Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online CONFERENCE PAPER ARCHIVE

IFAC-PapersOnLine 48-28 (2015) 680–685

Identification of Switched Linear Systems via Sparse Optimization ⋆

Weisen Jiang * **Hai-Tao Fang** *

* *Key Laboratory of Systems and Control, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, P. R. China(E-mail: jiangweisen12@mails.ucas.ac.cn, htfang@iss.ac.cn)*

Abstract: This paper addresses identification problem of switched linear(SL) systems from input-output data. The main challenge is the partitions of data points correspond to different subsystems are unavailable. Inspired by compressed sensing theory, we pursue the sparsity of estimation error and propose ℓ_0-norm optimization algorithm to identify parameters. Unfortunately, the computational complexity of this approach is intractable. To overcome this difficulty, we replace ℓ_0-norm by ℓ_1-norm, which retains sparse property. We not only provide recoverable conditions for identifying SL systems via ℓ_1-norm minimization program, but also show that ℓ_1-norm estimator is robust to bounded noise. Numerical experiments are included to demonstrate the performance of our algorithms.

Keywords: Sparse optimization, ℓ_1-norm minimization, ℓ_0-norm minimization, system identification, switched linear systems

1. INTRODUCTION

Switched linear(SL) systems are consisted of several linear subsystems and a switching law, the former arises from physical principles and the latter is governed by logical devices. These systems are important in practice because numerous biological and engineering systems are too complex to be described simply by a signal linear system, while many systems switch between several linear subsystems depending on different environments. In addition, SL systems are used to approximate nonlinear phenomenon, e.g. any nonlinear continuous function can be approximated by piecewise affine(PWA) functions with arbitrary accuracy. During the last two decades, SL systems have attracted increasing attention in control community and many theoretical results were obtained from various viewpoints, including control design(Ge and Sun (2005)), observability/stability analysis(Sun and Ge (2011); Vidal et al. (2003a); Gomez-Gutierrez et al. (2010)) and verification(Bemporad and Morari (1999)). However, most of these developments hinge on prior knowledge of system models or switching law, which are unavailable in many practical applications. In such situations, we need to identify system parameters in advanced from input-output data and some *a prior* structure information.

Identification of SL systems is a challenging problem since both system parameters and switching law are unknown. For the piecewise affine(PWA) systems, i.e. the regressors space is partitioned into polyhedra with affine system for each polyhedron, the switching law is determined by the regressor and continuous in the interior of polyhedrons. Based on this particular switching law, numerous identification algorithms have been proposed, e.g. clustering-based procedure(Ferrari-trecate et al. (2003)), mixed-integer programing(Roll et al. (2004)), bounded-error approach(Bemporad et al. (2005)) and recursive weighted least squares algorithm(Zhao and Zhou (2012)). When the switching law is arbitrary, most of existing identification algorithms are based on two-step procedures: classify the experimented data into several groups according to different subsystems, then identify each subsystem separately. To deal with identification and classification problem together, an algebraic geometric approach is proposed in Vidal et al. (2003b). In the noise-free case, under some conditions, this approach identifies system order, subsystem number, model parameters and classifies data exactly. Recently, inspired by the developments of compressed sensing community, Bako (2011) presented sparse optimization to identify SL systems and analyzed recoverable conditions without concerning noise .

In this paper, we study the performance of identification of SL systems via ℓ_0-norm and ℓ_1-norm minimization approaches. The intuition of ℓ_0-norm estimator is following: noise-free data points are lying on several hyperplanes and identification of the subsystem that contains the largest number of data points is equivalent to finding a parameter vector such that estimation error is sparsest. Unfortunately, the complexity of pursing sparsity under ℓ_0-norm criterion is intractable and Non-deterministic Polynomial-time (NP) hard. Inspired by recent developments in compressed sensing community(Candès and Tao (2005); Candès et al. (2006); Candès and Tao (2007)), we replace ℓ_0-norm by ℓ_1-norm, which promotes sparsity and remains computationally tractable. In Bako (2011), recov-

⋆ This work was supported by the National Key Basic Research Program of China (973 program) under Grant no. 2014CB845301/2/3 and the National Natural Science Foundation of China under Grant 61174143.

10.1016/j.ifacol.2015.12.208

erable conditions for ℓ_0-norm estimator are established on orthogonal projection matrix, which is difficult to verify, especially when the data set is large. Instead, our sufficient conditions for recovering via ℓ_0-norm and ℓ_1-norm estimators are built on data matrix directly, and appliable to some kinds of SL systems. When the measurements are contaminated with noise, unlike algebraic geometric approach(Vidal et al. (2003b)) that is sensitive to noise, ℓ_1-norm estimator is robust and satisfies an oracle inequality. Our main contributions are listed here.

- In the noise-free setting, sufficient conditions for recovering SL system parameters via ℓ_0-norm and ℓ_1-norm estimators are derived.
- Under some conditions, we prove that ℓ_1-norm estimator is robust to bounded noise.
- In the bounded noise case, ℓ_1-norm estimator achieves the performance of oracle estimator in some sense.
- Theoritical results are verified via simulations.

The organization of this paper is as follows. In Section 2, we formulate the identification problem of SL systems and give two illustrated examples. The ℓ_0-norm estimators are designed in Section 3, and identifiable conditions are established as well. Since ℓ_0-norm optimization program is intractable, we discuss ℓ_1-norm minimization program to estimate system parameters in Section 4. Both recoverable conditions and robustness are derived in this section. Several numerical experiments are presented in Section 5 to verify our theoretical developments, and Section 6 draws the conclusions.

Notation: A vector $x \in \mathbb{R}^n$ is viewed as a column vector, and x_i denotes the i-th component of x with an exception: system parameter vector $\theta_i \in \mathbb{R}^n$ and $\theta_{i,j}$ denotes the j-th component of θ_i. We use $\|x\|_2$ and $\|x\|_1$ to denote the standard ℓ_2 and ℓ_1 norms on vectors, i.e. $\|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$ and $\|x\|_1 = \sum_{i=1}^n |x_i|$. The set difference $\mathcal{I}_1 - \mathcal{I}_2$ is defined as $\mathcal{I}_1 - \mathcal{I}_2 \triangleq \{x : x \in \mathcal{I}_1, x \notin \mathcal{I}_2\}$. For a set $\mathcal{T}$, $|\mathcal{T}|$ is the cardinality of $\mathcal{T}$. A quasi-norm $\|x\|_0$ is defined as the number of nonzero components of x, and x is called k-sparse if $\|x\|_0 \leq k$. Given a matrix $A \in \mathbb{R}^{m\times n}$, for any index set $\mathcal{T} \subseteq \{1, 2, \cdots, n\}$, $A_{\mathcal{T}} \in \mathbb{R}^{n\times|\mathcal{T}|}$ denotes the sub-matrix of A consisting of columns of A indexed by $\mathcal{T}$. Similarly, $x_{\mathcal{T}} \in \mathbb{R}^{|\mathcal{T}|}$ denotes the sub-vector of x with components indexed by $\mathcal{T}$. The transpose of vector x and matrix A are denoted by x^T and A^T. Denote $\mathrm{Tr}(X)$ as the trace of matrix X.

2. PROBLEM FORMULATION

Consider SL systems

$$y(t) = \theta_{\sigma(t)}^T \varphi(t) + e(t), \tag{1}$$

where $\sigma(t) \in \mathcal{S} \triangleq \{1, \cdots, s\}$ is a switch variable, or the index of subsystem at time t, $\{\theta_1, \cdots, \theta_s\} \subseteq \mathbb{R}^n$ are system parameters to be identified, $\varphi(t)$ is regressor, $e(t)$ is measurements noise. The identification problem is to estimate $\{\theta_1, \cdots, \theta_s\}$ based on available input-output data $\{\varphi(t), y(t)\}_{t=1}^N$. Let $\mathcal{I} = \{1, 2, \cdots, N\}$ and $\mathcal{I}_i$ be the index set of data points correspond to i-th subsystem, denote $N_i = |\mathcal{I}_i|$. Without loss of generality, assume $N_1 \geq N_2 \geq \cdots \geq N_s$ and $\{\mathcal{I}_i\}_{i=1}^s$ are disjoint partition of $\mathcal{I}$.

When switch variable $\sigma(t)$ is known, we can apply traditional identification algorithms such as recursive least square and frequent domain approach to identify each subsystem separately instead of SL systems. Unfortunately, data partitions are unavailable in general, and this problem becomes rather difficult.

We give two illustrated examples and their identification problems are researched gradually in this paper.

Example 1. (PWA system). Consider a PWA system $f(\cdot) : [-1, 1] \to \mathbb{R}$ as follows:

$$f(x) = \begin{cases} x - 1 & \text{if } -1 \leq x \leq 0 \\ 0.5x + 1 & \text{if } 0 < x \leq 0.5 \end{cases}$$

and measurement

$$y = f(x) + e,$$

where e is the observation noise. Denote parameter vectors and regressor as

$$\theta_1 = [1,\ -1]^T,\ \theta_2 = [0.5,\ 1]^T,\ \varphi = [x,\ 1]^T,$$

then the switching law entirely depends on regressors, that is, $\sigma(t) = 1$ if $-1 \leq \varphi_1 \leq 0$; $\sigma(t) = 2$ if $0 < \varphi_2 \leq 0.5$.

Notice that the switching law of PWA system is a continuous function of regressors in the interior of each polyhedron. This property don't hold any more in general SL systems.

Example 2. (SL systems). Consider SL systems $f(\cdot) : \mathbb{R}^n \to \mathbb{R}$ consisting of three subsystems

$$f(\varphi(t)) = \theta_{\sigma(t)}^T \varphi(t),$$

and measurements

$$y(t) = f(\varphi(t)) + e(t).$$

Switching variables $\{\sigma(t)\}_{t=1}^N$ are independent identical distribution(i.i.d) random variables(r.v.s) with distribution

$$\mathbb{P}\{\sigma(t) = i\} = \begin{cases} \alpha_1 & \text{if } i = 1 \\ \alpha_2 & \text{if } i = 2 \\ \alpha_3 & \text{if } i = 3 \end{cases}$$

where $\{\alpha_i\}_{i=1}^3$ are positive numbers such that $\sum_{i=1}^3 \alpha_i = 1$. In such situation, the outputs may be different for the same regressors at different switching time.

3. ℓ_0-NORM MINIMIZATION APPROACH

In this section, we pursue the sparsity of estimation error and propose an ℓ_0-norm minimization program to identify system parameters. Since ℓ_0-norm is sensititve to noise, we adopt the following noise-free assumption thoughout this section.

Assumption 3. (Noise-free). $e(t) = 0$ for all t.

Define an estimate error vector

$$Error(\theta) = Y - X^T\theta,$$

where data matrix X and data vector Y are denoted as

$$X = [\varphi(1)\ \cdots\ \varphi(N)]$$
$$Y = [y(1)\ \cdots\ y(N)]^T.$$

According to system evolution (1), $Error(\theta_i)$ is $(N - N_i)$-sparse. Based on this observation and $N_1 \geq N_2$, we design

an estimator of θ_1 through solving ℓ_0-norm minimization problem

$$\hat{\theta} = \underset{\theta}{\text{minimize}} \ \|Y - X^T\theta\|_0. \tag{2}$$

To study the recoverability of this estimator, we introduce some basic notions come from compressed sensing.

Definition 4. (Donoho and Elad (2003)). Let X be $n \times N$ matrix with $n \leq N$. $spark(X)$ is the smallest number of columns of X that are linear dependent, $spark(X) = N+1$ if all columns of X are linear independent. X is full-spark if $spark(X) = n + 1$.

It is obviously that full-spark implies full-rank, but not converse. More specifically, $spark(X) \leq rank(X) + 1$. An equivalent statement of full-spark is any n columns of X are linear independent.

Example 5. Consider the $2 \times N$ data matrix X of PWA system introduced in Example 1

$$X = \begin{bmatrix} x(1) & x(2) & \cdots & x(N) \\ 1 & 1 & \cdots & 1 \end{bmatrix}. \tag{3}$$

If $x(i) \neq x(j)$ whenever $i \neq j$, then X is full-spark as well as full-rank.

Example 6. For the data matrix X generated by SL systems in Example 2, if $\{\varphi(t)\}_{t=1}^N$ are sampled independently from $\mathbb{R}^n$, then X is full-spark with probabiltiy one.

Example 7. For a $n \times (N+1)$ matrix $X = [e_1 \ I_N]$, where $e_1 = [1 \ 0 \ \cdots \ 0]^T$ and I_N is $N \times N$ identity matrix. Then $rank(X) = N$ but $spark(X) = 2$ since the first two columns are identical. Hence, $rank(X)$ is possible to be much larger than $spark(X)$.

The following theorem show ℓ_0-norm minimization program (2) recovers θ_1 exactly when N_1 is sufficient large.

Theorem 8. Under Assumption 3 and X is full-spark, if $N_1 \geq \frac{N+n}{2}$, then the solution $\hat{\theta}$ to (2) equals to θ_1.

Proof. Assume $\hat{\theta} \neq \theta_1$, then $\|Y - X^T\hat{\theta}\|_0 \leq \|Y - X^T\theta_1\|_0 \leq N - N_1$ since $\hat{\theta}$ is the optimal solution. Hence, $\|X^T(\theta_1 - \hat{\theta})\|_0 \leq 2(N - N_1)$, which implies that there are at least $2N_1 - N$ rows of X^T are orthogonal to $\theta_1 - \hat{\theta}$. Since $n \leq 2N_1 - N$, there exist at least n rows of X^T, or columns of X are orthogonal to $\theta_1 - \hat{\theta}$, which contradicts to X is full-spark.

When θ_1 is recovered, data points correspond to the 1-th subsystem are identified at the same time. In a sequel, we estimate θ_2 and then $\mathcal{I}_2$ based on remaining data points $\{\varphi(t), y(t)\}_{t \in \cup_{i=2}^s \mathcal{I}_i}$, and so on for $\{\theta_i, \mathcal{I}_i\}_{i=3}^s$. The procedure of ℓ_0-norm estimator is alternated between identification and classification:

$$\theta_1 \to \mathcal{I}_1 \to \theta_2 \to \mathcal{I}_2 \to \cdots \to \theta_s \to \mathcal{I}_s. \tag{4}$$

The ℓ_0-norm estimator is designed as follows.

Algorithm 1 ℓ_0-norm estimator

Input: $\{\varphi(t), y(t)\}_{t=1}^N$
Initialization: $\hat{\mathcal{I}}_0 = \emptyset$
while $1 \leq i \leq s$ **do**
1. Estimate θ_i:

$$\hat{\theta}_i = \arg\min \|Y_{\mathcal{I} - \cup_{j=0}^{i-1}\hat{\mathcal{I}}_j} - X^T_{\mathcal{I} - \cup_{j=0}^{i-1}\hat{\mathcal{I}}_j}\theta\|_0$$

2. Estimate $\mathcal{I}_i$:

$$\hat{\mathcal{I}}_i = \{t : y(t) - \hat{\theta}_i^T\varphi(t) = 0\}$$

end while
Output: $\{\hat{\theta}_i\}_{i=1}^s$

Corollary 9. Under Assumption 3 and X is full-spark. If $\{N_i\}_{i=1}^s$ satisfy chained inequalities:

$$\begin{aligned} N_1 &\geq \frac{N+n}{2} \\ N_2 &\geq \frac{N - N_1 + n}{2} \\ &\vdots \\ N_{s-1} &\geq \frac{N - N_1 - \cdots - N_{s-2} + n}{2} \\ N_s &\geq n, \end{aligned}$$

then $\{\theta_i, \mathcal{I}_i\}_{i=1}^s$ are recovered perfectly via Algorithm 1.

Proof. Suppose $\{N_i\}_{i=1}^s$ satisfy chained inequalities, then Theorem 8 implies that θ_1 and $\mathcal{I}_1$ are recovered via Algorithm 1. Follow a similar technique as Theorem 8, it is easy to verify that θ_2 is the unique solution to ℓ_0-norm minimization problem

$$\text{minimize} \ \|Y_{\mathcal{I} - \mathcal{I}_1} - X^T_{\mathcal{I} - \mathcal{I}_1}\theta\|_0,$$

and then $\mathcal{I}_2 = \{t : y(t) - \varphi(t)^T\theta_2 = 0\}$ is identified. Proceed Algorithm 1, we recover $\{\theta_i, \mathcal{I}_i\}_{i=1}^s$.

The significant challenge is how to design inputs such that X is full-spark, especially when $N \gg n$. This is an NP hard problem for deterministic inputs(Tillmann and Pfetsch (2014)). Fortunately, for random input sequence, full-spark property holds almost surely.

Assumption 10. (Inputs). $\{\varphi(t)\}_{t=1}^N$ are independent identical distribution(i.i.d) Gaussian random variables(r.v.s) with distribution $\mathcal{N}(0, I_n)$.

Theorem 11. Under Assumption 10, X is full-spark with probability one.

Proof. Since any n independent random vectors in $\mathbb{R}^n$ are linear independent with probability one, we have $spark(X) \geq n + 1$. In addition, X is full-rank almost surely, thus, $spark(X) \leq n + 1$, and we conclude that $spark(X) = n + 1$ with probability one.

Remark 12. For some SL systems, e.g. switched autoregressive system with exogenous inputs(SARX) and PWA systems, some components of X are dependent and deterministic, so Assumption 10 is violated. However, random inputs are possible to make X sufficiently disorder such that full-spark condition holds, as shown in next example.

Example 13. Consider the PWA system introduced in Example 1 when inputs $\{x(t)\}_{t=1}^N$ are i.i.d r.v.s with uniform distributions $\mathcal{U}(-1, 0.5)$. Since with probability 1, $x(t_1) \neq x(t_2)$ whenever $t_1 \neq t_2$, it follows from Example

5 that data matrix X is full-spark almost surely. As the number of data points N increasing, the probability that chained inequality holds tends to one.

Example 14. For the SL systems in Example 2, under the Assumption 10, if $\alpha_1 > 0.5$, $\alpha_2 > 0.25$ and $\alpha_3 > 0$, then excluding a zero measure set, $\{N_i\}_{i=1}^3$ satisfy chained inequalities when N is large enough.

4. ℓ_1-NORM MINIMIZATION APPROACHES

In the last section, we introduce ℓ_0-norm estimator and provide recoverable conditions. Unfortunately, sparse optimization problem is non-convex and intractable, we should replace ℓ_0-norm by other norms that retain sparse but tractable. A popular one is ℓ_1-norm and the minimization program is as follows:

$$\underset{\theta}{\text{minimize}} \|Y - X^T\theta\|_1. \tag{5}$$

Similar to spark property in ℓ_0-norm minimization program, we introduce a matrix property called k-balance.

Definition 15. An $m \times n$ matrix A is k-balance if there exists a constant $\varsigma > 0$, for any nonzero vector $z \in \mathbb{R}^n$, for any index set $\mathcal{K} \subseteq \mathcal{I}$ with $|\mathcal{K}| = k$, it holds that

$$\|(Az)_{\mathcal{K}}\|_1 + \varsigma\|z\|_2 < \|(Az)_{\mathcal{K}^c}\|_1. \tag{6}$$

We call A is weak k-balance if $\varsigma = 0$ in (6).

When A is weak k-balance, for any vector z, Az approximately assigns its length $\|Az\|_1$ to each components on average.

Remark 16. An equivalent expression of (6) is

$$\|(Az)_{\mathcal{K}}\|_1 + \varsigma\|z\|_2 < \frac{1}{2}\|Az\|_1. \tag{7}$$

Theorem 17. Under Assumption 3, if X^T is weak $(N - N_1)$-balance, then the solution $\hat{\theta}$ to (5) is equal to θ_1.

Proof. For any $z \in \mathbb{R}^n$,

$$\begin{aligned}\|Y - X^Tz\|_1 =& \|(Y - X^T\theta_1) + X^T(\theta_1 - z)\|_1\\ \geq& \|Y - X^T\theta_1\|_1 - \|(X^T(\theta_1 - z))_{\mathcal{I}_1^c}\|_1\\ &+ \|(X^T(\theta_1 - z))_{\mathcal{I}_1}\|_1\\ \geq& \|Y - X^T\theta_1\|_1 + \varsigma\|\theta_1 - z\|\\ \geq& \|Y - X^T\theta_1\|_1,\end{aligned} \tag{8}$$

where the second inequality is followed from (6) with $\mathcal{K}$ replaced by $\mathcal{I}_1^c$. Since the equality of (8) holds only if $z = \theta_1$, we conclude that θ_1 is the unique solution to (5).

Corollary 18. Under Assumption 3, $\{\theta_1, \cdots, \theta_s\}$ are recovered via ℓ_1-norm minimization programs if $\Phi_i^T \triangleq X^T_{\mathcal{I}-\bigcup_{j=1}^{i-1}\mathcal{I}_j}$ is weak $(N - \sum_{j=1}^{i} N_j)$-balance for all i.

Proof. Follow the procedure as proof of Corollary 9 and we omit it here.

For ℓ_0-norm estimators, all the analysis are based on noise-free assumption since ℓ_0-norm is sensitive to disturbance. However, ℓ_1-norm estimator is robust to bounded noise and also optimal in certain sense.

Theorem 19. Assume $\sup_t |e(t)| \leq \rho$ and X^T is $(N - N_1)$-balance. Let $\hat{\theta}_1$ be the solution to (5), then $\|\theta_1 - \hat{\theta}_1\|_2 \leq c_0\rho$, where c_0 is a constant independent of noise.

Proof. Since $\hat{\theta}_1$ is the optimal solution to (5),

$$\|Y - X^T\hat{\theta}_1\|_1 \leq \|Y - X^T\theta_1\|_1. \tag{9}$$

Let $V \triangleq Y - X^T\theta_1$ and rewrite (9) as

$$\|X^T(\theta_1 - \hat{\theta}_1) + V\|_1 \leq \|V\|_1. \tag{10}$$

Since

$$\begin{aligned}\|V\|_1 =& \|(V)_{\mathcal{I}_1^c}\|_1 + \|(V)_{\mathcal{I}_1}\|_1\\ \leq& \|(Y - X^T\hat{\theta}_1)_{\mathcal{I}_1^c}\|_1\\ &+ \|(X^T(\theta_1 - \hat{\theta}_1))_{\mathcal{I}_1^c}\|_1 + \rho|\mathcal{I}_1|\end{aligned} \tag{11}$$

and

$$\begin{aligned}&\|X^T(\theta_1 - \hat{\theta}_1) + V\|_1\\ &\geq \|(X^T(\theta_1 - \hat{\theta}_1) + V)_{\mathcal{I}_1^c}\|_1\\ &+ \|(X^T(\theta_1 - \hat{\theta}_1))_{\mathcal{I}_1}\|_1 - \rho|\mathcal{I}_1|.\end{aligned} \tag{12}$$

substitute (11) and (12) into (10) implies

$$\begin{aligned}&\|(X^T(\theta_1 - \hat{\theta}_1))_{\mathcal{I}_1}\|_1\\ &\leq \|(X^T(\theta_1 - \hat{\theta}_1))_{\mathcal{I}_1^c}\|_1 + 2\rho|\mathcal{I}_1|\end{aligned}$$

or

$$\frac{1}{2}\|X^T(\theta_1 - \hat{\theta}_1)\|_1 - \|(X^T(\theta_1 - \hat{\theta}_1))_{\mathcal{I}_1^c}\|_1 \leq \rho|\mathcal{I}_1|.$$

Since X^T is $(N - N_1)$-balance and $\mathcal{I}_1^c = N - N_1$, apply (7) to above inequality with z and $|\mathcal{K}|$ replaced by $\theta_1 - \hat{\theta}_1$ and $\mathcal{I}_1$ respectively, we obtain

$$\|\theta_1 - \hat{\theta}_1\|_2 \leq \frac{|\mathcal{I}_1|}{\varsigma}\rho = c_0\rho,$$

where $c_0 = \frac{N_1}{\varsigma}$.

Theorem 19 shows that the square error of ℓ_1-norm estimator is proportional to the noise level ρ^2, it is of interest to see that this performance can't be improved any more in the sense of ignoring the factor c_0. Assume $\{e(t)\}_{t=1}^N$ are i.i.d bounded noise with variance $\tilde{c}_0\rho^2$ and $\sup_t |e(t)|_\infty \leq \rho$, where $\tilde{c}_0$ is a constant. Suppose the eigenvalues of $X_{\mathcal{I}_1}X_{\mathcal{I}_1}^T$ lies in the interval $[\sqrt{N_1} - \sqrt{n} - \delta, \sqrt{N_1} + \sqrt{n} + \delta]$, where $\delta \in [0, \sqrt{N_1} - \sqrt{n})$ is independent of $\mathcal{K}$ but depends on N_1. We remark that under the Assumption 10, this property holds with probability at least $1 - e^{-\delta^2/2}$, see Davidson and Szarek (2001) for details. Provided we had available an oracle letting us know the index sets $\{\mathcal{I}_j\}_{j=1}^s$ in advance. Then oracle estimator (linear least square estimator) of the 1-st subsystem is given by

$$\hat{\theta}_1^{\text{oracle}} = (X_{\mathcal{I}_1}X_{\mathcal{I}_1}^T)^{-1}X_{\mathcal{I}_1}Y_{\mathcal{I}_1}$$

and its mean square error is

$$\begin{aligned}\mathbb{E}\,\|\theta_1 - \hat{\theta}_1^{\text{oracle}}\|_2^2 &= \mathbb{E}\,\|(X_{\mathcal{I}_1}X_{\mathcal{I}_1}^T)^{-1}X_{\mathcal{I}_1}E_{\mathcal{I}_1}\|_2^2\\ &= \tilde{c}_0\rho^2\,\text{Tr}((X_{\mathcal{I}_1}X_{\mathcal{I}_1}^T)^{-1}).\end{aligned}$$

It follows from $\text{Tr}((X_{\mathcal{I}_1}X_{\mathcal{I}_1}^T)^{-1}) \geq \frac{n}{\sqrt{N_1}+\sqrt{n}+\delta_p}$ that

$$\mathbb{E}\,\|\theta_1 - \hat{\theta}_1^{\text{oracle}}\|_2^2 \geq \frac{n\tilde{c}_0}{\sqrt{N_1} + \sqrt{n} + \delta_p}\rho^2.$$

Hence, compared with Theorem 19, under balance conditions, ℓ_1-norm estimator achieves the performance as oracle does except a factor, which is viewed as the price we pay for loss of knowledge of switch variable.

After θ_1 is identified, we estimate $\mathcal{I}_1$ using threshold method

$$\hat{\mathcal{I}}_1 = \{t : |y(t) - \hat{\theta}_1^T\varphi(t)| \leq \eta\},$$

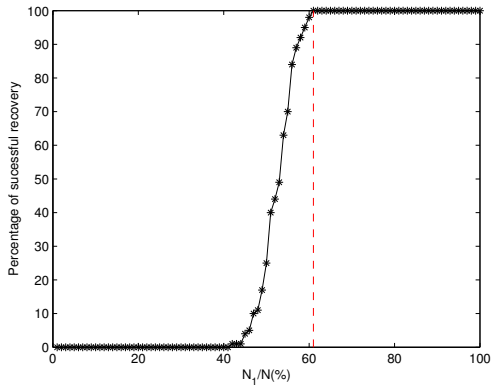

Fig. 1. Percentage of successful recovery for different N_1/N.

where $\eta > 0$ can be determinied by trajectory analysis(see Section 5). Similar to the noise-free case, estimator of θ_2 is given by

$$\hat{\theta}_2 = \arg\min \|Y_{\mathcal{I}-\hat{\mathcal{I}}_1} - X^T_{\mathcal{I}-\hat{\mathcal{I}}_1}\theta\|_1.$$

To sum up the ℓ_1-norm estimate procedure, we give an algorithm as follows.

Algorithm 2 ℓ_1-norm estimator

Input: $\{\varphi(t), y(t)\}_{t=1}^N$
Initialization: Threshold $\eta > 0$, $\hat{\mathcal{I}}_0 = \emptyset$
while $1 \le i \le s$ **do**
1. Estimate θ_i:

$$\hat{\theta}_i = \arg\min \|Y_{\mathcal{I}-\cup_{j=0}^{i-1}\hat{\mathcal{I}}_j} - X^T_{\mathcal{I}-\cup_{j=0}^{i-1}\hat{\mathcal{I}}_j}\theta\|_1$$

2. Select threshold η
3. Estimate $\mathcal{I}_i$:

$$\hat{\mathcal{I}}_i = \{t : |y(t) - \hat{\theta}_i^T\varphi(t)| \le \eta\}$$

end while
Output: $\{\hat{\theta}_i\}_{i=1}^s$

The remaining difficulty is how to design inputs such that X^T is k-balanced. Similar to *spark*, for deterministic inputs, this is a combinational problem. Fortunately, when the input signals are i.i.d Gaussian r.v.s, X^T is k-balanced with high probability.

Theorem 20. (Dwork et al. (2007)). Under Assumption 10, suppose $k = \beta N$ and β sufficient small, then X^T is k-balance with overwhelming probability $1 - Ce^{-cN}$, where c and C are positive constants independent of N.

5. NUMERICAL EXPERIMENTS

In this section, we design experiments to test the performance of ℓ_1-norm estimator. Consider SL systems

$$y(t) = \theta^T_{\sigma(t)}\varphi(t) + e(t),$$

where $\sigma(t) \in \{1, 2\}$ and

$$\theta_1 = [-1 \ \ 2 \ \ -1.3 \ \ 1.5]^T,$$
$$\theta_2 = [2.5 \ \ -0.7 \ \ -2 \ \ 1.2]^T.$$

We generate the data set under the following conditions:

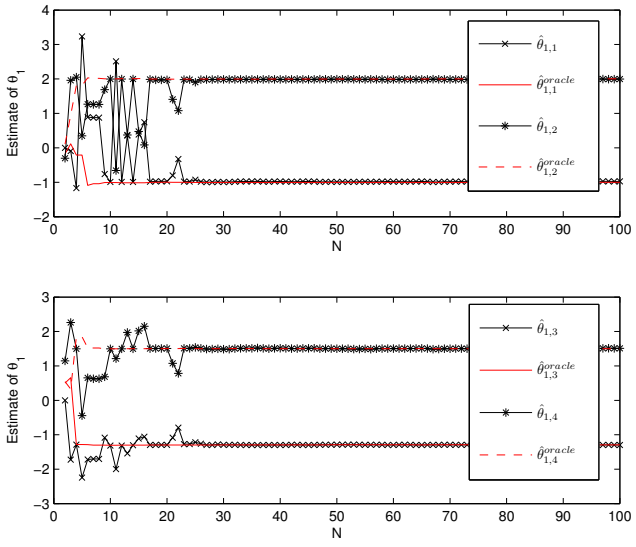

Fig. 2. Estimates of θ_1 via ℓ_1-norm and oracle estimators when measurements are contaminated with noise.

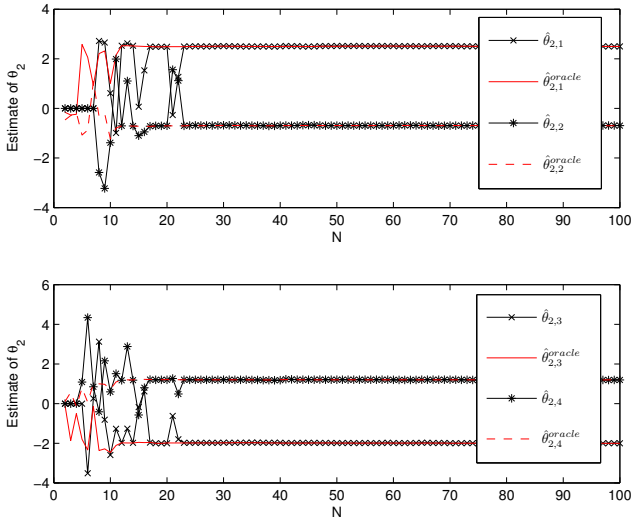

Fig. 3. Estimates of θ_2 via ℓ_1-norm and oracle estimators when measurements are contaminated with noise.

- The inputs $\{\varphi(t)\}_{t=1}^N$ are i.i.d r.v.s with standard normal distributions $\mathcal{N}(0, I_4)$.
- The noise $\{e(t)\}_{t=1}^N$ are i.i.d r.v.s. with uniform distribution $\mathcal{U}(-\rho, \rho)$. In addition, $e(t)$ and $\varphi(t)$ are independent.
- For the switch variable, we suppose

$$\sigma(t) = \begin{cases} 1 & \text{if } 1 \le t \le N_1 \\ 2 & \text{if } N_1 + 1 \le t \le N \end{cases}$$

In the first part, we verify the recoverability of ℓ_1-norm estimator in the noise-free case, i.e. $\rho = 0$. For a fixed value N_1, we solve ℓ_1-norm minimization program (5) for 100 times on independent simulations with $N = 1000$, and compute the percentage of recovering θ_1 successfully. In

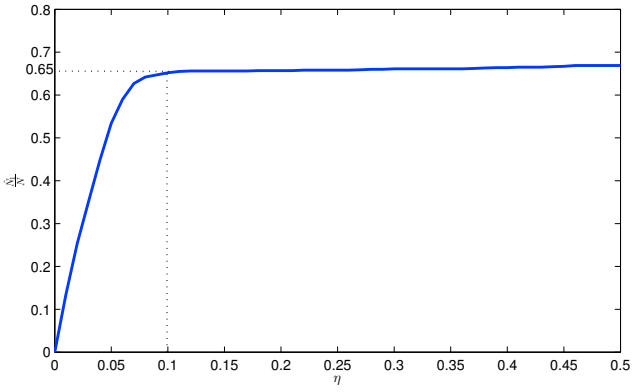

Fig. 4. When $\eta < 0.1$, the ratio $\frac{\hat{N}_1}{N}$ increases as η increases, but remains around 0.65 when $\eta > 0.1$. Hence, we select the threholds $\eta = 0.1$.

Fig. 1, ℓ_1-norm estimator recovers θ_1 successfully when $\frac{N_1}{N} > 61\%$, which is consistent with Theorem 17 and 20.

In the second part, we test the robustness of Algorithm 2. Assume noise level $\rho = 0.05$ and $N_1 = 0.65N$, run Algorithm 2 as N increasing. After Step 1, we obtain the estimation of first subsystem, which is shown in Fig. 2 together with oracle estimation. Then, follow Algorithm 2, we tracking the ratio $\frac{|\hat{\mathcal{I}}_1|}{\mathcal{I}}$(i.e. $\frac{\hat{N}_1}{N}$) as η increases when $N = 50$. As shown in Fig. 4, when $\eta < 0.1$, the value of $\frac{\hat{N}_1}{N}$ increases as η increases. However, when $\eta > 0.1$, $\frac{\hat{N}_1}{N}$ stops increasing and remains around 0.65. Hence, we choose threshold $\eta = 0.1$ to identify the data points belong to first subsystem. The estimates of θ_2 via ℓ_1-norm and oracle estimators are shown in Fig. 3, where our estimator achieves oracle level when $N > 25$. This remarkable performance is in accordance with Theorem 19 and discussions below it.

6. CONCLUSIONS

In this paper, we discusses the identification problem of SL systems from input-output datum. We transform this problem into pursuing the sparsity of estimate error, which is actually an ℓ_0-norm optimization. Since solving ℓ_0-norm minimization problem is intractable, we relax it and propose ℓ_1-norm minimization program. When the measurements are noise-free, sufficient conditions for recovering SL systems via ℓ_0-norm and ℓ_1-norm estimators are provided respectively. We also show that ℓ_1-norm estimator is robust to bounded noise and meets oracle inequality. Experiments are included to verify theoretical results and demonstrate the performance of ℓ_1-norm estimator. For further research, it is of interest to verify spark and k-balance conditions for general input sequences.

REFERENCES

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

Bemporad, A., Garulli, A., Paoletti, S., and Vicino, A. (2005). A bounded-error approach to piecewise affine system identification. *IEEE Trans. Autom. Control*, 50(10), 1567–1580.

Bemporad, A. and Morari, M. (1999). Verification of hybrid systems via mathematical programming. In *Hybrid Systems: Computation and Control, volume 1569 of Lecture Notes in Computer Science*, 31–45. Springer Verlag.

Candès, E.J., Romberg, J.K., and Tao, T. (2006). Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59(8), 1207–1223.

Candès, E.J. and Tao, T. (2005). Decoding by linear programming. *IEEE Trans. Inform. Theory*, 51(12), 4203–4215.

Candès, E.J. and Tao, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *Ann. Stat.*, 35(6), pp. 2313–2351.

Davidson, K.R. and Szarek, S.J. (2001). *Local operator theory, random matrices and Banach spaces*. Elsevier Science.

Donoho, D.L. and Elad, M. (2003). Optimally sparse representation in general (nonorthogonal) dictionaries via l_1 minimization. *Proc. Natl. Acad. Sci.*, 100(5), 2197–2202.

Dwork, C., McSherry, F., and Talwar, K. (2007). The price of privacy and the limits of lp decoding. In *Proc. 39th Annu. ACM Symp. Theory Computing (STOC)*, 85–94.

Ferrari-trecate, G., Liberati, M.M.D., Muselli, M., Liberati, D., and Morari, M. (2003). A clustering technique for the identification of piecewise affine systems. *Automatica*, 39, 205–217.

Ge, S.S. and Sun, Z. (2005). *Switched Linear Systems: Control and Design*. Communications and Control Engineering. Springer.

Gomez-Gutierrez, D., Ramirez-Prado, G., Ramirez-Trevino, A., and Ruiz-Leon, J. (2010). Observability of switched linear systems. *IEEE Trans. Inform. Theory*, 6(2), 127–135.

Roll, J., Bemporad, A., and Ljung, L. (2004). Identification of piecewise affine systems via mixed-integer programming. *Automatica*, 40(1), 37 – 50.

Sun, Z. and Ge, S.S. (2011). *Stability Theory of Switched Dynamical Systems*. Communications and Control Engineering. Springer.

Tillmann, A.M. and Pfetsch, M.E. (2014). The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing. *IEEE Trans. Inform. Theory*, 60(2), 1248–1259.

Vidal, R., Chiuso, A., Chiuso, R., Soatto, S., and Sastry, S. (2003a). Observability of linear hybrid systems. In *In Hybrid Systems: Computation and Control, LNCS*, 526–539. Springer Verlag.

Vidal, R., Soatto, S., Ma, Y., and Sastry, S. (2003b). An algebraic geometric approach to the identification of a class of linear hybrid systems. In *Proc. 42nd IEEE Conf. Dec. Control*, volume 1, 167–172 Vol.1.

Zhao, W. and Zhou, T. (2012). Weighted least squares based recursive parametric identification for the submodels of a pwarx system. *Automatica*, 48(6), 1190 – 1196.