbf{\Phi}^R_{i,t}, \mathbf{y}^R_{i,t}, \mathbf{\Phi}^C_{i,t}, \widehat{\mathbf{W}}^C_{i,t}, \mathbf{h}^C_{i,t}, \epsilon^u_{i,t}\} = \{\widehat{\mathbf{w}}_{i,\tau}, \mathbf{\Phi}^R_{i,\tau}, \mathbf{y}^R_{i,\tau}, \mathbf{\Phi}^C_{i,\tau}, \widehat{\mathbf{W}}^C_{i,\tau}, \mathbf{h}^C_{i,\tau}, \epsilon^u_{i,\tau}\}_{\tau=t-1}$

*some* true system parameter $\mathbf{w}$. The main difference between our algorithm and previous two-step algorithms mentioned in Section 3 is the incorporation of the second term, which makes assignment more robust.

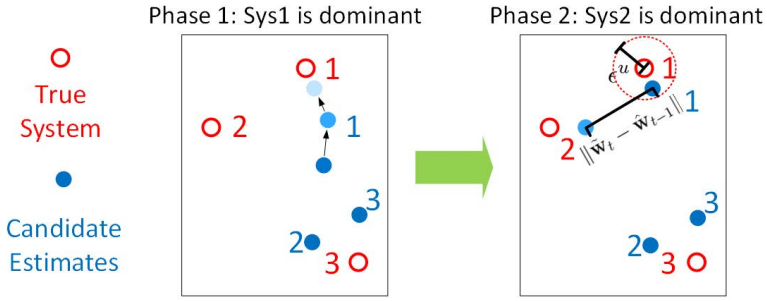


Figure 2: Idea of our algorithm

The idea behind the criterion is straightforward: letting $\alpha{=}1, \beta{=}1, \nu{=}0$, and replacing $\epsilon^u_{i,t-1}$ with $\|\boldsymbol{\epsilon}_{i,t-1}\|$, the second part of the criterion becomes $\max\left(1, \frac{\|\widetilde{\mathbf{w}}_{i,t} - \widehat{\mathbf{w}}_{i,t-1}\|}{2\|\boldsymbol{\epsilon}_{i,t-1}\|}\right)$. The numerator $\|\widetilde{\mathbf{w}}_{i,t} - \widehat{\mathbf{w}}_{i,t-1}\|$

is the magnitude of variation if we update the estimate $\widehat{\mathbf{w}}_{i,t-1}$ of candidate $i$ with $\{\boldsymbol{\phi}_t, y_t\}$. We could see that if the estimation error of $\widetilde{\mathbf{w}}_{i,t}$ doesn't increase compared with the error of $\widehat{\mathbf{w}}_{i,t-1}$, then we must have $\|\widetilde{\mathbf{w}}_{i,t} - \widehat{\mathbf{w}}_{i,t-1}\| \leq 2\|\boldsymbol{\epsilon}_{i,t-1}\|$. And if $\|\widetilde{\mathbf{w}}_{i,t} - \widehat{\mathbf{w}}_{i,t-1}\| > 2\|\boldsymbol{\epsilon}_{i,t-1}\|$, then the estimation error must get larger. Therefore, $\max\left(1, \frac{\|\widetilde{\mathbf{w}}_{i,t}-\widehat{\mathbf{w}}_{i,t-1}\|}{2\|\boldsymbol{\epsilon}_{i,t-1}\|}\right)$ works against candidates whose error would increase if we update using $\{\boldsymbol{\phi}_t, y_t\}$. The max operator shows that as long as $\|\widetilde{\mathbf{w}}_{i,t} - \widehat{\mathbf{w}}_{i,t-1}\| \leq 2\|\boldsymbol{\epsilon}_{i,t-1}\|$, we don't penalize any further. Since we don't know the true estimation error $\|\boldsymbol{\epsilon}_{i,t-1}\|$, we replace it with its estimated upper bound $\epsilon^u_{i,t-1}$ (we will show under some conditions, this is a valid upper bound in Theorem 9) computed in Algorithm 2.

This idea is illustrated with Fig. 2. Consider the same experimental setup as the toy example in Section 3. At time $t = 11$, candidate 1 will be within the ball region (denoted with the red dotted circle) estimated by upper bound $\epsilon^u$. Once the potential update magnitude of candidate 1 exceeds the diameter $2\epsilon^u$ of this region, which implies its estimation accuracy will become worse if we update candidate 1 with this data, so candidate 1 should be penalized when making the assignment.

We note that $\|\widetilde{\mathbf{w}}_{i,t} - \widehat{\mathbf{w}}_{i,t-1}\| \leq 2\|\boldsymbol{\epsilon}_{i,t-1}\|$ is only a necessary but not sufficient condition to ensure non-increasing estimation error. Even though we are relying on a necessary condition, our algorithm empirically achieves significantly improved performance over previous algorithms.

### 4.2 Candidate Estimate Updates

After the assignment is made, from Line 14 to Line 21 we update the estimate of the candidate $\widehat{z}_t$ to which the data has been assigned. Our updating approach is based on the randomized Kaczmarz method with a sliding window of data. We define window variables $\boldsymbol{\Phi}^R_{i,t}, \mathbf{y}^R_{i,t}$ to store previous $N_R$ data $\{\boldsymbol{\phi}, y\}$ assigned to candidate $i$. If we have collected $N_R$ data, i.e. $c_{\widehat{z}_t} \geq N_R$, we update the estimate with randomly picked historical data $\{\boldsymbol{\phi}^*_t, y^*_t\}$; otherwise we simply update using current data $\{\boldsymbol{\phi}^*_t, y^*_t\} = \{\boldsymbol{\phi}_t, y_t\}$. The idea behind the update rule is that we project the current estimate $\widehat{\mathbf{w}}_{\widehat{z}_t,t-1}$ onto the solution space of $\{\boldsymbol{\phi}^*_t, y^*_t\}$ such that $y^*_t = \widehat{\mathbf{w}}^\intercal_{\widehat{z}_t,t}\boldsymbol{\phi}^*_t$.

### 4.3 Computation of Error Upper Bound

We update the error upper bound estimate $\epsilon^u_{\widehat{z}_t,t}$ and related window variables $\boldsymbol{\Phi}^C_{\widehat{z}_t,t}, \widehat{\mathbf{W}}^C_{\widehat{z}_t,t}, \mathbf{h}^C_{\widehat{z}_t,t}$ of the chosen candidate in Line 23. The details of this update are given in Algorithm 2. If the window is not full, i.e. $c_{\widehat{z}_t} < N_C$, we simply follow the previous upper bound estimate, i.e. $\infty$; otherwise, we update according to a slightly complicated rule whose justification is given in Theorem 9.

## 5 Theoretical Results

Our main theorems are Theorem 12 and Theorem 17, which show the partial and local convergence guarantees for the algorithm respectively. Note that the proofs for all the lemmas, theorems, and corollaries are provided at the appendices.

To ease the exposition, we introduce some notation and concepts that are frequently used later:

- In Algorithm 1, we sample a column index $l_t$ from the matrix $\boldsymbol{\Phi}^R_{i,t}$ in Line 19 of Algorithm 1. Since $\boldsymbol{\Phi}^R_{i,t}$ is a matrix with columns being data vectors collected at different time, we essentially sampled a time index. Let $r_t(l_t)$ denote the true time corresponding to the collecting time of data of column $l_t$.
- Let $r(i,t)$ denote number of times subsystem $i$ is dominant up to time $t$.
- Setup(A): Assume hybrid SARX system only involves 1 subsystem, namely, subsystem $i$ with parameter $\mathbf{w}_i$. Then $\widehat{\mathbf{w}}_{i,t}$ is the only candidate estimate. We let $\boldsymbol{\epsilon}_{i,t} = \mathbf{w}_i - \widehat{\mathbf{w}}_{i,t}$ denote the estimation error.

**Algorithm 2:** UpdateUpperBound

Update window variables for chosen candidate:

$\mathbf{\Phi}^C_{\widehat{z}_t,t} = [\mathbf{\Phi}^C_{\widehat{z}_t,t-1}[:, 2:\text{end}], \boldsymbol{\phi}^*_t]$

$\widehat{\mathbf{W}}^C_{\widehat{z}_t,t} = [\widehat{\mathbf{W}}^C_{\widehat{z}_t,t-1}[:, 2:\text{end}], \widehat{\mathbf{w}}_{\widehat{z}_t,t}]$

$\mathbf{h}^C_{\widehat{z}_t,t} = [\mathbf{h}^C_{\widehat{z}_t,t-1}[2:\text{end}]; \eta^*_t]$

Update the error upper bound for chosen candidate:

**if** $c_{\widehat{z}_t} < N_C$ **then**

  $\epsilon^u_{\widehat{z}_t,t} = \epsilon^u_{\widehat{z}_t,t-1}$

**else**

  $\Delta\widehat{\mathbf{W}} = \widehat{\mathbf{w}}_{\widehat{z}_t,t}\mathbf{1}^\intercal_{N_C\times 1} - \widehat{\mathbf{W}}^C_{\widehat{z}_t,t}$

  $\Delta\widehat{\mathbf{w}} = \widehat{\mathbf{w}}_{\widehat{z}_t,t} - \widehat{\mathbf{w}}_{\widehat{z}_t,t-N_C}$

  $\mathbf{H} = \mathbf{diag}(\mathbf{h}^C_{\widehat{z}_t,t})$

  $\mathbf{A} = (\mathbf{\Phi}^C_{\widehat{z}_t,t}\mathbf{H}{\mathbf{\Phi}^C_{\widehat{z}_t,t}}^\intercal)^{-1}\mathbf{\Phi}^C_{\widehat{z}_t,t}\mathbf{H}$

  $\mathbf{b} = (\mathbf{\Phi}^C_{\widehat{z}_t,t}\mathbf{H}{\mathbf{\Phi}^C_{\widehat{z}_t,t}}^\intercal)^{-1}\left[\Delta\widehat{\mathbf{w}} - \mathbf{\Phi}^C_{\widehat{z}_t,t}\mathbf{H}\square\left(\mathbf{\Phi}^C_{\widehat{z}_t,t}, \Delta\widehat{\mathbf{W}}\right)\right]$

  Let $V = \{[\pm n_{\max}, \pm n_{\max}, \ldots, \pm n_{\max}]^\intercal_{N_C}\}$

  $\epsilon^u_{\widehat{z}_t,t} = \max_{\mathbf{n}\in V}\|A\mathbf{n} - \mathbf{b}\|$

**Remark:** $\square(A, B) \equiv [a_1^\intercal b_1, a_2^\intercal b_2, \ldots, a_n^\intercal b_n]^\intercal$, where $a_i, b_i$ are the $i$th columns of matrices $A, B$

### 5.1 Preliminary Results

In Section 5.1, we first present several lemmas that serve as the building blocks for later theorems.

**Lemma 6.** *$\forall i$, after $c_i \geq N_R$, since $\mathbf{\Phi}^R_{i,t}{\mathbf{\Phi}^R_{i,t}}^\intercal = \sum \boldsymbol{\phi}\boldsymbol{\phi}^\intercal$, and following Assumption 4, we know the singular values of $\mathbf{\Phi}^R_{i,t}$ is upper and lower bounded by $s_{\max}$ and $s_{\min}$ selectively. Following this, the following results hold trivially*

*(i) Let $F_{\max} = \sqrt{n}s_{\max}, F_{\min} = \sqrt{n}s_{\min}$, then we have*

$$F_{\min} \leq \|\mathbf{\Phi}^R_{i,t}\|_F \leq F_{\max} \tag{4}$$

*(ii) Let $\kappa(\mathbf{\Phi}^R_{i,t}) = \|\mathbf{\Phi}^R_{i,t}\|_F\|{\mathbf{\Phi}^R_{i,t}}^{-1}\|_2, \kappa_{\max} = \sqrt{((n-1)s^2_{\max} + s^2_{\min})/s^2_{\min}}$, and $\kappa_{\min} = \sqrt{n}$, where $-1$ denotes the right inverse, then we have*

$$\kappa_{\min} \leq \kappa(\mathbf{\Phi}^R_{i,t}) \leq \kappa_{\max} \tag{5}$$

*(iii) Let $\xi(\mathbf{\Phi}^R_{i,t}) = \|\mathbf{\Phi}^R_{i,t}\|_F / \|\mathbf{\Phi}^R_{i,t}\|_2$, $\xi_{\max} = \sqrt{n}$, and $\xi_{\min} = \sqrt{(s^2_{\max} + (n-1)s^2_{\min})/s^2_{\max}}$, then we have*

$$\xi_{\min} \leq \xi(\mathbf{\Phi}^R_{i,t}) \leq \xi_{\max} \tag{6}$$

**Lemma 7.** *Following Assumption 2, we have*

*(i)* $\mathbb{E}[n_{r_t(l_t)}] = 0, \mathbb{E}[n^2_{r_t(l_t)}] = \sigma^2_n$

*(ii)* $n_{r_t(l_t)}$ *and* $\boldsymbol{\phi}_{r_t(l_t)}$ *are uncorrelated*

*(iii)*

$$\frac{N_R}{F^2_{\max}}\sigma^2_n \leq \mathbb{E}\left[\frac{n^2_{r_t(l_t)}}{\|\boldsymbol{\phi}_{r_t(l_t)}\|^2}\right] \leq \frac{N_R}{F^2_{\min}}\sigma^2_n \tag{7}$$

The following Lemma 8 is extension of result in [12].

**Lemma 8.** *For any random vector $\mathbf{z} \in \mathbb{R}^n$, we have*

$$\kappa_{\max}^{-2}\mathbb{E}[\|\mathbf{z}\|^2] \leq \mathbb{E}\left[\left(\frac{\boldsymbol{\phi}_{r_t(l_t)}^{\mathsf{T}}\mathbf{z}}{\|\boldsymbol{\phi}_{r_t(l_t)}\|}\right)^2\right] \leq \xi_{\min}^{-2}\mathbb{E}[\|\mathbf{z}\|^2] \tag{8}$$

## 5.2 Valid Upper Bound

The following theorem gives justification for error upper bound $\epsilon_{\widehat{z}_t,t}^u$ computed from Line 9 to Line 15 in Algorithm 2. However, this is a restrictive result, as it requires that the candidate $\widehat{z}_t$ is updated with data from the same subsystem for last $N_C$ steps, i.e. the elements in $\boldsymbol{\Phi}_{\widehat{z}_t,t}^C, \widehat{\mathbf{W}}_{\widehat{z}_t,t}^C, \mathbf{h}_{\widehat{z}_t,t}^C$ are collected from the same subsystem.

**Theorem 9.** *Assume at some time $t$, $\widehat{z}_t = z_t = i$, $c_i \geq N_C$, and $\boldsymbol{\Phi}_{\widehat{z}_t,t}^C, \widehat{\mathbf{W}}_{\widehat{z}_t,t}^C, \mathbf{h}_{\widehat{z}_t,t}^C$ are constructed from data entirely from subsystem $i$, then $\epsilon_{i,t}^u$ is a valid upper bound for $\boldsymbol{\epsilon}_{i,t}$, i.e. $\epsilon_{i,t}^u \geq \|\boldsymbol{\epsilon}_{i,t}\| = \|\mathbf{w}_i - \widehat{\mathbf{w}}_{i,t}\|$.*

## 5.3 Partial Convergence Results

In this section, we list the results regarding partial convergence where we assume there is no misassignment, i.e. data generated from the same subsystem can all be assigned to one particular candidate. In this sense, to analyze the convergence properties, it suffices to consider the case where there is only one subsystem in the hybrid model, and one corresponding candidate, i.e. Setup(A).

Lemma 10 provides the convergence analysis at the beginning phase of the algorithm, when $t < N_R$ and Algorithm 1 executes Line 17. Then Lemma 11 provides the convergence analysis for the second phase of the algorithm, when $t \geq N_R$ and Algorithm 1 executes Line 19 and 20. Finally, we have the partial convergence result Theorem 12 simply by combining Lemma 10 and Lemma 11.

**Lemma 10.** *With Setup(A), we have*

$$\frac{\sigma_n^2}{\phi_{\max}^2} \leq \mathbb{E}\left[\|\boldsymbol{\epsilon}_{i,N_R-1}\|^2\right] \leq \|\boldsymbol{\epsilon}_{i,0}\|^2 + \frac{N_R - 1}{S_{\min}^2} \tag{9}$$

**Lemma 11.** *With Setup(A), for $t \geq N_R$ we have*

$$\mathbb{E}\left[\|\boldsymbol{\epsilon}_{i,t}\|^2\right] \leq \left(1 - \kappa_{\max}^{-2}\right)^{t-N_R+1}\mathbb{E}\left[\|\boldsymbol{\epsilon}_{i,N_R-1}\|^2\right] + N_R\frac{\kappa_{\max}^2}{F_{\min}^2}\left[1 - \left(1 - \kappa_{\max}^{-2}\right)^{t-N_R+1}\right]\sigma_n^2 \tag{10}$$

$$\mathbb{E}\left[\|\boldsymbol{\epsilon}_{i,t}\|^2\right] \geq \left(1 - \xi_{\min}^{-2}\right)^{t-N_R+1}\mathbb{E}\left[\|\boldsymbol{\epsilon}_{i,N_R-1}\|^2\right] + N_R\frac{\xi_{\min}^2}{F_{\max}^2}\left[1 - \left(1 - \xi_{\min}^{-2}\right)^{t-N_R+1}\right]\sigma_n^2 \tag{11}$$

**Theorem 12** (Partial Convergence)**.** *WLOG, assume for any $i$, data generated by subsystem $i$ will all be assigned to candidate $i$. Let $\boldsymbol{\epsilon}_{i,t} = \mathbf{w}_i - \widehat{\mathbf{w}}_{i,t}$ denote the estimation error of candidate $i$ at time $t$. Then $\forall i, t$ such that $r(i,t) \geq N_R$, we have*

$$\mathbb{E}\left[\|\boldsymbol{\epsilon}_{i,t}\|^2\right] \leq \left(1-\kappa_{\max}^{-2}\right)^{r(i,t)-N_R+1}\left(\|\boldsymbol{\epsilon}_{i,0}\|^2+\frac{(N_R-1)}{S_{\min}^2}\right) + N_R\frac{\kappa_{\max}^2}{F_{\min}^2}\left[1 - \left(1 - \kappa_{\max}^{-2}\right)^{r(i,t)-N_R+1}\right]\sigma_n^2 \tag{12}$$

$$\mathbb{E}\left[\|\boldsymbol{\epsilon}_{i,t}\|^2\right] \geq \left(1 - \xi_{\min}^{-2}\right)^{r(i,t)-N_R+1}\frac{\sigma_n^2}{\phi_{\max}^2} + N_R\frac{\xi_{\min}^2}{F_{\max}^2}\left[1 - \left(1 - \xi_{\min}^{-2}\right)^{r(i,t)-N_R+1}\right]\sigma_n^2 \tag{13}$$

*If as $t \to \infty$, we have $r(i,t) \to \infty$, i.e. subsystem $i$ can dominate infinitely often, then as $t \to \infty$, we shall have*

$$N_R\frac{\xi_{\min}^2}{F_{\max}^2}\sigma_n^2 \leq \mathbb{E}\left[\|\boldsymbol{\epsilon}_{i,t}\|^2\right] \leq N_R\frac{\kappa_{\max}^2}{F_{\min}^2}\sigma_n^2 \tag{14}$$

### 5.4 Local Convergence Results

In this section, we present results regarding local convergence. Lemma 16 shows that when all candidates have accurate enough estimates, then the next assignment will be correct. Lemma 15, which is derived from Lemma 14, gives a lower bound on the probability that the estimates will stay accurate enough during the algorithm assuming assignments are correct. By Lemma 15 and Lemma 16, we could obtain the local convergence result Theorem 17.

For Lemma 14 to hold, we need a technical assumption given below to guarantee that estimation errors form a supermartingale, which allows us to use supermartingale maxima inequality to get the probability bound in Lemma 14. Note that Assumption 13 is not mandatory for the algorithm to work, instead, it's solely for analysis purposes. Also note that if there is no noise, even though Assumption 13 fails, the estimation errors still form a supermartingale, which enables us to proceed with the analysis. Local convergence for the noiseless case is provided in Corollary 18.

**Assumption 13.** *Assume there is an upper bound on the SNR:* $\forall t, \frac{\|\boldsymbol{\phi}_t\|}{|n_t|} \leq S_{\max}$, *which satisfies* $S_{\max} \leq \kappa_{\max} S_{\min}$.

**Lemma 14.** *Assume Assumption 13 holds. With Setup(A), for* $\forall t \geq N_R$ *and some* $\epsilon' > 0$, *we have*

$$P\left(\bigcap_{\tau=N_R}^{t} \left\{\|\boldsymbol{\epsilon}_{i,\tau}\|^2 \leq \epsilon'^2\right\}\right) \geq 1 - \frac{\mathbb{E}\left[\|\boldsymbol{\epsilon}_{i,N_R-1}\|^2\right]}{\epsilon'^2} \tag{15}$$

**Lemma 15.** *Assume Assumption 13 holds. With Setup(A), for some* $\epsilon' > 0$, *assume* $\|\boldsymbol{\epsilon}_{i,0}\| \leq \epsilon_0$ *such that* $\sqrt{N_R\left(\epsilon_0^2 + \frac{N_R}{S_{\min}^2}\right)} \leq \epsilon'$, *then for* $\forall t$ *we have*

$$P\left(\bigcap_{\tau=1}^{t} \left\{\|\boldsymbol{\epsilon}_{i,\tau}\|^2 \leq \epsilon'^2\right\}\right) \geq 1 - 2\sqrt{\frac{N_R}{\epsilon'^2}\left(\epsilon_0^2 + \frac{N_R}{S_{\min}^2}\right)} \tag{16}$$

**Lemma 16.** *Let* $\epsilon' = \frac{1}{2\phi_{\max}}\left(\psi - \frac{n_{\max}}{\nu S_{\min}} - 3n_{\max}\right)$, $\alpha = 2$, *and* $\beta = 1$. *Assume at time* $t$, *candidates are one-to-one* $\epsilon'$*-close to subsystems. WLOG, we could assume* $\forall i, \|\boldsymbol{\epsilon}_{i,t-1}\| \equiv \|\mathbf{w}_i - \widehat{\mathbf{w}}_{i,t-1}\| \leq \epsilon'$. *Furthermore, we assume that all assignments prior to time* $t$ *are made correctly, i.e.* $\forall s < t, \widehat{z}_s = z_s$. *Then at time* $t$, *we will also assign data correctly, i.e.* $\widehat{z}_t = z_t$.

**Theorem 17** (Local Convergence)**.** *Assume Assumption 13 holds. Let* $\epsilon' = \frac{1}{2\phi_{\max}}\left(\psi - \frac{n_{\max}}{\nu S_{\min}} - 3n_{\max}\right)$, $\alpha = 2$, *and* $\beta = 1$. *Let* $\boldsymbol{\epsilon}_{i,t} = \mathbf{w}_i - \widehat{\mathbf{w}}_{i,t}$ *denote the estimation error of candidate* $i$ *at time* $t$. *WLOG, assume* $\forall i, \|\boldsymbol{\epsilon}_{i,0}\| \leq \epsilon_0$ *such that* $\sqrt{N_R\left(\epsilon_0^2 + \frac{N_R}{S_{\min}^2}\right)} \leq \epsilon'$. *Then* $\forall i, t$ *such that* $r(i,t) \geq N_R$, *with probability at least* $1 - 2m\sqrt{\frac{N_R}{\epsilon'^2}\left(\epsilon_0^2 + \frac{N_R}{S_{\min}^2}\right)}$, *we have the following results: (i). We can correctly identify the switching sequence, i.e.* $\forall t, \widehat{z}_t = z_t$. *In another way,* $\forall i, t$, $\{\boldsymbol{\phi}_t, y_t\}$ *from subsystem* $i$ *will be assigned to candidate* $i$. *(ii). Results for* (12), (13) *and* (14) *will hold.*

**Corollary 18** (Local Convergence Without Noise)**.** *Let* $n_t = 0$, *i.e. there is no noise. Let* $\epsilon' = \frac{\psi}{2\phi_{\max}}$, $\alpha = 2$, *and* $\beta = 1$. *Let* $\boldsymbol{\epsilon}_{i,t} = \mathbf{w}_i - \widehat{\mathbf{w}}_{i,t}$ *denote the estimation error of candidate* $i$ *at time* $t$, *and assume* $\forall i, \|\boldsymbol{\epsilon}_{i,0}\| \leq \epsilon_0$ *such that* $\sqrt{N_R\epsilon_0^2} \leq \epsilon'$.

*Then* $\forall i, t$ *such that* $r(i,t) \geq N_R$, *with probability at least* $1 - 2m\sqrt{\frac{N_R}{\epsilon'^2}\epsilon_0^2}$, *we have the following results: (i). We can correctly identify the switching sequence, i.e.* $\forall t, \widehat{z}_t = z_t$. *In another way,* $\forall t, \forall i$,

$\{\boldsymbol{\phi}_t, y_t\}$ *from subsystem* $i$ *will be assigned to candidate* $i$. *(ii). we have the following convergence results:* $\forall i, t$ *such that* $r(i,t) \geq N_R$

$$\mathbb{E}\left[\|\boldsymbol{\epsilon}_{i,t}\|^2\right] \leq \left(1-\kappa_{\max}^{-2}\right)^{r(i,t)-N_R+1} \|\boldsymbol{\epsilon}_{i,0}\|^2 \tag{17}$$

*If as* $t\to\infty$, *we have* $r(i,t)\to\infty$, *i.e. subsystem* $i$ *can dominate infinitely often, then as* $t\to\infty$, *we have* $\mathbb{E}\left[\|\boldsymbol{\epsilon}_{i,t}\|^2\right]=0$.

# 6 Discussions and Extensions

## 6.1 Poles, Condition Number, and Convergence Rate

Systems with poles close to the unit circle are not preferable as they are close to be unstable. In algorithm convergence analysis, Hessian matrix or objective function with large condition number is usually not preferable as the convergence rate tends to get small. In this section, we will show how these two facts meet consistently in our algorithm. That is, as the system poles getting closer to the unit circle, the condition number of Hessian matrix will get larger, and the convergence rate of upper bound in (12) will get smaller.

To study the convergence rate, it suffices to study a single subsystem without any switching. We drop the subsystem subscript, and let $\boldsymbol{\epsilon}_t = \mathbf{w} - \widehat{\mathbf{w}}_t$ denote the estimation error. Since the goal of this section is to provide insight into the relations between poles, condition number, and convergence rate, so several steps involve approximation. And when study how poles affect the condition number, we only consider a toy system with order 3, since it is challenging to find nice analytical expressions for systems with higher order.

First we consider how condition number influences convergence the rate of upper bound in (12).

### 6.1.1 Condition Number vs. Convergence Rate

The expression for single ARX system is given by $y_t = \sum_{j=1}^{n_a} a_j y_{t-j} + \sum_{k=1}^{n_c} c_k u_{t-k} + n_t = \mathbf{w}^\intercal \boldsymbol{\phi}_t + n_t$ following our notations in Section 2. In Assumption 4, we have $s_{\min}^2 I_n \preceq \sum_{t\in S} \boldsymbol{\phi}_t \boldsymbol{\phi}_t^\intercal \preceq s_{\max}^2 I_n$. We will see this equation is related to the correlation matrix $\mathbf{R} \equiv \mathbb{E}[\boldsymbol{\phi}_t \boldsymbol{\phi}_t^\intercal]$, if it exists.

In [6], we could know for the ARX system given above, if (i) poles of system are within the unit circle, and (ii) noise is white Gaussian and input is wide-sense stationary, then there exists $\mathbf{R}$ such that $\lim_{t\to\infty} \mathbb{E}[\boldsymbol{\phi}_t \boldsymbol{\phi}_t^\intercal] = \mathbf{R}$.

When $N_R$ is large, according to law of large numbers, equation (3) and the result above, we have $\lim_{\min(S)\to\infty} \sum_{t\in S} \boldsymbol{\phi}_t \boldsymbol{\phi}_t^\intercal \approx N_R \mathbf{R}$. We let $\lambda_{\max}, \lambda_{\min}$ denote the maximum and minimum eigenvalue of $\mathbf{R}$. Now dropping the "lim" and replace "$\approx$" with "$=$", we could get $N_R \lambda_{\min} I_n \preceq \sum_{t\in S} \boldsymbol{\phi}_t \boldsymbol{\phi}_t^\intercal \preceq N_R \lambda_{\max} I_n$. So $N_R\lambda_{\min}$ and $N_R\lambda_{\max}$ are equivalent to $s_{\min}^2$ and $s_{\max}^2$ defined in Assumption 4. Then according to Lemma 6, we could have $\kappa_{\max} = \sqrt{(n-1)\lambda_{\max}/\lambda_{\min}+1}$ and $\xi_{\min} = \sqrt{(n-1)\lambda_{\min}/\lambda_{\max}+1}$. So, for the asymptotic convergence upper bounds in (10), (12), and (17) which all involve $\kappa_{\max}$, when the condition number of $\mathbf{R}$, $\lambda_{\max}/\lambda_{\min}$, increases, $\kappa_{\max}$ will increase, and the convergence rate in upper bounds will decrease.

### 6.1.2 Poles vs. Condition Number

We consider a toy example of system with order 3: $y_t = a_1 y_{t-1} + a_2 y_{t-2} + c_1 u_{t-1} + n_t = \mathbf{w}^\intercal \boldsymbol{\phi}_t + n_t$ where $\mathbf{w} \equiv [a_1, a_2, c_1]$ and $\boldsymbol{\phi}_t \equiv [y_{t-1}, y_{t-2}, u_{t-1}]$. We assume all poles are within the unit circle,

$u_t \sim \mathcal{N}(0, \sigma_u^2)$, $n_t \sim \mathcal{N}(0, \sigma_n^2)$, $u_t \perp n_t$, $u_t \perp u_s$, $n_t \perp n_s, \forall t, s$, and $\sigma_u \gg \sigma_n$. Following [6], we have

$$\mathbf{R} = \mathbb{E}\left[\boldsymbol{\phi}_t \boldsymbol{\phi}_t^\mathsf{T}\right] = \begin{bmatrix} r(0) & r(1) & 0 \\ r(1) & r(0) & 0 \\ 0 & 0 & \sigma_u^2 \end{bmatrix} \tag{18}$$

where $r(0), r(1)$ can be computed by solving

$$\begin{bmatrix} r(0) \\ r(1) \\ r(2) \end{bmatrix} = \begin{bmatrix} 1 & -a_1 & -a_2 \\ -a_1 & 1-a_2 & 0 \\ -a_2 & -a_1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} \sigma_n^2 + c_1 \sigma_u^2 \\ 0 \\ 0 \end{bmatrix} \tag{19}$$

So, we have

$$\mathbf{R} = \mathbb{E}\left[\boldsymbol{\phi}_t \boldsymbol{\phi}_t^\mathsf{T}\right] = \begin{bmatrix} (a_2-1)c & (-a_1)c & 0 \\ (-a_1)c & (a_2-1)c & 0 \\ 0 & 0 & \sigma_u^2 \end{bmatrix} \tag{20}$$

where $c = \frac{\sigma_n^2 + c_1 \sigma_u^2}{(a_2+1)(a_1+a_2-1)(a_1-a_2+1)}$. We will drop $\sigma_n^2$ in the following computation as $\sigma_u \gg \sigma_n$. The eigenvalues of $\mathbf{R}$ are given by

$$\lambda_1 = -\frac{c_1 \sigma_u^2}{(a_2+1)(a_1+a_2-1)} \tag{21}$$

$$\lambda_2 = \frac{c_1 \sigma_u^2}{(a_2+1)(a_1-a_2+1)} \tag{22}$$

$$\lambda_3 = \sigma_u^2 \tag{23}$$

Note that the poles $p_1, p_2$ satisfy $p_1 + p_2 = a_1$ and $p_1 p_2 = -a_2$, so we have

$$\lambda_1 = \frac{c_1 \sigma_u^2}{(1-p_1 p_2)(1-p_1)(1-p_2)} \tag{24}$$

$$\lambda_2 = \frac{c_1 \sigma_u^2}{(1-p_1 p_2)(1+p_1)(1+p_2)} \tag{25}$$

$$\lambda_3 = \sigma_u^2 \tag{26}$$

Let $c_1 \geq 1$, since $p_1, p_2 < 1$, we can see the condition number of $\mathbf{R}$ will have the following lower bound

$$\frac{\lambda_{\max}}{\lambda_{\min}} \geq \frac{\lambda_1}{\lambda_3} = \frac{1}{(1-p_1 p_2)(1-p_1)(1-p_2)} \tag{27}$$

It's easy to see as poles get closer to the unit circle, this lower bound will get larger and the condition number is likely to increase as well.

### 6.1.3 Poles vs. Convergence Rate

Finally, by combining the two results we just showed, we could see as the system poles getting closer to unit circle, the convergence rate of upper bound in (12) will decrease.

There are two comments regarding this conclusion. (i) Even though this result only involves the rate of *upper bound*, empirical results show the true convergent rate follow accordingly; (ii) Our algorithm favors stable system which is a little counterintuitive as unstable system tends to have higher SNR.

### 6.2 Unbounded Noise and Monte Carlo Method

Note that we compute the error upper bound $\epsilon^u_{\widehat{z}_t,t}$ in Line 15 of Algorithm 2 by finding the maximum $\|A\mathbf{n}-\mathbf{b}\|$ from cube vertices $V$ defined by the noise magnitude upper bound $n_{\max}$. However, if $n_{\max}$ is unknown or the noise itself is unbounded, e.g. Gaussian, Algorithm 2 is not applicable to evaluate $\epsilon^u_{\widehat{z}_t,t}$. In this case, if we could have samples of noise instead, an alternative approach is to use Monte Carlo method to evaluate $\epsilon^u_{\widehat{z}_t,t}$. Specifically, if we have $N_t$ samples of noise vector $\mathbf{n}_t$ (defined in the proof for Theorem 9) given by $\{\mathbf{n}_t^{(i)}\}_{i=1}^{N_t}$, we could let $\epsilon^u_{\widehat{z}_t,t} = \max \|A\mathbf{n}_t^{(i)} - \mathbf{b}\|$. Due to the Monte Carlo nature, this is not necessarily a valid upper bound. In another way, the result in Theorem 9 doesn't hold, i.e. $\epsilon^u_{\widehat{z}_t,t} < \|\boldsymbol{\epsilon}_{\widehat{z}_t,t}\|$. Practically, algorithm still has satisfactory performance when using this Monte Carlo method, but theoretically, this may not guarantee local convergence since local convergence result Theorem 17 implicitly Theorem 9, i.e. $\epsilon^u_{\widehat{z}_t,t} \geq \|\boldsymbol{\epsilon}_{\widehat{z}_t,t}\|$ , to hold for every time step.

If we prefer the theoretical guarantees to practical implementation, by subtly designing the number of Monte Carlo samples $N_t$ at time $t$, there could be some probability guarantee to ensure $\epsilon^u_{\widehat{z}_t,t}$ is a valid upper bound at every time step.

For ease of illustration, we assume there is only one subsystem, then we could drop the subsystem index subscript, and replace $\epsilon^u_{\widehat{z}_t,t}$ with $\epsilon^u_t$, $\boldsymbol{\epsilon}_{\widehat{z}_t,t}$ with $\boldsymbol{\epsilon}_t$. And we assume the Monte Carlo method starts at time 1. Then we have the following theorem:

**Theorem 19.** *If we use Monte Carlo method above to compute $\epsilon^u_t$, for some $\zeta_1, \zeta_2 \in (0,1)$, let $N_t \geq \frac{\zeta_2 t}{2\zeta_1^{2t}}$, then*

$$P\left(P\left(\bigcap_{t=1}^{\infty}\{\|\boldsymbol{\epsilon}_t\|\leq\epsilon_t^u\}\right)\geq 1-\frac{\zeta_1}{1-\zeta_1}\right)\geq 1-\frac{\exp(-\zeta_2)}{1-\exp(-\zeta_2)} \tag{28}$$

*In another word, the probability that every $\epsilon^u_t$ is a valid upper bound is large with a large probability.*

The proof for this theorem is again in the appendices. We can immediately see from Theorem 19 that in order to make the probabilities large, the number of Monte Carlo samples need to increase exponentially with respect to time, which makes implementation intractable when time is long.

### 6.3 Extension to MIMO Case

So far we have been a considering SISO system in (1), where all the $y_t$ and $u_t$ are scalars. However, our algorithm can be applied to MIMO systems with some transformation of the system equation, and we will provide a potential direction in this section.

Let $\mathbf{y}_t \in \mathbb{R}^{n_y}, \mathbf{u}_t \in \mathbb{R}^{n_u}$, then the MIMO SARX system is given by

$$\mathbf{y}_t = \sum_{j=1}^{n_a} \mathbf{A}_j(z_t)\mathbf{y}_{t-j} + \sum_{k=1}^{n_c} \mathbf{C}_k(z_t)\mathbf{u}_{t-k} + \mathbf{n}_t \tag{29}$$

where $\{\mathbf{A}_j(z_t)\}_{j=1}^{n_a}$, $\{\mathbf{C}_j(z_t)\}_{j=1}^{n_c}$ are the parameters of subsystem $z_t$. Let $\mathbf{W}_{z_t} = [\mathbf{A}_1(z_t), \ldots, \mathbf{A}_{n_a}(z_t), \mathbf{C}_1(z_t), \ldots \mathbf{C}_{n_c}(z_t)]^\intercal$, $\boldsymbol{\phi}_t = [\mathbf{y}_{t-1}^\intercal, \ldots, \mathbf{y}_{t-n_a}^\intercal, \mathbf{u}_{t-1}^\intercal, \ldots, \mathbf{u}_{t-n_c}^\intercal]^\intercal$. Let $\mathbf{w}_{z_t,i}$ denote the $i$th column of $\mathbf{W}_{z_t}$, and let $y_{t,i}, n_{t,i}$ denote the $i$th element in $\mathbf{y}_t$ and $\mathbf{n}_t$. Then the MIMO system can be broken into a set of equations: $\forall i \in [n_y]$,

$$y_{t,i} = \mathbf{w}_{z_t,i}^\intercal \boldsymbol{\phi}_t + n_{t,i} \tag{30}$$

which has the same form as (1). So we could modify our algorithm to estimate each $\mathbf{w}_{z_t,i}$ in a parallel way, then combine them to estimate $\mathbf{W}_{z_t}$.

### 6.4 Multiple $N_C$'s and Forgetting Factor

#### 6.4.1 Multiple $N_C$'s

Note that in Algorithm 2, we have window variables $\mathbf{\Phi}_{i,t}^C \in \mathbb{R}^{n\times N_C}, \widehat{\mathbf{W}}_{i,t}^C \in \mathbb{R}^{n\times N_C}, \mathbf{h}_{i,t}^C \in \mathbb{R}^{N_C}$ for some window length $N_C$ to compute the error upper bound $\epsilon_{\widehat{z}_t,t}^u$. Theorem 9 says when all data stored in the window are from the same subsystem, then $\epsilon_{\widehat{z}_t,t}^u$ will be a valid error upper bound with respect this subsystem. However, if there enters some *outlier* data (data generated by subsystem that is different from the subsystem that generates the majority of data in the window variables), $\epsilon_{\widehat{z}_t,t}^u$ computed using window variables might be an invalid upper bound. If the window length is too large, since the window is sliding, the effect of outlier will stay a longer time, but the correction effect of the majority of the inlier data might reduce the effect of outlier. On the contrary, if the window length is too small, the effect of outlier will quickly vanish, but the correction effect from the inlier data will reduce as well and we may have even worse $\epsilon_{\widehat{z}_t,t}^u$ during the stay of outlier.

Practically, we could use multiple $N_C$'s and corresponding window variables. Each set of window variables compute $\epsilon_{\widehat{z}_t,t}^u$ separately, and we pick the maximum of them as the final decision. In this way, the disadvantages of large and small window lengths might cancel out each other thus making $\epsilon_{\widehat{z}_t,t}^u$ more robust to misassignment.

#### 6.4.2 Forgetting Factor

One interesting fact about our algorithm is, in Line 19 of Algorithm 1, instead of using the latest data, we pick randomly from previous data to update the estimates. This idea is initially proposed in [12]. The reason we incorporate this randomization into the algorithm is to acquire the asymptotic convergence result via Assumption 4, Lemma 6, Lemma 7, and Lemma 8.

If no randomization scheme is utilized, the algorithm on a single subsystem is equivalent to the Kaczmarz algorithm or the normalized least mean squares (NLMS) algorithm in [6]. This type of algorithm, however, does not have satisfactory convergence results yet. One linear convergence result provided in [6] is valid only when the step size in estimate update is very small, which makes it little practical use. The difficulty to derive nice convergence results is that nearby data $\boldsymbol{\phi}_t$ could be highly correlated, which can be seen from the definition, and updating the estimate with data in chronological order aggravates the situation. The randomized scheme picks data for update randomly and independently, which brings independence into the algorithm and makes analysis tractable.

Empirically, if we don't incorporate the random selection in Line 19, and always use latest data as Line 17, the performance can sometimes be slightly better. This is potentially because when sampling previous data vectors, it's likely that we sample one data multiple times possibly due to its large norm, and, generally speaking, previously used data may not provide as much information as some new data.

One potential way to balance between establishing theoretical results and exploiting new data is to incorporate a forgetting factor $\gamma$. Specifically, in Line 19, we sample data according to the following distribution

$$P(l_t = i) \begin{cases} \gamma \|\mathbf{\Phi}_{\widehat{z}_t,t}^R[:,\, i]\|^2 \frac{1}{F} \text{ if } i = N_R \\ (1-\gamma)\|\mathbf{\Phi}_{\widehat{z}_t,t}^R[:,\, i]\|^2 \frac{1}{F} \text{ if } \text{ if } i < N_R \end{cases} \tag{31}$$

where $\gamma > 0.5$, and $F = \gamma\|\mathbf{\Phi}_{\widehat{z}_t,t}^R[:,\, N_R]\|^2 + (1-\gamma)\sum_{i=1}^{N_R-1} \|\mathbf{\Phi}_{\widehat{z}_t,t}^R[:,\, i]\|^2$ is the normalization factor. With this distribution, we can see the probability of choosing the latest data ($l_t = N_R$) is larger compared with the distribution in Algorithm 1. As $\gamma$ gets closer to 1, we are more likely to sample the latest data.

As for the convergence result, it suffices to only consider how the building block lemmas will change with this new distribution, and the main theorems will follow these lemmas. In the building block lemmas, only the expectations in Lemma 7 (iii), and Lemma 8 involve the data sampling process. With the new sampling distribution, it's not difficult to see (7) and (8) will become

$$\tilde{\gamma}^{-1}\frac{N_R}{F_{\max}^2}\sigma_n^2 \leq \mathbb{E}\left[\frac{n_{r_t(l_t)}^2}{\|\boldsymbol{\phi}_{r_t(l_t)}\|^2}\right] \leq \tilde{\gamma}\frac{N_R}{F_{\min}^2}\sigma_n^2 \tag{32}$$

$$\tilde{\gamma}^{-1}\kappa_{\max}^{-2}\mathbb{E}[\|\mathbf{z}\|^2] \leq \mathbb{E}\left[\left(\frac{\boldsymbol{\phi}_{r_t(l_t)}^{\intercal}\mathbf{z}}{\|\boldsymbol{\phi}_{r_t(l_t)}\|}\right)^2\right] \leq \tilde{\gamma}\xi_{\min}^{-2}\mathbb{E}[\|\mathbf{z}\|^2] \tag{33}$$

where $\tilde{\gamma} = \frac{\gamma}{1-\gamma} > 1$. The rest of the lemmas, theorems, corollaries follow from these new results.

# 7 Numerical Results

In this section, we use simulation examples to evaluate the theoretical results as well as the performance of our algorithm.

## 7.1 Evaluation of Asymptotic Convergence Bounds

Since it is not convenient to visualize the convergence bounds for SARX system with multiple subsystems, and (10) and (11) give tighter performance than (13) and (14), we will evaluate the lower and upper asymptotic convergence bounds in (10) and (11) on single ARX system by comparing the bounds with the actual convergence behavior.

Consider a specific system

$$y_t = 0.7y_{t-1} - 0.12y_{t-2} + u_{t-1} + n_t \tag{34}$$

where $n_t \sim \mathcal{N}(0, \sigma_n^2)$, $\sigma_n = 10^{-4}$, $u_t \sim \mathcal{N}(0,1)$. According to Section 6.1, the correlation matrix is given by

$$R = \begin{bmatrix} 1.67 & 1.04 & 0 \\ 1.04 & 1.67 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{35}$$

, and its minimum and maximum eigenvalues are $\lambda_{\min} = 0.63$ and $\lambda_{\max} = 2.71$.

Since it's difficult to have exact knowledge of $\kappa_{\max}$ and $\xi_{\min}$, we will use the approximate values defined in Section 6.1, i.e. $\kappa_{\max} = \sqrt{(n-1)\lambda_{\max}/\lambda_{\min}+1}$ and $\xi_{\min} = \sqrt{(n-1)\lambda_{\min}/\lambda_{\max}+1}$. And similarly, we could have $F_{\max} = \sqrt{nN_R\lambda_{\max}}$ and $F_{\min} = \sqrt{nN_R\lambda_{\min}}$.

We set $N_R = 10$ and simulation time horizon $T = 1000$. To evaluate the expectation $\mathbb{E}\left[\|\boldsymbol{\epsilon}_{i,t}\|^2\right]$ in (10) and (11), we run the algorithm 50 times with different realizations of input $u_t$, noise $n_t$, and random data selection in Line 19 of Algorithm 1, and take the average of estimation errors as the expectation.

The simulation results are given in Fig. 3. We can see the estimation error $\mathbb{E}\left[\|\boldsymbol{\epsilon}_{i,t}\|^2\right]$ can be successfully bounded by the upper and lower bound in (10) and (11).

## 7.2 Robust Behavior of our Algorithm

### 7.2.1 Single Realization Experiment

First we evaluate our algorithm and compare it with the OBE algorithm in [5] using SARX system given below

- Subsystem 1: $y_t = 0.2y_{t-1} + 0.24y_{t-2} + 2u_{t-1} + n_t$
- Subsystem 2: $y_t = 0.7y_{t-1} - 0.12y_{t-2} + 1u_{t-1} + n_t$

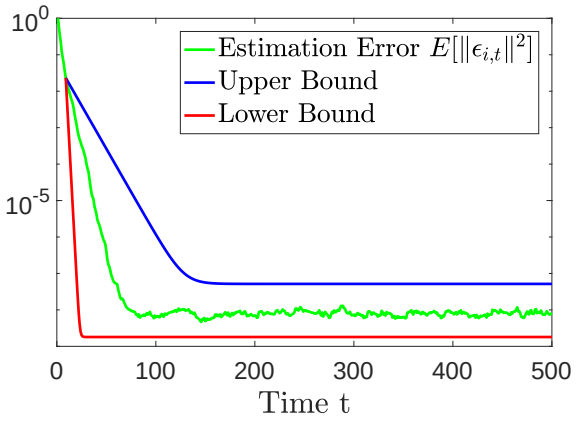


Figure 3: Evaluation of convergence bounds

- Subsystem 3: $y_t = -1.4y_{t-1} - 0.53y_{t-2} + 1u_{t-1} + n_t$
- Subsystem 4: $y_t = 1.7y_{t-1} - 0.72y_{t-2} + 0.5u_{t-1} + n_t$

where $u_t \sim \mathcal{N}(0,1)$. $n_t$ follows $\mathcal{N}(0,\sigma_n^2)$ truncated to region $[-3\sigma_n, 3\sigma_n]$ where $\sigma_n = 10^{-4}$, so noise is bounded with $n_{\max} = 3\sigma_n$.

To fully evaluate the performance, we consider 3 different switching patterns of subsystems: (i) *Slow Switching* (SS): subsystem 1 dominates from 1 to 500, subsystem 2 dominates from 501 to 1000, subsystem 3 dominates from 1001 to 1500, and subsystem 4 dominates from 1501 to 2000. (ii) *Minimum Dwell Time* (MD): each subsystem dominates 30 time steps, and then the time it takes to switch to a new subsystem is a random variable following the geometric distribution with parameter 1/16. When the subsystem switches, all subsystems are equally likely to be switched to, and after the switching, this process restarts again. (iii) *Fast Switching* (FS): at every time step, every subsystem dominates with equal probabilities.

In our algorithm, we set $N_R = 3, N_C = 20, \alpha = 4, \beta = 3, \nu = 10^{-4}$, and simulation time horizon $T = 2000$. The candidates are initialized with standard multivariate Gaussian distribution. After the algorithms completes all $T$ time steps, we first relabel the candidates with a bijective mapping $h(\cdot) : [m] \to [m]$ such that $\sum_{i\in[m]} \|\mathbf{w}_i - \widehat{\mathbf{w}}_{h(i),T}\|$ is minimized. *In the following, the candidates are referring to the relabeled candidates.*

We compute all the estimation errors, i.e. $\boldsymbol{\epsilon}_{i,t} = \mathbf{w}_i - \widehat{\mathbf{w}}_{i,t}, \forall i, t$, which measure the distance between candidate $i$ and subsystem $i$ during the algorithm.

Fig. 4 depicts the simulation results. The dots in the plots represent each $\|\boldsymbol{\epsilon}_{\widehat{z}_t,t}\|, \forall t$, which means there is only one dot plotted for one time step. Different colors correspond to different candidates and corresponding true subsystems. For example, if there exists a blue dot at time $t = 1400$, this means we assign data generated at time $t = 1400$ to candidate 3, and current error between candidate 3 and subsystem 3 is given by the y-axis value of the dot.

From the plots, we see that our algorithm converges more quickly than the OBE algorithm. Since none of the colors have a sharp increase in error, we could claim the phenomenon described in Section 3 is effectively avoided in these realizations. For the OBE algorithm, the performance is obviously worse: in the FS case, none of the candidates even converge, and the algorithm even stops halfway due to numerical instability. In the plots for SS-OBE, we can see the undesired phenomenon described in Section 3: candidate 2 has converged to the vicinity of subsystem 2 from $t = 501$ to $t = 1000$, but after time $t \geq 1500$, the error goes large again. This is because we are assigning data generated by subsystem 4 to candidate 2, making candidate 2 move towards subsystem 4. From these plots, we see

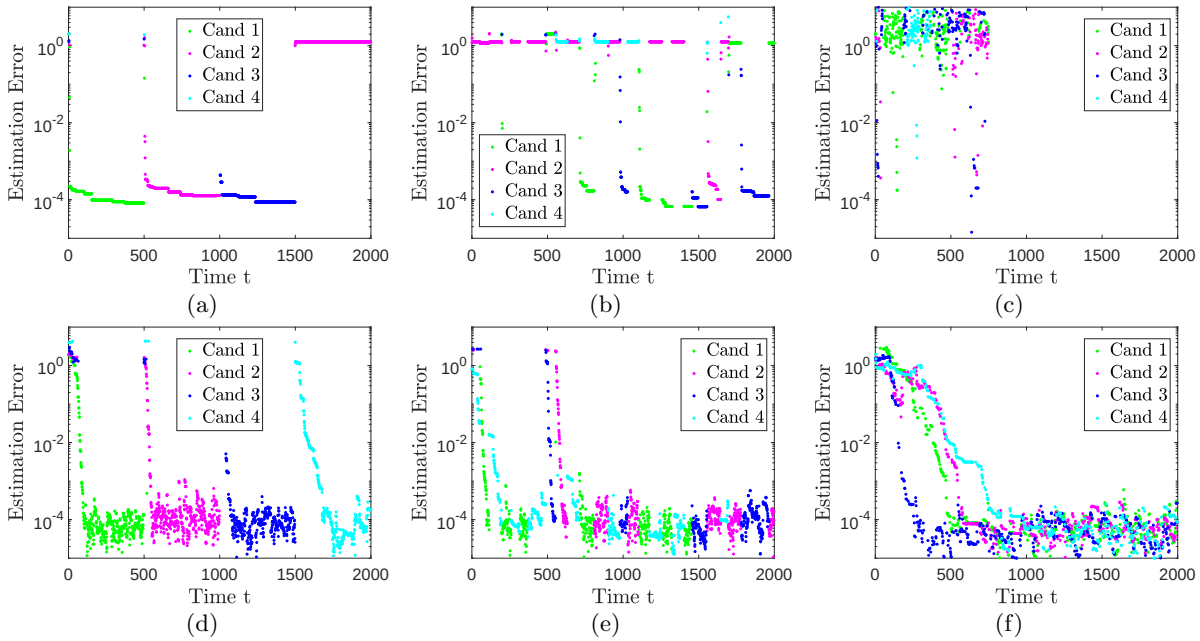


Figure 4: Estimation errors of OBE algorithm and our algorithm: (a) SS-OBE; (b) MD-OBE; (c) FS-OBE; (d) SS-Ours; (e) MD-OBE; (f) FS-Ours

that our algorithm outperforms OBE algorithm for all switching patterns.

### 7.2.2 Multiple Realizations Experiment

Since a single realization cannot comprehensively evaluates the performance, we further compare our algorithm with the OBE algorithm using multiple realizations. Specifically, we consider 9 experiment setups, given by all the combinations of switching patterns {SS, MD, FS} and noise level $\sigma_n \in \{10^{-1}, 10^{-2}, 10^{-3}\}$, and for each of the experiment setup, we run $M$=100 realizations. Each subsystem parameters are generated randomly in each realization: we first sample 2 real poles on $[-1, 1]$ uniformly, and then compute the parameters from the sampled poles. The rest of the setups, e.g. number and orders of subsystems, algorithm parameters, etc., follow the previous single realization experiment.

For realization $i$, we define the two metrics: $\text{FE}(i) = \frac{1}{m}\sum_{j=1}^{m} \|\boldsymbol{\epsilon}_{j,T}\|$ and $\text{CER}(i) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\{z_t \neq \widehat{z}_t\}$. FE measures the final estimation error and CER is the classification error rate. Table 1 lists the average FE and CER values over all 100 realizations. We could see, our algorithm exhibits better performance in each setup.

## 8 Conclusions

In this paper, we introduced a robust algorithm to solve online switched system identification problem. Our algorithm follows the conventional two-step framework, but the modified assignment criterion leads to a more robust assignment process. After we assign the data to some candidate, we update the candidate estimate based on the idea of randomized Kaczmarz algorithm. We showed partial and local convergence results. The partial convergence result is: assuming there is no misassignment, then the estimation error converges geometrically to some quantity related to noise variance in the expectation square sense. The local convergence result is: assuming all candidates have good enough initialization, with some probability, it can be guaranteed that no misassignment will be made, and

Table 1: Results for Multiple Realizations Experiment

| | Ours | OBE | Ours | OBE |
|---|---|---|---|---|
| | FE | FE | CER | CER |
| SS, $10^{-1}$ | 8.4x$10^{-1}$ | 8.7x$10^{-1}$ | 56.3% | 59.1% |
| SS, $10^{-2}$ | 2.8x$10^{-2}$ | 8.2x$10^{-1}$ | 22.1% | 55.5% |
| SS, $10^{-3}$ | 9.0x$10^{-2}$ | 8.2x$10^{-1}$ | 8.35% | 56.4% |
| MD, $10^{-1}$ | 4.3x$10^{-1}$ | 5.2x$10^{-1}$ | 47.5% | 50.3% |
| MD, $10^{-2}$ | 4.0x$10^{-2}$ | 2.8x$10^{-1}$ | 11.3% | 31.3% |
| MD, $10^{-3}$ | 9.4x$10^{-3}$ | 2.4x$10^{-1}$ | 4.91% | 28.8% |
| FS, $10^{-1}$ | 2.6x$10^{-1}$ | 6.8x$10^{-1}$ | 39.3% | 53.9% |
| FS, $10^{-2}$ | 6.0x$10^{-2}$ | 1.5x$10^{-1}$ | 11.7% | 22.1% |
| FS, $10^{-3}$ | 5.8x$10^{-2}$ | 1.8x$10^{-1}$ | 8.93% | 18.9% |

the estimation error will converge geometrically as in the partial result. Numerical results verify the asymptotic convergence bounds we developed, and shows the efficiency of our proposed algorithm in comparison with the existing OBE algorithm.

For future work, there are several aspects that we would focus on.

- As for theories, we would seek to relax Assumption 13, and analyze the local convergence in a more general setting. Also, we could relax even further to analyze the global convergence without good initialization requirement.
- We plan to apply our algorithm to advanced and real world examples to further evaluate its applicability.
- Our current algorithm finds the upper bound of estimation error by searching all the cube vertices $V$ defined in Algorithm 2, which leads to heavy computation burden when $N_C$ is large. In the future, we would seek a way to estimate the error more efficiently without sacrificing theoretical guarantees.
- Since so far we don't consider the case in which we have control over the system input, another interesting extension would be designing certain input, possibly closed-loop or open loop but with certain distribution, given which the system parameters can be learned faster.

## Acknowledgement

The authors thank Yan Shuo Tan for suggesting the use of super-martingale theory, which proved to be crucial in the local convergence analysis.

## References

[1] L. Bako. Identification of switched linear systems via sparse optimization. *Automatica*, 47(4):668–677, Apr. 2011.

[2] L. Bako, K. Boukharouba, E. Duviella, and S. Lecoeuche. A recursive identification algorithm for switched linear/affine models. *Nonlinear Analysis: Hybrid Systems*, 5(2):242–253, May 2011.

[3] V. Bezruck, Y. N. Belov, O. Voitovych, K. Netrebenko, V. Tikhonov, G. Rudnev, G. Khlopov, and S. Khomenko. Application of autoregressive model for recognition of meteorological objects. In *Radar Symposium (IRS), 2010 11th International*, pages 1–3. IEEE, 2010.

[4] J. H. Cochrane. Time series for macroeconomics and finance. *Manuscript, University of Chicago*, 2005.

[5] A. Goudjil, M. Pouliquen, E. Pigeon, and O. Gehan. Convergence analysis of a real-time identification algorithm for switched linear systems with bounded noise. In *IEEE 55th Conference on Decision and Control (CDC)*, pages 2957–2962, 2016.

[6] S. Haykin and B. Widrow, editors. *[Simon Haykin] Least-Mean-Square Adaptive Filters*. Wiley series in adaptive and learning systems for signal processing, communication, and control. Wiley-Interscience, Hoboken, N.J, 2003.

[7] F. Kozin. Autoregressive moving average models of earthquake records. *Probabilistic Engineering Mechanics*, 3(2):58–63, 1988.

[8] Y. Ma and R. Vidal. Identification of deterministic switched ARX systems via identification of algebraic varieties. In *International Workshop on Hybrid Systems: Computation and Control*, pages 449–465. Springer, 2005.

[9] T. Ogawa, H. Sonoda, S. Ishiwa, and Y. Shigeta. An application of autoregressive model to pattern discrimination of brain electrical activity mapping. *Brain topography*, 6(1):3–11, 1993.

[10] N. Ozay, C. Lagoa, and M. Sznaier. Set membership identification of switched linear systems with known number of subsystems. *Automatica*, 51:180–191, Jan. 2015.

[11] N. Ozay, M. Sznaier, C. M. Lagoa, and O. I. Camps. A Sparsification Approach to Set Membership Identification of Switched Affine Systems. *IEEE Trans. on Aut. Control*, 57(3):634–648, Mar. 2012.

[12] T. Strohmer and R. Vershynin. A randomized kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15(2):262–278, 2009.

[13] R. Vidal. Recursive identification of switched ARX systems. *Automatica*, 44(9):2274–2287, Sept. 2008.

[14] R. Vidal, S. Soatto, Y. Ma, and S. Sastry. An algebraic geometric approach to the identification of a class of linear hybrid systems. In *Proceedings. 42nd IEEE Conference on Decision and Control*, volume 1, pages 167–172. IEEE, 2003.

# A Proofs for Preliminary Results in Section 5.1

## A.1 Proof for Lemma 7

**Proof** Let $[N_R] = \{1, \ldots, N_R\}$. From Assumption 2, $\forall t, \mathbb{E}[n_t] = 0, \mathbb{E}[n_t^2] = \sigma_n^2$, and since $\forall i \in [N_R]$, $r_t(i)$ is a deterministic time step, so we have $\mathbb{E}[n_{r_t(i)}] = 0$ and $\mathbb{E}[n_{r_t(i)}^2] = \sigma_n^2$. Therefore,

$$\begin{aligned}
\mathbb{E}[n_{r_t(l_t)}] &= \mathbb{E}[\mathbb{E}[n_{r_t(l_t)} | \Phi^R_{\widehat{z}_t,t}, l_t = i]] \\
&= \mathbb{E}[\mathbb{E}[n_{r_t(i)} | \Phi^R_{\widehat{z}_t,t}, l_t = i]] \\
&= \mathbb{E}[\mathbb{E}[n_{r_t(i)} | \Phi^R_{\widehat{z}_t,t}]] \\
&= \mathbb{E}[n_{r_t(i)}] \\
&= 0
\end{aligned} \tag{36}$$

where the third equality holds since given $\Phi^R_{\widehat{z}_t,t}$, whether $l_t$ is chosen to be $i$ is independent of $n_{r_t(i)}$. Similarly,

$$\begin{aligned}
\mathbb{E}[n^2_{r_t(l_t)}] &= \mathbb{E}[\mathbb{E}[n^2_{r_t(l_t)}|\Phi^R_{\widehat{z}_t,t}, l_t = i]] \\
&= \mathbb{E}[\mathbb{E}[n^2_{r_t(i)}|\Phi^R_{\widehat{z}_t,t}, l_t = i]] \\
&= \mathbb{E}[\mathbb{E}[n^2_{r_t(i)}|\Phi^R_{\widehat{z}_t,t}]] \\
&= \mathbb{E}[n^2_{r_t(i)}] \\
&= \sigma_n^2
\end{aligned} \tag{37}$$

So (i) is proved.

From Assumption 2, we know $\forall t$, $u_t$ and $n_t$ are independent, so $n_t$ is also independent of $\boldsymbol{\phi}_t$ from (1). Since $\forall i \in [N_R]$, $r_t(i)$ is a deterministic time step, we know $n_{r_t(i)}$ is independent of $\boldsymbol{\phi}_{r_t(i)}$. Therefore

$$\begin{aligned}
\mathbb{E}[\boldsymbol{\phi}_{r_t(l_t)} n_{r_t(l_t)}] &= \mathbb{E}[\mathbb{E}[\boldsymbol{\phi}_{r_t(l_t)} n_{r_t(l_t)}|\Phi^R_{\widehat{z}_t,t}, l_t = i]] \\
&= \mathbb{E}[\mathbb{E}[\boldsymbol{\phi}_{r_t(i)} n_{r_t(i)}|\Phi^R_{\widehat{z}_t,t}, l_t = i]] \\
&= \mathbb{E}[\mathbb{E}[\boldsymbol{\phi}_{r_t(i)} n_{r_t(i)}|\Phi^R_{\widehat{z}_t,t}]] \\
&= \mathbb{E}[\boldsymbol{\phi}_{r_t(i)} n_{r_t(i)}] \\
&= \mathbb{E}[\boldsymbol{\phi}_{r_t(i)}]\mathbb{E}[n_{r_t(i)}] \\
&= 0 \\
&= \mathbb{E}[\boldsymbol{\phi}_{r_t(l_t)}]\mathbb{E}[n_{r_t(l_t)}]
\end{aligned} \tag{38}$$

Therefore, $n_{r_t(l_t)}$ and $\boldsymbol{\phi}_{r_t(l_t)}$ are uncorrelated, and (ii) is proved.

From (i) and (ii), we can see

$$\begin{aligned}
\mathbb{E}\left[\frac{n^2_{r_t(l_t)}}{\|\boldsymbol{\phi}_{r_t(l_t)}\|^2}\right] &= \mathbb{E}[n^2_{r_t(l_t)}]\mathbb{E}\left[\frac{1}{\|\boldsymbol{\phi}_{r_t(l_t)}\|^2}\right] \\
&= \sigma_n^2 \mathbb{E}\left[\mathbb{E}\left[\frac{1}{\|\boldsymbol{\phi}_{r_t(l_t)}\|^2}\middle| \Phi^R_{\widehat{z}_t,t}\right]\right] \\
&= \sigma_n^2 \mathbb{E}\left[\sum_{i\in[N_R]} \frac{1}{\|\boldsymbol{\phi}_{r_t(i)}\|^2} P(l_t = i|\Phi^R_{\widehat{z}_t,t})\right] \\
&= \sigma_n^2 \mathbb{E}\left[\sum_{i\in[N_R]} \frac{1}{\|\boldsymbol{\phi}_{r_t(i)}\|^2} \frac{\|\boldsymbol{\phi}_{r_t(i)}\|^2}{\|\boldsymbol{\Phi}^R_{\widehat{z}_t,t}\|_F^2}\right] \\
&= \sigma_n^2 \mathbb{E}\left[\frac{N_R}{\|\boldsymbol{\Phi}^R_{\widehat{z}_t,t}\|_F^2}\right] \\
&= \sigma_n^2 N_R \mathbb{E}\left[\frac{1}{\|\boldsymbol{\Phi}^R_{\widehat{z}_t,t}\|_F^2}\right]
\end{aligned} \tag{39}$$

From Lemma 6, we have $F_{\min} \le \|\boldsymbol{\Phi}^R_{i,t}\|_F \le F_{\max}$, so (iii) is proved. □

## A.2 Proof for Lemma 8

**Proof** First we prove the lower bound. Let ${\boldsymbol{\Phi}^R_{\widehat{z}_t,t}}^{-1}$ denote the right inverse of $\boldsymbol{\Phi}^R_{\widehat{z}_t,t}$, then accordingly

${\boldsymbol{\Phi}^R_{\widehat{z}_t,t}}^{-1^\mathsf{T}}$ is the left inverse of ${\boldsymbol{\Phi}^R_{\widehat{z}_t,t}}^\mathsf{T}$. As for $\|{\boldsymbol{\Phi}^R_{\widehat{z}_t,t}}^{-1}\|_2$, by definition of matrix norm, we have, for $\forall\, \mathbf{z}$,

$$\|{\boldsymbol{\Phi}^R_{\widehat{z}_t,t}}^{-1}\|_2 = \|{\boldsymbol{\Phi}^R_{\widehat{z}_t,t}}^{-1^\mathsf{T}}\|_2 \geq \frac{\|{\boldsymbol{\Phi}^R_{\widehat{z}_t,t}}^{-1^\mathsf{T}} {\boldsymbol{\Phi}^R_{\widehat{z}_t,t}}^\mathsf{T}\mathbf{z}\|}{\|{\boldsymbol{\Phi}^R_{\widehat{z}_t,t}}^\mathsf{T}\mathbf{z}\|} \tag{40}$$

which gives

$$\|{\boldsymbol{\Phi}^R_{\widehat{z}_t,t}}^\mathsf{T}\mathbf{z}\|^2 \geq \frac{\|\mathbf{z}\|^2}{\|{\boldsymbol{\Phi}^R_{\widehat{z}_t,t}}^{-1}\|_2^2} \tag{41}$$

Expanding LHS and dividing both sides by $\|\boldsymbol{\Phi}^R_{\widehat{z}_t,t}\|_F^2$, we have

$$\sum_{i\in[N_R]} \frac{1}{\|\boldsymbol{\Phi}^R_{\widehat{z}_t,t}\|_F^2} \left(\boldsymbol{\phi}^\mathsf{T}_{r_t(i)}\mathbf{z}\right)^2 \geq \frac{\|\mathbf{z}\|^2}{\|\boldsymbol{\Phi}^R_{\widehat{z}_t,t}\|_F^2 \|{\boldsymbol{\Phi}^R_{\widehat{z}_t,t}}^{-1}\|_2^2} \tag{42}$$

Use the definition $\kappa(\boldsymbol{\Phi}^R_{\widehat{z}_t,t}) = \|\boldsymbol{\Phi}^R_{\widehat{z}_t,t}\|_F \|{\boldsymbol{\Phi}^R_{\widehat{z}_t,t}}^{-1}\|_2$ in Lemma 6, then

$$\sum_{i\in[N_R]} \frac{\|\boldsymbol{\phi}_{r_t(i)}\|^2}{\|\boldsymbol{\Phi}^R_{\widehat{z}_t,t}\|_F^2} \left(\frac{\boldsymbol{\phi}^\mathsf{T}_{r_t(i)}\mathbf{z}}{\|\boldsymbol{\phi}_{r_t(i)}\|}\right)^2 \geq \kappa(\boldsymbol{\Phi}^R_{\widehat{z}_t,t})^{-2}\|\mathbf{z}\|^2 \tag{43}$$

Note that the LHS is equal to $\mathbb{E}\left[(\frac{\boldsymbol{\phi}^\mathsf{T}_{r_t(l_t)}\mathbf{z}}{\|\boldsymbol{\phi}_{r_t(l_t)}\|})^2 \middle| \mathbf{z}, \boldsymbol{\Phi}^R_{\widehat{z}_t,t}\right]$, so

$$\mathbb{E}\left[\left(\frac{\boldsymbol{\phi}^\mathsf{T}_{r_t(l_t)}\mathbf{z}}{\|\boldsymbol{\phi}_{r_t(l_t)}\|}\right)^2 \middle| \mathbf{z}, \boldsymbol{\Phi}^R_{\widehat{z}_t,t}\right] \geq \kappa(\boldsymbol{\Phi}^R_{\widehat{z}_t,t})^{-2}\|\mathbf{z}\|^2 \tag{44}$$

Now taking expectation of both sides again and using smoothing property of expectation, we have

$$\mathbb{E}\left[\left(\frac{\boldsymbol{\phi}^\mathsf{T}_{r_t(l_t)}\mathbf{z}}{\|\boldsymbol{\phi}_{r_t(l_t)}\|}\right)^2\right] \geq \mathbb{E}\left[\kappa(\boldsymbol{\Phi}^R_{\widehat{z}_t,t})^{-2}\|\mathbf{z}\|^2\right] \tag{45}$$

From Lemma 6, we have $\kappa(\boldsymbol{\Phi}^R_{i,t}) \leq \kappa_{\max}$, so

$$\mathbb{E}\left[\left(\frac{\boldsymbol{\phi}^\mathsf{T}_{r_t(l_t)}\mathbf{z}}{\|\boldsymbol{\phi}_{r_t(l_t)}\|}\right)^2\right] \geq \kappa_{\max}^{-2}\mathbb{E}\left[\|\mathbf{z}\|^2\right] \tag{46}$$

As for the upper bound, note that for $\forall\mathbf{z}$,

$$\|\boldsymbol{\Phi}^R_{\widehat{z}_t,t}\|_2 = \|{\boldsymbol{\Phi}^R_{\widehat{z}_t,t}}^\mathsf{T}\|_2 \geq \frac{\|{\boldsymbol{\Phi}^R_{\widehat{z}_t,t}}^\mathsf{T}\mathbf{z}\|}{\|\mathbf{z}\|} \tag{47}$$

which gives

$$\|{\boldsymbol{\Phi}^R_{\widehat{z}_t,t}}^\mathsf{T}\mathbf{z}\|^2 \leq \|\mathbf{z}\|^2\|\boldsymbol{\Phi}^R_{\widehat{z}_t,t}\|_2^2 \tag{48}$$

Then using similar technique as the proof for lower bound, we could have

$$\mathbb{E}\left[\left(\frac{\boldsymbol{\phi}^\mathsf{T}_{r_t(l_t)}\mathbf{z}}{\|\boldsymbol{\phi}_{r_t(l_t)}\|}\right)^2\right] \leq \xi_{\min}^{-2}\mathbb{E}[\|\mathbf{z}\|^2] \tag{49}$$

□

# B Proofs for Valid Upper Bound Results in Section 5.2

## B.1 Proof for Theorem 9

**Proof** From the setup statement in Theorem 9, we could see that to show the theorem, it suffices to consider there's only one subsystem, namely subsystem i, in the hybrid SARX model. Then, $c_i = t$, and the setup condition in theorem statement can be met automatically when $t \geq N_C$.

When $t \geq N_C$, i.e. $c_i \geq N_C$: the data $\boldsymbol{\phi}_t^*, y_t^*, \eta_t^*$ we choose to update the candidate in Line 21 of Algorithm 1 is formed in Line 20, where we sample a column index $l_t$ from the matrix $\boldsymbol{\Phi}_{i,t}^R$ in Line 19 of Algorithm 1. Since $\boldsymbol{\Phi}_{i,t}^R$ is a matrix with columns being data vectors collected at different time, we essentially sampled a time index. Let $r_t(l_t)$ denote the true time index corresponding to the column $l_t$ we sample at time $t$. So $\boldsymbol{\phi}_t^* = \boldsymbol{\phi}_{r_t(l_t)}, y_t^* = y_{r_t(l_t)}$. In addition, we let $n_t^* = n_{r_t(l_t)}$.

Plugging definition $\boldsymbol{\epsilon}_{i,t} = \mathbf{w}_i - \widehat{\mathbf{w}}_{i,t}$ and system equation $y_t^* = \mathbf{w}_i^\mathsf{T}\boldsymbol{\phi}_t^* + n_t^*$ into update rule $\widehat{\mathbf{w}}_t = \widehat{\mathbf{w}}_{t-1} - \eta_t^*\boldsymbol{\phi}_t^*(\widehat{\mathbf{w}}_{t-1}^\mathsf{T}\boldsymbol{\phi}_t^* - y_t^*)$, we could have

$$\begin{aligned}\boldsymbol{\epsilon}_{i,t} &= (I - \eta_t^*\boldsymbol{\phi}_t^*\boldsymbol{\phi}_t^{*\mathsf{T}})\boldsymbol{\epsilon}_{i,t-1} - \eta_t^*\boldsymbol{\phi}_t^* n_t^* \\ &= \boldsymbol{\epsilon}_{i,t-1} - \eta_t^*\boldsymbol{\phi}_t^*\boldsymbol{\phi}_t^{*\mathsf{T}}\boldsymbol{\epsilon}_{i,t-1} - \eta_t^*\boldsymbol{\phi}_t^* n_t^*\end{aligned} \tag{50}$$

Replacing the first term $\boldsymbol{\epsilon}_{i,t-1}$ on the RHS of (50) by $\boldsymbol{\epsilon}_{i,t-1} = \boldsymbol{\epsilon}_{i,t-2} - \eta_{t-1}^*\boldsymbol{\phi}_{t-1}^*\boldsymbol{\phi}_{t-1}^{*\mathsf{T}}\boldsymbol{\epsilon}_{i,t-1} - \eta_{t-1}^*\boldsymbol{\phi}_{t-1}^* n_{t-1}^*$ and repeat this procedure recursively, we could finally have

$$\boldsymbol{\epsilon}_{i,t} = \boldsymbol{\epsilon}_{i,t-N_C} - \sum_{j=t}^{t-(N_C-1)} \eta_j^*\boldsymbol{\phi}_j^*\boldsymbol{\phi}_j^{*\mathsf{T}}\boldsymbol{\epsilon}_{i,j-1} - \sum_{j=t}^{t-(N_C-1)} \eta_j^*\boldsymbol{\phi}_j^* n_j^* \tag{51}$$

Consider the LHS of (51), by addition and subtraction, we could see

$$\boldsymbol{\epsilon}_{i,t} = \sum_{j=t}^{t-(N_C-1)} \eta_j^*\boldsymbol{\phi}_j^*\boldsymbol{\phi}_j^{*\mathsf{T}}\boldsymbol{\epsilon}_{i,t} + \boldsymbol{\epsilon}_{i,t} - \sum_{j=t}^{t-(N_C-1)} \eta_j^*\boldsymbol{\phi}_j^*\boldsymbol{\phi}_j^{*\mathsf{T}}\boldsymbol{\epsilon}_{i,t} \tag{52}$$

Combining (51) and (52), we have

$$\sum_{j=t}^{t-(N_C-1)} \eta_j^*\boldsymbol{\phi}_j^*\boldsymbol{\phi}_j^{*\mathsf{T}}\boldsymbol{\epsilon}_{i,t} = \left((\widehat{\mathbf{w}}_{i,t} - \widehat{\mathbf{w}}_{i,t-N_C}) - \sum_{j=t}^{t-(N_C-1)} \eta_j^*\boldsymbol{\phi}_j^*\boldsymbol{\phi}_j^{*\mathsf{T}}(\widehat{\mathbf{w}}_{i,t} - \widehat{\mathbf{w}}_{i,j-1})\right) - \sum_{j=t}^{t-(N_C-1)} \eta_j^*\boldsymbol{\phi}_j^* n_j^* \tag{53}$$

Now using the notations and operator defined in Line 9 to Line 13 in Algorithm 2, and let $\mathbf{n}_t = [n_{t-(N_C-1)}, \ldots, n_{t-1}, n_t]^\mathsf{T}$ we have a neat form:

$$\boldsymbol{\Phi}_{i,t}^C\mathbf{H}\boldsymbol{\Phi}_{i,t}^{C\,\mathsf{T}}\boldsymbol{\epsilon}_{i,t} = \left[\Delta\widehat{\mathbf{w}} - \boldsymbol{\Phi}_{i,t}^C \square \left(\boldsymbol{\Phi}_{i,t}^C, \Delta\widehat{\mathbf{W}}\right)\right] - \boldsymbol{\Phi}_{i,t}^C\mathbf{H}\mathbf{n}_t \tag{54}$$

Now, we want to show the invertibility of matrix $\boldsymbol{\Phi}_{i,t}^C\mathbf{H}\boldsymbol{\Phi}_{i,t}^{C\,\mathsf{T}}$. Define

$$\widetilde{\boldsymbol{\Phi}} = \left[\sqrt{\eta_{t-(N_C-1)}^*}\boldsymbol{\phi}_{t-(N_C-1)}^*, \ldots, \sqrt{\eta_{t-1}^*}\boldsymbol{\phi}_{t-1}^*, \sqrt{\eta_t^*}\boldsymbol{\phi}_t^*\right]_{n\times N_C} \tag{55}$$

, then we could see that

$$\boldsymbol{\Phi}_{i,t}^C\mathbf{H}\boldsymbol{\Phi}_{i,t}^{C\,\mathsf{T}} = \sum_{j=t}^{t-(N_C-1)} \eta_j^*\boldsymbol{\phi}_j^*\boldsymbol{\phi}_j^{*\mathsf{T}} = \widetilde{\boldsymbol{\Phi}}\widetilde{\boldsymbol{\Phi}}^\mathsf{T} \tag{56}$$

, so it suffices to show $\widetilde{\mathbf{\Phi}}$ has $n$ linearly independent columns. Note that $\eta^* > 0$ and $\mathbf{\Phi}_{i,t}^C = [\boldsymbol{\phi}_{t-(N_C-1)}^*, \ldots, \boldsymbol{\phi}_{t-1}^*, \boldsymbol{\phi}_t^*]$, so it further suffices to show $\mathbf{\Phi}_{i,t}^C$ has $n$ linearly independent columns.

Since $\mathbf{\Phi}_{i,t}^C$ is composed of $N_C$ columns sample from different matrices $\mathbf{\Phi}_{i,t}^R \in \mathbb{R}^{n\times N_R}$ from time $t-(N_C-1)$ to $t$, then the condition $N_C \geq N_R^2$ requirement in Algorithm 1 guarantees that there are at least $N_R$ columns in $\mathbf{\Phi}_{i,t}^C$ such that their generating time are different. Then from Assumption 4, we know these $n$ columns must be linearly independent, and so are the corresponding columns in $\widetilde{\mathbf{\Phi}}$. Therefore, $\mathbf{\Phi}_{i,t}^C \mathbf{H} \mathbf{\Phi}_{i,t}^{C\,\mathsf{T}}$ is invertible.

With this result, (54) becomes:

$$\boldsymbol{\epsilon}_{i,t} = (\mathbf{\Phi}_{i,t}^C \mathbf{H} \mathbf{\Phi}_{i,t}^{C\,\mathsf{T}})^{-1} \left[\Delta\widehat{\mathbf{w}} - \mathbf{\Phi}_{i,t}^C \square \left(\mathbf{\Phi}_{i,t}^C, \Delta\widehat{\mathbf{W}}\right)\right] - (\mathbf{\Phi}_{i,t}^C \mathbf{H} \mathbf{\Phi}_{i,t}^{C\,\mathsf{T}})^{-1} \mathbf{\Phi}_{i,t}^C \mathbf{H} \mathbf{n}_t \tag{57}$$

then using definition of $\mathbf{A}$ and $\mathbf{b}$ in Algorithm 2, we have

$$\boldsymbol{\epsilon}_{i,t} = \mathbf{b} - \mathbf{A}\mathbf{n}_t \tag{58}$$

Since $\|\mathbf{n}_t\|_\infty \leq n_{\max}$, if we define the set of vertices $V = \{[\pm n_{\max}, \pm n_{\max}, \ldots, \pm n_{\max}]_{N_C}^{\mathsf{T}}\}$, then it's easy to see

$$\|\mathbf{b} - \mathbf{A}\mathbf{n}_t\| \leq \max_{\mathbf{n}\in V} \|A\mathbf{n} - \mathbf{b}\| \tag{59}$$

And since $\epsilon_{i,t}^u \equiv \max_{\mathbf{n}\in V} \|A\mathbf{n} - \mathbf{b}\|$, we could finally see

$$\epsilon_{i,t}^u \geq \|\boldsymbol{\epsilon}_{i,t}\| \tag{60}$$

□

# C Proofs for Partial Convergence in Section 5.3

## C.1 Proof for Lemma 10

**Proof** According Algorithm 1, when $t \leq N_R - 1$, we know $c_i < N_R$, and the update rule is given by $\widehat{\mathbf{w}}_{i,t} = \widehat{\mathbf{w}}_{i,t-1} - \eta_t^* \boldsymbol{\phi}_t (\widehat{\mathbf{w}}_{i,t}^{\mathsf{T}} \boldsymbol{\phi}_t - y_t)$. Since $\boldsymbol{\epsilon}_{i,t} = \mathbf{w}_i - \widehat{\mathbf{w}}_{i,t}$ and $\mathbf{w}_i^{\mathsf{T}} \boldsymbol{\phi}_t + n_t = y_t$, we can derive the following error dynamics through simple algebra:

$$\boldsymbol{\epsilon}_{i,t} = (I - \eta_t^* \boldsymbol{\phi}_t \boldsymbol{\phi}_t^{\mathsf{T}}) \boldsymbol{\epsilon}_{i,t-1} - \eta_t^* \boldsymbol{\phi}_t n_t \tag{61}$$

Notice that in Algorithm 1, we set $\eta_t^* = \|\boldsymbol{\phi}_t\|^{-2}$, so

$$\boldsymbol{\epsilon}_{i,t} = \left(I - \frac{\boldsymbol{\phi}_t \boldsymbol{\phi}_t^{\mathsf{T}}}{\|\boldsymbol{\phi}_t\|^2}\right) \boldsymbol{\epsilon}_{i,t-1} - \frac{\boldsymbol{\phi}_t n_t}{\|\boldsymbol{\phi}_t\|^2} \tag{62}$$

Taking norm squares of both sides,

$$\|\boldsymbol{\epsilon}_{i,t}\|^2 = \left\| \left(I - \frac{\boldsymbol{\phi}_t \boldsymbol{\phi}_t^{\mathsf{T}}}{\|\boldsymbol{\phi}_t\|^2}\right) \frac{\boldsymbol{\epsilon}_{i,t-1}}{\|\boldsymbol{\epsilon}_{i,t-1}\|} \right\|^2 \|\boldsymbol{\epsilon}_{i,t-1}\|^2 + \frac{n_t^2}{\|\boldsymbol{\phi}_t\|^2} \tag{63}$$

of which the cross term vanishes because it's equal to 0. Now consider the first term in (63),

$$\begin{aligned} &\left\| \left(I - \frac{\boldsymbol{\phi}_t \boldsymbol{\phi}_t^{\mathsf{T}}}{\|\boldsymbol{\phi}_t\|^2}\right) \frac{\boldsymbol{\epsilon}_{i,t-1}}{\|\boldsymbol{\epsilon}_{i,t-1}\|} \right\|^2 \\ =&\frac{\boldsymbol{\epsilon}_{i,t-1}^{\mathsf{T}}}{\|\boldsymbol{\epsilon}_{i,t-1}\|} \left(I - \frac{\boldsymbol{\phi}_t \boldsymbol{\phi}_t^{\mathsf{T}}}{\|\boldsymbol{\phi}_t\|^2}\right) \frac{\boldsymbol{\epsilon}_{i,t-1}}{\|\boldsymbol{\epsilon}_{i,t-1}\|} \\ =&1 - \left(\frac{\boldsymbol{\phi}_t^{\mathsf{T}} \boldsymbol{\epsilon}_{i,t-1}}{\|\boldsymbol{\phi}_t\| \|\boldsymbol{\epsilon}_{i,t-1}\|}\right)^2 \end{aligned} \tag{64}$$

Plugging (64) into (63), then

$$\|\boldsymbol{\epsilon}_{i,t}\|^2 = \left[1 - \left(\frac{\boldsymbol{\phi}_t^\intercal \boldsymbol{\epsilon}_{i,t-1}}{\|\boldsymbol{\phi}_t\|\|\boldsymbol{\epsilon}_{i,t-1}\|}\right)^2\right] \|\boldsymbol{\epsilon}_{i,t-1}\|^2 + \frac{n_t^2}{\|\boldsymbol{\phi}_t\|^2} \tag{65}$$

Since $0 \leq (\frac{\boldsymbol{\phi}_t^\intercal \boldsymbol{\epsilon}_{i,t-1}}{\|\boldsymbol{\phi}_t\|\|\boldsymbol{\epsilon}_{i,t-1}\|})^2 \leq 1$, we have

$$\frac{n_t^2}{\|\boldsymbol{\phi}_t\|^2} \leq \|\boldsymbol{\epsilon}_{i,t}\|^2 \leq \|\boldsymbol{\epsilon}_{i,t-1}\|^2 + \frac{n_t^2}{\|\boldsymbol{\phi}_t\|^2} \tag{66}$$

Since $\|\boldsymbol{\phi}_t\|^2 \leq \phi_{\max}^2$ and $\frac{\|\boldsymbol{\phi}_t\|}{|n_t|} \geq S_{\min}$ according to Assumption 3, then

$$\frac{n_t^2}{\phi_{\max}^2} \leq \|\boldsymbol{\epsilon}_{i,t}\|^2 \leq \|\boldsymbol{\epsilon}_{i,t-1}\|^2 + \frac{1}{S_{\min}^2} \tag{67}$$

Now taking expectation of both sides of (67),

$$\frac{\sigma_n^2}{\phi_{\max}^2} \leq \mathbb{E}\left[\|\boldsymbol{\epsilon}_{i,t}\|^2\right] \leq \mathbb{E}\left[\|\boldsymbol{\epsilon}_{i,t-1}\|^2\right] + \frac{1}{S_{\min}^2} \tag{68}$$

Now if we apply (68) recursively, we could finally prove (9) in the lemma. □

## C.2 Proof for Lemma 11

**Proof** From Algorithm 1, when $t \geq N_R$, $c_i \geq N_R$. And to update estimate, we first sample a column index $l_t$ from the matrix $\boldsymbol{\Phi}_{i,t}^R$ in Line 19 of Algorithm 1. Since $\boldsymbol{\Phi}_{i,t}^R$ is a matrix with columns being data vectors collected at different time, we essentially sampled a time index. Let $r_t(l_t)$ denote the true time index corresponding to the column $l_t$ we sample at time $t$. So the corresponding $\boldsymbol{\phi}_t^*, y_t^*$ are actually $\boldsymbol{\phi}_{r_{t(l)}}, y_{r_t(l_t)}$. And we have the update rule $\widehat{\mathbf{w}}_{i,t} = \widehat{\mathbf{w}}_{i,t-1} - \eta_t^* \boldsymbol{\phi}_{r_t(l_t)}(\widehat{\mathbf{w}}_{t-1}^\intercal \boldsymbol{\phi}_{r_t(l_t)} - y_{r_t(l_t)})$. So following (63) in proof for Lemma 10, we have

$$\|\boldsymbol{\epsilon}_{i,t}\|^2 = \left[1 - \left(\frac{\boldsymbol{\phi}_{r_t(l_t)}^\intercal \boldsymbol{\epsilon}_{i,t-1}}{\|\boldsymbol{\phi}_{r_t(l_t)}\|\|\boldsymbol{\epsilon}_{i,t-1}\|}\right)^2\right] \|\boldsymbol{\epsilon}_{i,t-1}\|^2 + \frac{n_{r_t(l_t)}^2}{\|\boldsymbol{\phi}_{r_t(l_t)}\|^2} \tag{69}$$

Now take expectation of both sides of (69),

$$\mathbb{E}[\|\boldsymbol{\epsilon}_{i,t}\|^2] = \mathbb{E}\left[\|\boldsymbol{\epsilon}_{i,t-1}\|^2\right] - \mathbb{E}\left[\left(\frac{\boldsymbol{\phi}_{r_t(l_t)}^\intercal \boldsymbol{\epsilon}_{i,t-1}}{\|\boldsymbol{\phi}_{r_t(l_t)}\|}\right)^2\right] + \mathbb{E}\left[\frac{n_{r_t(l_t)}^2}{\|\boldsymbol{\phi}_{r_t(l_t)}\|^2}\right] \tag{70}$$

Applying Lemma 8 and Lemma 7(iii), we have

$$\begin{cases} \mathbb{E}\left[\|\boldsymbol{\epsilon}_{i,t}\|^2\right] \geq \left(1 - \xi_{\min}^{-2}\right) \mathbb{E}\left[\|\boldsymbol{\epsilon}_{i,t-1}\|^2\right] + \frac{N_R}{F_{\max}^2}\sigma_n^2 \\ \mathbb{E}\left[\|\boldsymbol{\epsilon}_{i,t}\|^2\right] \leq \left(1 - \kappa_{\max}^{-2}\right) \mathbb{E}\left[\|\boldsymbol{\epsilon}_{i,t-1}\|^2\right] + \frac{N_R}{F_{\min}^2}\sigma_n^2 \end{cases} \tag{71}$$

Finally, apply (71) recursively, we could end up getting (11) and (10) in Lemma 11 □

## C.3 Proof for Theorem 12

**Proof** When there is only one subsystem, Lemma 10 and Lemma 11 selectively characterize the behavior of estimation error when $t < N_R$ and $t \geq N_R$. By combining them and replacing the universal time index $t$ in Lemma 10 and Lemma 11 with the individual time index $r(i,t)$ for subsystem $i$, we can have this theorem. □

# D Proofs for Local Convergence in Section 5.4

## D.1 Proof for Lemma 14

**Proof** Let $\boldsymbol{\chi}_t$ be "All data $\{\boldsymbol{\phi}, y\}$ assigned to candidate i up to time t and all the data $\{\boldsymbol{\phi}^*, y^*\}$ we used to update $\widehat{\mathbf{w}}_i$ up to time $t$". Then we can see $\boldsymbol{\chi}_t \subset \boldsymbol{\chi}_{t+1}$ and

$$\mathbb{E}\left[\|\boldsymbol{\epsilon}_{i,t}\|^2 \middle| \boldsymbol{\chi}_t\right] = \mathbb{E}\left[\|\mathbf{w}_i - \widehat{\mathbf{w}}_{i,t}\|^2 \middle| \boldsymbol{\chi}_t\right] = \|\mathbf{w}_i - \widehat{\mathbf{w}}_{i,t}\|^2 = \|\boldsymbol{\epsilon}_{i,t}\|^2 \tag{72}$$

where the second equality holds since knowing $\boldsymbol{\chi}_t$ we know update process of $\widehat{\mathbf{w}}_{i,0}, \widehat{\mathbf{w}}_{i,1}, \ldots, \widehat{\mathbf{w}}_{i,t}$ completely. (72) says the randomness of $\|\boldsymbol{\epsilon}_{i,t}\|^2$ completely comes from $\boldsymbol{\chi}_t$.

When $t \geq N_R$, (69) characterizes the error dynamics, and we restate it here:

$$\|\boldsymbol{\epsilon}_{i,t}\|^2 = \|\boldsymbol{\epsilon}_{i,t-1}\|^2 - \left(\frac{\boldsymbol{\phi}_{r_t(l_t)}^\intercal \boldsymbol{\epsilon}_{i,t-1}}{\|\boldsymbol{\phi}_{r_t(l_t)}\|}\right)^2 + \frac{n_{r_t(l_t)}^2}{\|\boldsymbol{\phi}_{r_t(l_t)}\|^2} \tag{73}$$

Now take $\mathbb{E}[\cdot|\boldsymbol{\chi}_{t-1}]$ on both sides of (73), we have

$$\mathbb{E}\left[\|\boldsymbol{\epsilon}_{i,t}\|^2 | \boldsymbol{\chi}_{t-1}\right] = \|\boldsymbol{\epsilon}_{i,t-1}\|^2 - \mathbb{E}\left[\left(\frac{\boldsymbol{\phi}_{r_t(l_t)}^\intercal \boldsymbol{\epsilon}_{i,t-1}}{\|\boldsymbol{\phi}_{r_t(l_t)}\|}\right)^2 \middle| \boldsymbol{\chi}_{t-1}\right] + \mathbb{E}\left[\frac{n_{r_t(l_t)}^2}{\|\boldsymbol{\phi}_{r_t(l_t)}\|^2} \middle| \boldsymbol{\chi}_{t-1}\right] \tag{74}$$

First consider $\mathbb{E}\left[\left(\frac{\boldsymbol{\phi}_{r_t(l_t)}^\intercal \boldsymbol{\epsilon}_{i,t-1}}{\|\boldsymbol{\phi}_{r_t(l_t)}\|}\right)^2 \middle| \boldsymbol{\chi}_{t-1}\right]$, we have

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{\boldsymbol{\phi}_{r_t(l_t)}^\intercal \boldsymbol{\epsilon}_{i,t-1}}{\|\boldsymbol{\phi}_{r_t(l_t)}\|}\right)^2 \middle| \boldsymbol{\chi}_{t-1}\right] \\
=&\mathbb{E}\left[\left(\frac{\boldsymbol{\phi}_{r_t(l_t)}^\intercal \boldsymbol{\epsilon}_{i,t-1}}{\|\boldsymbol{\phi}_{r_t(l_t)}\|}\right)^2 \middle| \boldsymbol{\chi}_{t-1}, \boldsymbol{\epsilon}_{i,t-1}\right] \\
=&\mathbb{E}\left[\left(\frac{\boldsymbol{\phi}_{r_t(l_t)}^\intercal \boldsymbol{\epsilon}_{i,t-1}}{\|\boldsymbol{\phi}_{r_t(l_t)}\|}\right)^2 \middle| \boldsymbol{\Phi}_{i,t-1}^R, \boldsymbol{\epsilon}_{i,t-1}\right]
\end{aligned} \tag{75}$$

where the first equality holds as $\boldsymbol{\epsilon}_{i,t-1}$ is nonrandom given $\boldsymbol{\chi}_{t-1}$; the second equality holds for the following reason: $\boldsymbol{\Phi}_{i,t-1}^R$ can be determined from $\boldsymbol{\chi}_{t-1}$, and $\boldsymbol{\phi}_{r_t(l_t)}$ is drawn from $\boldsymbol{\Phi}_{i,t-1}^R$ in an independent experiment, so $\boldsymbol{\phi}_{r_t(l_t)}$ depends on $\boldsymbol{\chi}_{t-1}$ only through $\boldsymbol{\Phi}_{i,t-1}^R$. Note that RHS in (75) can follow similar argument from (44) to (49), then

$$\kappa_{\max}^{-2}\|\boldsymbol{\epsilon}_{i,t-1}\|^2 \leq \mathbb{E}\left[\left(\frac{\boldsymbol{\phi}_{r_t(l_t)}^\intercal \boldsymbol{\epsilon}_{i,t-1}}{\|\boldsymbol{\phi}_{r_t(l_t)}\|}\right)^2 \middle| \boldsymbol{\chi}_{t-1}\right] \leq \xi_{\min}^{-2}\|\boldsymbol{\epsilon}_{i,t-1}\|^2 \tag{76}$$

Then consider $\mathbb{E}\left[\frac{n_{r_t(l_t)}^2}{\|\boldsymbol{\phi}_{r_t(l_t)}\|^2} \middle| \boldsymbol{\chi}_{t-1}\right]$. By Assumption 3, we have

$$\mathbb{E}\left[\frac{n_{r_t(l_t)}^2}{\|\boldsymbol{\phi}_{r_t(l_t)}\|^2} \middle| \boldsymbol{\chi}_{t-1}\right] \leq \frac{1}{S_{\min}^2} \tag{77}$$

Applying (77) and (76) to (74), we have

$$\mathbb{E}\left[\|\boldsymbol{\epsilon}_{i,t}\|^2 | \boldsymbol{\chi}_{t-1}\right] \leq \left(1 - \kappa_{\max}^{-2}\right)\|\boldsymbol{\epsilon}_{i,t-1}\|^2 + \frac{1}{S_{\min}^2} \tag{78}$$

Now we want to show $\mathbb{E}\left[\|\boldsymbol{\epsilon}_{i,t}\|^2|\boldsymbol{\chi}_{t-1}\right] \leq \|\boldsymbol{\epsilon}_{i,t-1}\|^2$. The general form of (73) for $\forall t$ is

$$\|\boldsymbol{\epsilon}_{i,t}\|^2 = \left[1 - \left(\frac{\boldsymbol{\phi}_t^{*\intercal}\boldsymbol{\epsilon}_{i,t-1}}{\|\boldsymbol{\phi}_t^*\|\|\boldsymbol{\epsilon}_{i,t-1}\|}\right)^2\right]\|\boldsymbol{\epsilon}_{i,t-1}\|^2 + \frac{n_t^{*2}}{\|\boldsymbol{\phi}_t^*\|^2} \tag{79}$$

where $\boldsymbol{\phi}_t^*$ is defined in Line 17 and 20 in Algorithm 1, and we let $n_t^*$ denote the noise corresponding to data $\{\boldsymbol{\phi}_t^*, y_t^*\}$. Since $\left(\frac{\boldsymbol{\phi}_t^{*\intercal}\boldsymbol{\epsilon}_{i,t-1}}{\|\boldsymbol{\phi}_t^*\|\|\boldsymbol{\epsilon}_{i,t-1}\|}\right)^2 \leq 1$ and by Assumption 13, then $\forall t$

$$\|\boldsymbol{\epsilon}_{i,t}\|^2 \geq \frac{n_t^{*2}}{\|\boldsymbol{\phi}_t^*\|^2} \geq \frac{1}{S_{\max}^2} \geq \frac{1}{\kappa_{\max}^2 S_{\min}^2} \tag{80}$$

So for $\forall t \geq 2$,

$$\|\boldsymbol{\epsilon}_{i,t-1}\|^2 \geq \frac{1}{\kappa_{\max}^2 S_{\min}^2} \tag{81}$$

Following (81), (78) gives

$$\mathbb{E}\left[\|\boldsymbol{\epsilon}_{i,t}\|^2|\boldsymbol{\chi}_{t-1}\right] \leq \|\boldsymbol{\epsilon}_{i,t-1}\|^2 \tag{82}$$

So, we can see $\{\|\boldsymbol{\epsilon}_{i,t}\|^2, t \geq N_R - 1\}$ is a supermartingale with respect to $\{\boldsymbol{\chi}_t, t \geq N_R - 1\}$. Finally, using supermartingale maxima inequality, we have (15) directly. □

## D.2 Proof for Lemma 15

**Proof** In the first phase of the algorithm when $t \leq N_R - 1$, using (68) recursively, we have

$$\mathbb{E}\left[\|\boldsymbol{\epsilon}_{i,t}\|^2\right] \leq \|\boldsymbol{\epsilon}_{i,0}\|^2 + \frac{t}{S_{\min}^2} \leq \epsilon_0^2 + \frac{N_R}{S_{\min}^2} \tag{83}$$

Define $\epsilon_a$ such that $\epsilon_a^2 = \sqrt{\epsilon'^2 N_R\left(\epsilon_0^2 + \frac{N_R}{S_{\min}^2}\right)}$, then from the condition in the statement of Lemma 15, we can see $\epsilon_a^2 \leq \epsilon'^2$. According to Markov inequality, we have

$$P\left(\|\boldsymbol{\epsilon}_{i,t}\|^2 \leq \epsilon_a^2\right) \geq 1 - \frac{1}{\epsilon_a^2}\left(\epsilon_0^2 + \frac{N_R}{S_{\min}^2}\right) \tag{84}$$

Then using union bound, we have

$$P\left(\bigcap_{\tau=1}^{N_R-1}\left\{\|\boldsymbol{\epsilon}_{i,\tau}\|^2 \leq \epsilon_a^2\right\}\right) \geq 1 - \frac{N_R}{\epsilon_a^2}\left(\epsilon_0^2 + \frac{N_R}{S_{\min}^2}\right) \tag{85}$$

Now for $t \geq N_R$, we have

$$
\begin{aligned}
& P\left(\bigcap_{\tau=1}^{t}\left\{\|\boldsymbol{\epsilon}_{i,\tau}\|^2 \leq \epsilon'^2\right\}\right) \\
\geq & P\left(\bigcap_{\tau=1}^{N_R-1}\left\{\|\boldsymbol{\epsilon}_{i,\tau}\|^2 \leq \epsilon_a^2\right\}, \bigcap_{\tau=N_R}^{t}\left\{\|\boldsymbol{\epsilon}_{i,\tau}\|^2 \leq \epsilon'^2\right\}\right) \\
= & P\left(\bigcap_{\tau=1}^{N_R-1}\left\{\|\boldsymbol{\epsilon}_{i,\tau}\|^2 \leq \epsilon_a^2\right\}\right) \cdot P\left(\bigcap_{\tau=N_R}^{t}\left\{\|\boldsymbol{\epsilon}_{i,\tau}\|^2 \leq \epsilon'^2\right\} \,\middle|\, \|\boldsymbol{\epsilon}_{i,N_R-1}\|^2 \leq \epsilon_a^2, \bigcap_{\tau=1}^{N_R-2}\left\{\|\boldsymbol{\epsilon}_{i,\tau}\|^2 \leq \epsilon_a^2\right\}\right) \\
\geq & \left[1-\frac{N_R}{\epsilon_a^2}\left(\epsilon_0^2+\frac{N_R}{S_{\min}^2}\right)\right]\left[1-\frac{\epsilon_a^2}{\epsilon'^2}\right] \\
\geq & 1-\frac{N_R}{\epsilon_a^2}\left(\epsilon_0^2+\frac{N_R}{S_{\min}^2}\right)-\frac{\epsilon_a^2}{\epsilon'^2} \\
= & 1-2\sqrt{\frac{N_R}{\epsilon'^2}\left(\epsilon_0^2+\frac{N_R}{S_{\min}^2}\right)}
\end{aligned} \tag{86}
$$

where the first inequality holds since $\epsilon_a^2 \leq \epsilon'^2$; the third inequality holds by applying (85) and (15) in Lemma 14; the last line holds by plugging in the definition of $\epsilon_a$. □

### D.3 Proof for Lemma 16

**Proof** In Line 11 of Algorithm 1, we make assignment according to

$$
\widehat{z}_t = \arg\min_i \; r_i \cdot \max\left(1, \alpha\frac{\|\widetilde{\mathbf{w}}_{i,t}-\widehat{\mathbf{w}}_{i,t-1}\|}{2(\epsilon_{i,t-1}^u+\nu)}\right)^\beta \tag{87}
$$

So, if data $\{\boldsymbol{\phi}_t, y_t\}$ is generated by subsystem $i$, i.e. $z_t = i$, and we want it to be assigned to candidate $i$ according to Line 9 in Algorithm 1, it suffices to have $\forall j \neq i$

$$
r_j \cdot \max\left(1, \alpha\frac{\|\widetilde{\mathbf{w}}_{j,t}-\widehat{\mathbf{w}}_{j,t-1}\|}{2(\epsilon_{j,t-1}^u+\nu)}\right)^\beta > \; r_i \cdot \max\left(1, \alpha\frac{\|\widetilde{\mathbf{w}}_{i,t}-\widehat{\mathbf{w}}_{i,t-1}\|}{2(\epsilon_{i,t-1}^u+\nu)}\right)^\beta \tag{88}
$$

From Line 9 in Algorithm 1, we can see $\|\widetilde{\mathbf{w}}_{i,t}-\widehat{\mathbf{w}}_{i,t-1}\| = \|\boldsymbol{\phi}_t\|^{-1}|\widehat{\mathbf{w}}_{i,t-1}^\intercal\boldsymbol{\phi}_t - y_t|$. So, (88) is equivalent to

$$
r_j \cdot \max\left(1, \alpha\frac{\|\boldsymbol{\phi}_t\|^{-1}|\widehat{\mathbf{w}}_{j,t-1}^\intercal\boldsymbol{\phi}_t - y_t|}{2(\epsilon_{j,t-1}^u+\nu)}\right)^\beta > \; r_i \cdot \max\left(1, \alpha\frac{\|\boldsymbol{\phi}_t\|^{-1}|\widehat{\mathbf{w}}_{i,t-1}^\intercal\boldsymbol{\phi}_t - y_t|}{2(\epsilon_{i,t-1}^u+\nu)}\right)^\beta \tag{89}
$$

Since the LHS of (89) is larger than or equal to $r_j$, to show (89), it suffices to show

$$
r_j > \; r_i \cdot \max\left(1, \frac{\alpha}{2}\cdot\frac{\|\boldsymbol{\phi}_t\|^{-1}|\widehat{\mathbf{w}}_{i,t-1}^\intercal\boldsymbol{\phi}_t - y_t|}{\epsilon_{i,t-1}^u+\nu}\right)^\beta \tag{90}
$$

Note that we could have the following

$$\begin{aligned}
&\frac{\|\boldsymbol{\phi}_t\|^{-1}|\widehat{\mathbf{w}}_{i,t-1}^{\mathsf{T}}\boldsymbol{\phi}_t - y_t|}{\epsilon_{i,t-1}^u + \nu} \\
=&\|\boldsymbol{\phi}_t\|^{-1}\frac{|n_t + \boldsymbol{\epsilon}_{i,t-1}^{\mathsf{T}}\boldsymbol{\phi}_t|}{\epsilon_{i,t-1}^u + \nu} \\
\leq&\|\boldsymbol{\phi}_t\|^{-1}\frac{|n_t| + \|\boldsymbol{\epsilon}_{i,t-1}\|\|\boldsymbol{\phi}_t\|}{\epsilon_{i,t-1}^u + \nu} \\
=&\frac{|n_t|}{\|\boldsymbol{\phi}_t\|(\epsilon_{i,t-1}^u + \nu)} + \frac{\|\boldsymbol{\epsilon}_{i,t-1}\|}{\epsilon_{i,t-1}^u + \nu} \\
\leq&\frac{|n_t|}{\|\boldsymbol{\phi}_t\|(\|\boldsymbol{\epsilon}_{i,t-1}\| + \nu)} + 1
\end{aligned} \tag{91}$$

where the first line holds since $y_t = \mathbf{w}_i^{\mathsf{T}}\boldsymbol{\phi}_t + n_t$, and $\boldsymbol{\epsilon}_{i,t-1} = \mathbf{w}_i - \widehat{\mathbf{w}}_{i,t-1}$; the last line holds since $\epsilon_{i,t-1}^u > \|\boldsymbol{\epsilon}_{i,t-1}\|$ from Theorem 9. Since we let $\alpha = 2, \beta = 1$, to ensure (90) holds, it suffices to ensure the following holds:

$$r_j > r_i\left(\frac{|n_t|}{\|\boldsymbol{\phi}_t\|(\|\boldsymbol{\epsilon}_{i,t-1}\| + \nu)} + 1\right) \tag{92}$$

Since $r_i = \|\boldsymbol{\phi}_t\|^{-1}|y_t - \widehat{\mathbf{w}}_{i,t-1}^{\mathsf{T}}\boldsymbol{\phi}_t|, r_j = \|\boldsymbol{\phi}_t\|^{-1}|y_t - \widehat{\mathbf{w}}_{j,t-1}^{\mathsf{T}}\boldsymbol{\phi}_t|$, $\boldsymbol{\epsilon}_{j,t-1} = \mathbf{w}_j - \widehat{\mathbf{w}}_{j,t-1}$ and $y_t = \mathbf{w}_i^{\mathsf{T}}\boldsymbol{\phi}_t + n_t$, we have

$$r_i = \|\boldsymbol{\phi}_t\|^{-1}|n_t + \boldsymbol{\epsilon}_{i,t-1}^{\mathsf{T}}\boldsymbol{\phi}_t| \tag{93}$$

$$r_j = \|\boldsymbol{\phi}_t\|^{-1}|n_t + (\mathbf{w}_i - \mathbf{w}_j)^{\mathsf{T}}\boldsymbol{\phi}_t + \boldsymbol{\epsilon}_{j,t-1}^{\mathsf{T}}\boldsymbol{\phi}_t| \tag{94}$$

So, (92) is equivalent to

$$|n_t + (\mathbf{w}_i - \mathbf{w}_j)^{\mathsf{T}}\boldsymbol{\phi}_t + \boldsymbol{\epsilon}_{j,t-1}^{\mathsf{T}}\boldsymbol{\phi}_t| > |n_t + \boldsymbol{\epsilon}_{i,t-1}^{\mathsf{T}}\boldsymbol{\phi}_t|\left(\frac{|n_t|}{\|\boldsymbol{\phi}_t\|(\|\boldsymbol{\epsilon}_{i,t-1}\| + \nu)} + 1\right) \tag{95}$$

Note that in (95), we can see

$$LHS \geq |(\mathbf{w}_i - \mathbf{w}_j)^{\mathsf{T}}\boldsymbol{\phi}_t| - |n_t| - |\boldsymbol{\epsilon}_{j,t-1}^{\mathsf{T}}\boldsymbol{\phi}_t| \tag{96}$$

$$\begin{aligned}
RHS &\leq |n_t| + |\boldsymbol{\epsilon}_{i,t-1}^{\mathsf{T}}\boldsymbol{\phi}_t| + \frac{(|n_t| + \|\boldsymbol{\epsilon}_{i,t-1}\|\|\boldsymbol{\phi}_t\|)|n_t|}{\|\boldsymbol{\phi}_t\|(\|\boldsymbol{\epsilon}_{i,t-1}\| + \nu)} \\
&= |n_t| + |\boldsymbol{\epsilon}_{i,t-1}^{\mathsf{T}}\boldsymbol{\phi}_t| + \frac{\|\boldsymbol{\epsilon}_{i,t-1}\||n_t|}{\|\boldsymbol{\epsilon}_{i,t-1}\| + \nu} + \frac{|n_t|^2}{\|\boldsymbol{\phi}_t\|(\|\boldsymbol{\epsilon}_{i,t-1}\| + \nu)} \\
&< |n_t| + |\boldsymbol{\epsilon}_{i,t-1}^{\mathsf{T}}\boldsymbol{\phi}_t| + |n_t| + \frac{|n_t|^2}{\|\boldsymbol{\phi}_t\|\nu} \\
&= 2|n_t| + |\boldsymbol{\epsilon}_{i,t-1}^{\mathsf{T}}\boldsymbol{\phi}_t| + \frac{|n_t|^2}{\|\boldsymbol{\phi}_t\|\nu}
\end{aligned} \tag{97}$$

Considering (96) and (97), we can see to ensure (95) holds, it suffices to let

$$|\boldsymbol{\epsilon}_{i,t-1}^{\mathsf{T}}\boldsymbol{\phi}_t| + |\boldsymbol{\epsilon}_{j,t-1}^{\mathsf{T}}\boldsymbol{\phi}_t| + 3|n_t| - |(\mathbf{w}_i - \mathbf{w}_j)^{\mathsf{T}}\boldsymbol{\phi}_t| + \frac{|n_t|^2}{\|\boldsymbol{\phi}_t\|\nu} \leq 0 \tag{98}$$

From Assumption 2, 3, 5, we have $\|\boldsymbol{\phi}_t\| \leq \phi_{\max}, |n_t| \leq n_{\max}, |(\mathbf{w}_i - \mathbf{w}_j)^{\mathsf{T}}\boldsymbol{\phi}_t| \geq \psi, \frac{|n_t|}{\|\boldsymbol{\phi}_t\|} \leq \frac{1}{S_{\min}}$. Applying these bounds to (98), we can see to ensure (95) holds, it suffices to let

$$\left(\|\boldsymbol{\epsilon}_{i,t-1}\| + \|\boldsymbol{\epsilon}_{j,t-1}\|\right)\phi_{\max} + 3n_{\max} - \psi + \frac{n_{\max}}{S_{\min}\nu} \leq 0 \tag{99}$$

So, to ensure (99) holds, it suffices to have $\forall i \in [m]$

$$\|\boldsymbol{\epsilon}_{i,t-1}\| \leq \frac{1}{2\phi_{\max}}\left(\psi - \frac{n_{\max}}{\nu S_{\min}} - 3n_{\max}\right) = \epsilon' \tag{100}$$

Tracing all the way back, we can see when (100) holds for $\forall i \in [m]$, (88) would hold, therefore we could assign data $\{\boldsymbol{\phi}_t, y_t\}$ generated by subsystem $i$ to candidate $i$, i.e. $\widehat{z}_t = z_t$. □

### D.4 Proof for Theorem 17

**Proof** For ease of explanation, we let **Correct Assignment Always (CAA)** be the event of correct assignment at every time step, which is exactly result (i). Note that in the claims of this theorem, result (ii) is a direct consequence of result (i) according to Theorem 12. To prove this theorem, it suffices to prove **CAA** happens with probability at least $1 - 2m\sqrt{\frac{N_R}{\epsilon'^2}\left(\epsilon_0^2 + \frac{N_R}{S_{\min}^2}\right)}$ It's difficult to evaluate **CAA** directly, so we will evaluate **CAA** only on the **perfect event trajectory (PET)**: "at every time step (including 0), all candidates have accurate enough estimate after updates such that we can make correct assignment at next time step according to Lemma 16; at every time step, we can make correct assignment". Since **CAA** occurs whenever **PET** occurs, a lower bound on $P(\mathbf{PET})$ would also be a lower bound on $P(\mathbf{CAA})$. We will show that to evaluate $P(\mathbf{PET})$, it suffices to study each candidate separately and then combine them altogether. We will illustrate this with a toy example and then generalize it to general cases.

Table 2: Perfect Event Trajectory (**PET**)

| Time Indices | | Correct Assign | | Accurate Enough Estimation After Update |
|---|---|---|---|---|
| $(t=0)$: | | NA | | $\|\boldsymbol{\epsilon}_{1,0}\|^2, \|\boldsymbol{\epsilon}_{2,0}\|^2 \leq \epsilon'^2$ |
| $(t=1)$: | $\xrightarrow{wp1}$ | $\widehat{z}_1 = 1$ | $\longrightarrow$ | $\|\boldsymbol{\epsilon}_{1,1}\|^2, \|\boldsymbol{\epsilon}_{2,1}\|^2 \leq \epsilon'^2$ after update of $\widehat{\mathbf{w}}_{1,0}$ |
| | | | $\vdots$ | |
| $(t=t_1-1)$: | $\xrightarrow{wp1}$ | $\widehat{z}_{t_1-1} = 1$ | $\longrightarrow$ | $\|\boldsymbol{\epsilon}_{1,t_1-1}\|^2, \|\boldsymbol{\epsilon}_{2,t_1-1}\|^2 \leq \epsilon'^2$ after update of $\widehat{\mathbf{w}}_{1,t_1-2}$ |
| $(t=t_1)$: | $\xrightarrow{wp1}$ | $\widehat{z}_{t_1} = 2$ | $\longrightarrow$ | $\|\boldsymbol{\epsilon}_{1,t_1}\|^2, \|\boldsymbol{\epsilon}_{2,t_1}\|^2 \leq \epsilon'^2$ after update of $\widehat{\mathbf{w}}_{1,t_1-1}$ |
| | | | $\vdots$ | |
| $(t=t_2-1)$: | $\xrightarrow{wp1}$ | $\widehat{z}_{t_2-1} = 2$ | $\longrightarrow$ | $\|\boldsymbol{\epsilon}_{1,t_2-1}\|^2, \|\boldsymbol{\epsilon}_{2,t_2-1}\|^2 \leq \epsilon'^2$ after update of $\widehat{\mathbf{w}}_{1,t_2-2}$ |
| $(t=t_2)$: | $\xrightarrow{wp1}$ | $\widehat{z}_{t_2} = 1$ | $\longrightarrow$ | $\|\boldsymbol{\epsilon}_{1,t_2}\|^2, \|\boldsymbol{\epsilon}_{2,t_2}\|^2 \leq \epsilon'^2$ after update of $\widehat{\mathbf{w}}_{1,t_2-1}$ |
| | | | $\vdots$ | |

Assume there are only two subsystems 1 and 2 in the hybrid SARX system. Subsystem 1 dominates at time $\{1, 2, \ldots, t_1 - 1, t_2, t_2 + 1, \ldots\}$ and subsystem 2 dominates at $\{t_1, t_1 + 1, \ldots, t_2 - 1\}$. This is to say, there is a switching from 1 to 2 at time $t_1$, and 2 back to 1 at time $t_2$. Now, consider the **PET** in Table 2. In this table, time indices are listed on the left of the vertical separator, and the events occur at different time steps are listed on the right. The "*Correct Assign*" column lists the events of making correct assignment at different time steps. The "*Accurate Enough Estimation After Update*" column lists the events of accurate enough (below $\epsilon'$) estimation after update. "$\xrightarrow{wp1}$" means event $\{\|\boldsymbol{\epsilon}_{1,t-1}\|^2, \|\boldsymbol{\epsilon}_{2,t-1}\|^2 \leq \epsilon'^2$ after update of $\widehat{\mathbf{w}}_{i,t-2}$ for some i$\}$ at time $t-1$ will lead to event $\{\widehat{z}_t = z_t\}$, i.e. making correct assignment at time $t$, with probability 1, whose justification is given in Lemma 16. "$\longrightarrow$" means with certain probability, current correct assignment will make estimates accurate enough after update. From Table 2, we can see the randomness in **PET** only come from all the events in the "*Accurate Enough Estimation After Update*" column. In another way, to evaluate $P(\mathbf{PET})$, it's

equivalent to evaluate the probability that for every time step, after updating estimate with data from correct subsystem, the new estimation error will be smaller than $\epsilon'^2$. To see things more clearly, the events we want to evaluate have the following properties

1. At time $t = 1, 2, \ldots, t_1 - 1$, only $\|\boldsymbol{\epsilon}_{1,t}\|^2$ is changing while $\|\boldsymbol{\epsilon}_{2,t}\|^2 = \|\boldsymbol{\epsilon}_{2,0}\|^2$ is unchanged. And we always have $\|\boldsymbol{\epsilon}_{1,t}\|^2, \|\boldsymbol{\epsilon}_{2,t}\|^2 \leq \epsilon'^2$
2. At time $t = t_1, t_1 + 1, t_2 - 1$, only $\|\boldsymbol{\epsilon}_{2,t}\|^2$ is changing while $\|\boldsymbol{\epsilon}_{1,t}\|^2 = \|\boldsymbol{\epsilon}_{1,t_1-1}\|^2$ is unchanged. And we always have $\|\boldsymbol{\epsilon}_{1,t}\|^2, \|\boldsymbol{\epsilon}_{2,t}\|^2 \leq \epsilon'^2$
3. At time $t = t_2, t_2 + 1, \ldots,$ only $\|\boldsymbol{\epsilon}_{1,t}\|^2$ is changing while $\|\boldsymbol{\epsilon}_{2,t}\|^2 = \|\boldsymbol{\epsilon}_{2,t_2-1}\|^2$ is unchanged. And we always have $\|\boldsymbol{\epsilon}_{1,t}\|^2, \|\boldsymbol{\epsilon}_{2,t}\|^2 \leq \epsilon'^2$
4. Additionally, we have $\|\boldsymbol{\epsilon}_{1,0}\|^2, \|\boldsymbol{\epsilon}_{2,0}\|^2 \leq \epsilon'^2$

Now consider the following fictitious Scenario (A): with $\|\boldsymbol{\epsilon}_{1,0}\|^2, \|\boldsymbol{\epsilon}_{2,0}\|^2 \leq \epsilon'^2$, let $\boldsymbol{\Phi}_1$ and $\boldsymbol{\Phi}_2$ denote the data we assigned to candidate 1 $\widehat{\mathbf{w}}_{1,t}$ and candidate 2 $\widehat{\mathbf{w}}_{2,t}$ respectively; then we first update $\widehat{\mathbf{w}}_{1,0}$ using $\boldsymbol{\Phi}_1$ and then update $\widehat{\mathbf{w}}_{2,0}$ with $\boldsymbol{\Phi}_2$ as if there is always only one subsystem during this course. With the properties listed above, we can see $P(\mathbf{PET}) = P(C_1 \bigcap C_2)$ where $C_i$ is the event "candidate i always have error smaller than $\epsilon'^2$ in Scenario (A)". Since $C_1$, $C_2$ corresponds to applying algorithm to single subsystem, we can see $P(C_1)$ and $P(C_2)$ can be lower bounded by the probability in Lemma 15. Therefore, we have

$$P(\mathbf{CAA}) \geq P(\mathbf{PET}) \geq 1 - P(C_1^c) - P(C_2^c) \geq 1 - 2 \cdot 2\sqrt{\frac{N_R}{\epsilon'^2}\left(\epsilon_0^2 + \frac{N_R}{S_{\min}^2}\right)} \tag{101}$$

Now for the more general SARX model with $m$ subsystems, we could generalize the argument above, and end up getting

$$P(\mathbf{CAA}) \geq P(\mathbf{PET}) \geq 1 - m \cdot 2\sqrt{\frac{N_R}{\epsilon'^2}\left(\epsilon_0^2 + \frac{N_R}{S_{\min}^2}\right)} \tag{102}$$

Finally, with the argument we made at the beginning of the proof, we can see the proof for this theorem is done. □

### D.5 Proof for Corollary 18

**Proof** When there is no noise, Assumption 13 which is the building block to the local convergence result Theorem 17 is no longer valid as $S_{\min}$ and $S_{\max}$ will both go to $\infty$ and $\kappa_{\max} \geq \frac{S_{\max}}{S_{\min}}$ is no longer well defined. However, in this case, we can prove variants of Lemma 14 and Lemma 15 without relying on Assumption 13.

Specifically, in the proof of Lemma 14, if $n_t = 0$, the last term $\frac{1}{S_{\min}^2}$ in (78) would vanish and we proved the supermartingale directly and thus Lemma 14 holds without relying on Assumption 13. And in the proof of Lemma 15, all the term $\frac{N_R}{S_{\min}^2}$ would vanish due to the absence of noise. So the claim of Lemma 15 would be assume $\|\boldsymbol{\epsilon}_{i,0}\| \leq \epsilon_0$ such that $\sqrt{N_R\epsilon_0^2} \leq \epsilon'$, then for $\forall t$ we have

$$P\left(\bigcap_{\tau=1}^{t}\left\{\|\boldsymbol{\epsilon}_{i,\tau}\|^2 \leq \epsilon'^2\right\}\right) \geq 1 - 2\sqrt{\frac{N_R}{\epsilon'^2}\epsilon_0^2} \tag{103}$$

We could apply this variant of Lemma 15 to proof for Theorem 17 directly and get the probability bound $1 - 2m\sqrt{\frac{N_R}{\epsilon'^2}\epsilon_0^2}$. Finally we can get (17) in the corollary simply by letting $\sigma_n = 0$ in the partial convergence result Theorem 12. □

# E Proofs for Extension Results Theorem 19

## E.1 Proof for Theorem 19

**Proof** From the (58), we see $\boldsymbol{\epsilon}_t = \mathbf{b} - \mathbf{A}\mathbf{n}_t$ for some $\mathbf{A}$ and $\mathbf{b}$. For some $\tilde{\epsilon} > 0$, we have

$$\begin{aligned} P(\|\boldsymbol{\epsilon}_t\| \leq \tilde{\epsilon}) &= P(\mathbb{1}\{\|\mathbf{A}\mathbf{n}_t - \mathbf{b}\| \leq \tilde{\epsilon}\}) \\ &= \mathbb{E}[\mathbb{1}\{\|\mathbf{A}\mathbf{n}_t - \mathbf{b}\| \leq \tilde{\epsilon}\}] \end{aligned} \tag{104}$$

With Monte Carlo samples of $\mathbf{n}_t, \{\mathbf{n}_t^{(i)}\}_{i=1}^{N_t}$, according to Hoeffding's inequality, we have

$$\begin{aligned} &P\left(\frac{1}{N_t}\sum_{i=1}^{N}\mathbb{1}\{\|\mathbf{A}\mathbf{n}_t^{(i)} - \mathbf{b}\| \leq \tilde{\epsilon}\} - P(\|\boldsymbol{\epsilon}_t\| \leq \tilde{\epsilon}) \leq \zeta_1^t\right) \\ &\geq 1 - \exp(-2N_t\zeta_1^{2t}) \end{aligned} \tag{105}$$

Let $\tilde{\epsilon} = \epsilon_t^u = \max \|A\mathbf{n}_t^{(i)} - \mathbf{b}\|$, and note that $N_t \geq \frac{\zeta_2 t}{2\zeta_1^{2t}}$, we have

$$P\left(P(\|\boldsymbol{\epsilon}_t\| \leq \tilde{\epsilon}) \geq 1 - \zeta_1^t\right) \geq 1 - \exp(-\zeta_2 t) \tag{106}$$

Using union bound, we have

$$P\left(\bigcap_{t=1}^{\infty}\left\{P(\|\boldsymbol{\epsilon}_t\| \leq \tilde{\epsilon}) \geq 1 - \zeta_1^t\right\}\right) \geq 1 - \frac{\exp(-\zeta_2)}{1 - \exp(-\zeta_2)} \tag{107}$$

For the event inside $P(\cdot)$, according to union bound, we have

$$\bigcap_{t=1}^{\infty}\left\{P(\|\boldsymbol{\epsilon}_t\| \leq \tilde{\epsilon}) \geq 1 - \zeta_1^t\right\} \Rightarrow P\left(\bigcap_{t=1}^{\infty}\{\|\boldsymbol{\epsilon}_t\| \leq \tilde{\epsilon}\} \geq 1 - \frac{\zeta_1}{1 - \zeta_1}\right) \tag{108}$$

Therefore, plugging (108) into (107), we could get (28) □