Proceedings of the 33rd Chinese Control Conference
July 28-30, 2014, Nanjing, China

Identification for Wiener System with Discontinuous Piece-wise Linear Function 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 presents a new approach to the identification of Wiener system consisted of an ARX subsystem followed by a static discontinuous piece-wise linear subsystem. We show this problem can be transformed into an ℓ_0-norm optimization problem, which is intractable(NP hard). To overcome this difficulty, we consider ℓ_1-norm convex relaxation inspired by compressed sensing. In the noise-free case , sufficient conditions are provided for recovering unknown parameters via ℓ_0-norm and ℓ_1-norm minimization programs. Numerical experiments demonstrate our novel algorithms perform well in noisy measurements case.

Key Words: System identification, Wiener system, sparse optimization, ARX model, compressed sensing

1 Introduction

Many practical systems can be modeled by the system composed of a linear subsystem cascaded with a static nonlinear subsystem. Wiener system is a system consisted of a linear subsystem followed by a nonlinear subsystem, and Hammerstein system is a reversed structure of Wiener system, that is, a nonlinear subsystem is followed by a linear subsystem.

Because of the importance of these kinds of systems in practical applications (see [1], [2] and [3]), the identification problem has been an active research topic for many years. Both parametric and nonparametric approaches are utilized according to the representation of nonlinear subsystem, e.g. [4], [5], [6] and [7] for the former, [8], [9], [10] and [11] for the latter. In the parametric approach, the nonlinearity is considered as a linear combination of known functions, or is a piece-wise function in this paper. Hence, the system can be transformed into a linear regression form with respect to coefficients of linear subsystem and products of coefficients in both nonlinear and linear functions. In the nonparametric approach, the nonlinear subsystem is usually estimated at an arbitrary point. In this case, identification is equivalent to estimating unknown coefficients in the series expansion.

The main contribution of this paper is that we propose a novel method to the identification problem of Wiener system with nonlinearity being a discontinuous piece-wise linear function, which has been studied in [4], [6], [12], [13], etc. Both [4] and [6] provide recursive algorithms by using stochastic approximation approach, and strongly consistent analysis is given as well. The main drawback of these algorithms is a huge amounts of data points should be obtained to guarantee numerical convergence. This becomes intractable when financing cost of each experiment is expensive, or consuming time is long, e.g. the chemical process. In contrast to these algorithms, our proposed method called sparse optimization can overcome this difficulty. The intuition of our method is that data points generated by such Wiener system lie in the union of several hyperplanes (see Section

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.

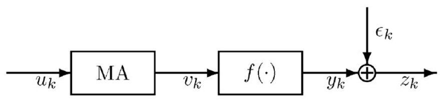

Fig. 1: Wiener Model

3.1). Therefore, the identification problem is equivalent to estimating the hyperplanes that contain most of data points, which is known as subspace clustering problem. When a set of data points that lie in several subspaces is given, subspace clustering focuses on the problem of estimating the number of subspaces, the dimension of each subspace, and the segmentation of data points corresponding to each subspace. In [14] and [15], the authors proposed a novel approach called sparse subspace clustering(SSC) to solve this problem, which is based on the observation that the sparsest representation of a vector would only choose vectors from the subspace in which it happens to lie in. Here, sparse representation means we used a few number of vectors for representation. However, to obtain sparsest representation we need solve ℓ_0-norm optimization problem, which is nonconvex and intractable. Instead, we utilize a classical approach called ℓ_1-norm minimization to relax this problem and solve it approximately. We provide a sufficient condition for recovering unknown parameters via ℓ_1-norm minimization in noise-free case. Even though the condition is not satisfied, or the measurements are corrupted with noise, this method performs well by using re-weighted ℓ_1-norm minimization technique [16].

The rest of the paper is organized as follows. Problem formulation is given in Section 2. In Section 3, we reformulate the identification problem as an ℓ_0-norm optimization problem, and utilize ℓ_1-norm minimization method to solve this problem approximately. To verify the performance of algorithm proposed in Section 3, we demonstrate some simulations in Section 4. Some conclusions are presented in Section 5.

2 Problem Formulation

Consider the Wiener system expressed by the block diagram shown in Figure 1. The nonlinearity part of the system

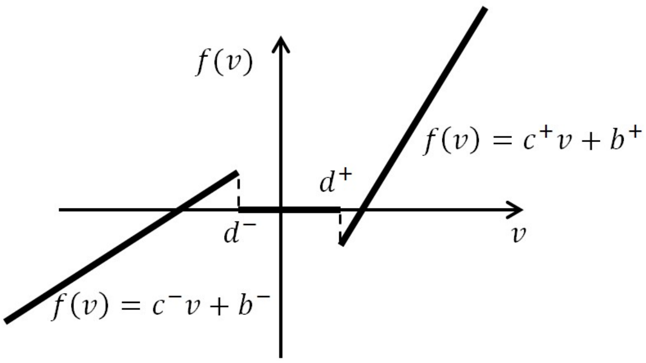

Fig. 2: Nonlinearity

is defined by a static piece-wise linear function:

$$f(v) = \begin{cases} c^+v + b^+, & \text{if } v > d^+ \\ 0, & \text{if } -d^- \le v \le d^+ \\ c^-v + b^-, & \text{if } v < d^- \end{cases}, \quad (1)$$

which is shown in Figure 2. Here we assume that $d^- < d^+$, both c^+ and c^- are nonzero.

Let the linear subsystem be described by a moving average (MA) model:

$$v_k = C(z)u_k, \quad (2)$$

where

$$C(z) = 1 + c_1 z + c_2 z^2 + \cdots + c_q z^q,$$

and z is a time delay operator, that is $zu(k) = u(k-1)$. Here we assume the order q of linear subsystem is known. The nonlinear output y_k is observed with additive noise ϵ_k, that is

$$z_k = y_k + \epsilon_k. \quad (3)$$

The parameters contained in linear subsystem and piece-wise nonlinear function are unknown. The identification problem is how to estimate parameters c^+, c^-, d^+, d^-, b^+, b^-, $c_i, i = 1, \cdots, q$ based on the observations $\{z_k\}_{k=1}^N$ and $\{u_k\}_{k=1}^N$.

Remark 1. As shown in Fig.2, we notice the nonlinear subsystem is consisted of three linear blocks, and the outputs of these blocks are overlapped. Hence, our model is more general than that in [4], [6] and [17]. In addition, to estimate d^+ and d^-, the methods proposed in these paper are not available since all of them need the assumption that the outputs of different linear block are disjoint.

3 Main results

In this section, we present our main results of this paper. We begin with an observation that identification problem can be reformulated as a sparse optimization problem. Instead of solving this problem directly, we use ℓ_1-norm convex relaxation to approximate the solution. Then we present an algorithm to estimate the parameters. Throughout the first three subsections, we assume the measurements are noise-free, that is $z_k = y_k$. The noisy measurements case is discussed in Section 3.4 and Section 3.5.

3.1 Sparse Optimization Method

In this subsection, we transform the identification problem into a sparse optimization problem. Firstly, we rewrite the system as a linear regression form with respect to coefficients of linear subsystem and products of coefficients in both nonlinear and linear functions. Then, the outputs of system are exactly lying in three hyperplanes, and the identification problem is equivalent to estimating the coefficients of hyperplanes. Finally, we proposed a sparse optimization method to solve it.

According to (2), substituting the output v_k of MA subsystem into (1) implies

$$y_k = \begin{cases} c^+u_k + \cdots + c^+c_q u_{k-q} + b^+, & \text{if } v_k > d^+ \\ c^-u_k + \cdots + c^-c_q u_{k-q} + b^-, & \text{if } v_k < d^- \\ 0, & \text{otherwise} \end{cases}. \quad (4)$$

Now, we rewrite (4) as a compact form

$$y_k = \begin{cases} \theta_1^T \phi_k, & \text{if } v_k > d^+ \\ \theta_2^T \phi_k, & \text{if } v_k < d^- \\ 0, & \text{otherwise} \end{cases}, \quad (5)$$

where

$$\theta_1^T \triangleq [c^+, c^+c_1, \cdots, c^+c_q, b^+], \quad (6)$$
$$\theta_2^T \triangleq [c^-, c^-c_1, \cdots, c^-c_q, b^-], \quad (7)$$
$$\phi_k^T \triangleq [u_k, u_{k-1}, \cdots, u_{k-q},\ 1]. \quad (8)$$

Here, θ_1 and θ_2 are unknown parameters, and all but the last component of ϕ_k are input signals, which can be designed. When θ_1 and θ_2 are identified, estimation of $c^+, c^-, b^+, b^-, c_i, i = 1, \cdots, q$ are followed from (6) and (7) with simple computations. Therefore, in the rest of this paper, we focus our works on identification of θ_1 and θ_2.

From (5), notice that each data point (ϕ_k, y_k) is lying on one of the three hyperplanes. In other words, data points $\{\phi_k, y_k\}_{k=1}^N$ are sampled from three hyperplanes. Hence, such Wiener system can also be seen as a linear switched systems, and the switch time are unknown since they depend on v_k, which are also unknown.

We split data set $\{\phi_k, y_k\}_{k=1}^N$ into three parts according to which hyperplane they are lying on:

$$\mathcal{A}_1 \triangleq \{k : y_k = \theta_1^T \phi_k, 1 \le k \le N\},$$
$$\mathcal{A}_2 \triangleq \{k : y_k = \theta_2^T \phi_k, 1 \le k \le N\},$$
$$\mathcal{A}_3 \triangleq \{k : y_k = 0, 1 \le k \le N\},$$

and N_1, N_2, N_3 are their corresponding cardinality numbers. Note that $\mathcal{A}_1$ and $\mathcal{A}_2$ are unknown, but $\mathcal{A}_3$ is known. Without loss of generality, suppose that $N_1 > N_2$ and set $N \doteq N - N_3$, otherwise we can delete data points lying on hyperplane $y = 0$. Here, $a \doteq b$ means set the value of b to a.

Construct data matrix X and vector Y as follow:

$$X = [\phi_1, \phi_2, \cdots, \phi_N], \ Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}. \quad (9)$$

Then, according to (5) and $N_3 = 0$,

$$\|Y - X^T\theta_1\|_0 \le N - N_1,\ \|Y - X^T\theta_2\|_0 \le N - N_2, \quad (10)$$

where $\|x\|_0$ is the number of nonzero components of x. Define an estimation error vector $E(\theta)$ depends on θ as $E(\theta) = Y - X^T\theta$. It follows from (5) that $E(\theta_1)$ and $E(\theta_2)$ are sparse vectors if N_1 and N_2 are enough large. Here we call a vector is sparse if most of its components are zero.

Use the above observations and $N_1 > N_2$, we estimate θ_1 by solving l_0 optimization problem:

$$\hat{\theta}_1 = \arg\min_{\theta} \|Y - X^T\theta\|_0. \quad (11)$$

At the same time, the data set $\mathcal{A}_1$ is estimated by

$$\hat{\mathcal{A}}_1 = \{k : (Y - X^T\hat{\theta}_1)_k = 0\},$$

where $(x)_k$ denotes the kth component of x.

In order to estimate θ_2, we construct a data matrix $X_{\hat{\mathcal{A}}_1^c}$ and measurement vector $Y_{\hat{\mathcal{A}}_1^c}$ as

$$X_{\hat{\mathcal{A}}_1^c} = [\phi_{i_1}, \phi_{i_2}, \cdots,],\ Y_{\hat{\mathcal{A}}_1^c} = \begin{bmatrix} y_{i_1} \\ y_{i_2} \\ \vdots \end{bmatrix}, \quad (12)$$

where $i_j \in \{1, 2, \cdots, N\} - \hat{\mathcal{A}}_1$. Now, θ_2 is estimated by solving l_0 optimization problem:

$$\hat{\theta}_2 = \arg\min_{\theta} \|Y_{\hat{\mathcal{A}}_1^c} - X^T_{\hat{\mathcal{A}}_1^c}\theta\|_0, \quad (13)$$

and the estimation for data set $\mathcal{A}_2$ is

$$\hat{\mathcal{A}}_2 = \{k : (Y - X^T\hat{\theta}_2)_k = 0\}.$$

From (11) and (13), notice that the methods to estimate θ_1 and θ_2 are similar. Therefore, most of our analysis in the following can be extended to θ_2 directly.

3.2 A Sufficient Condition

In this subsection, we introduce some basic notions and results about compressed sensing and deduce sufficient conditions for recovering θ_1 and θ_2 by solving ℓ_0-norm optimization problems (11) and (13) respectively.

Definition 1 (Spark)**.** The spark of a matrix $\Phi \in \mathbb{R}^{m\times n}$ is defined as

$$\mathrm{spark}(\Phi) = \min_{x\in\mathcal{N}(\Phi)\setminus\mathbf{0}} \|x\|_0.$$

The $\mathrm{spark}(\Phi)$ is also seen as the smallest number of columns of Φ that are linear dependent. Recall that $\mathrm{rank}(\Phi)$ is the maximal number of columns from Φ that are linear independent. It turns out that $\mathrm{spark}(\Phi) \le \mathrm{rank}(\Phi) + 1$, and the following example show that $\mathrm{spark}(\Phi)$ can be much smaller than $\mathrm{rank}(\Phi)$. However, when Φ is a random matrix, e.g. all the entries of Φ are sampled from Gaussian random variables with independent identical distribution, the equality holds almost surely(a.s.) [18].

Example 1. Let

$$A = \begin{bmatrix} 1 & & \\ 0 & & \\ \vdots & \mathbf{I} & \\ 0 & & \end{bmatrix}$$

where I is a $p \times p$ identity matrix. Then $\mathrm{spark}(A) = 2$, but $\mathrm{rank}(A) = p$.

Definition 2 (Mutual coherence)**.** The mutual coherence $\mu(\Phi)$ of a matrix $\Phi \in \mathbb{R}^{m\times n}$ is the largest absolute value of the cross-correlations between the columns of Φ:

$$\mu(\Phi) = \max_{1\le i<j\le n} \frac{|\langle a_i, a_j\rangle|}{\|a_i\|_2\|a_j\|_2},$$

where a_i is the ith column of Φ, and without loss of generality, we assume that all the columns of Φ are nonzero.

Notice that $\mu(\Phi)$ measures the smallest angle between any two columns of Φ. Both $\mathrm{spark}(\Phi)$ and $\mu(\Phi)$ can be seen as metrics to measure how rich the data contained in columns of Φ. The smaller $\mu(\Phi)$ is (or the larger $\mathrm{spark}(\Phi)$ is), the richer the data is. A relationship between these two metrics are below.

Lemma 1 (see [19])**.** *For any matrix Φ, it holds that*

$$\mathrm{spark}(\Phi) \ge 1 + \frac{1}{\mu(\Phi)}.$$

Without loss of generality, assume that X^T is a full column rank matrix (data set is sufficient large). Since $Y = X^T\theta_1 + E(\theta_1)$, if we can determine $E(\theta_1)$, then $\theta_1 = (XX^T)^{-1}X(Y - E(\theta_1))$. This problem is also known as decoding in coding theory [20], where θ_1 is plaintext, $X^T\theta_1$ is ciphertext. The receiver observes $X^T\theta_1$ with an additive sparse noise $E(\theta_1)$, and wishes to recover θ_1 when X and Y are known.

Now, we give our first result on recovering θ_1 by solving a l_0 optimization problem.

Theorem 1. *If there exists a matrix Φ such that $\Phi X^T = 0$ and satisfies $\mathrm{spark}(\Phi) > 2(N - N_1)$, then ℓ_0-norm optimization (11) recovers θ_1 exactly.*

Proof. By contradiction. Assume the solution of (11) is $\hat{\theta}$ and $\theta_1 \ne \hat{\theta}$. Since $\|E(\theta_1)\|_0 \le N - N_1$ and $\hat{\theta}$ is the optimal solution, $\|Y - X^T\hat{\theta}\|_0 \le N - N_1$. Let $\hat{E} = Y - X^T\hat{\theta}$, then $E(\theta_1) \ne \hat{E}$ because X^T is full column rank and $\theta_1 \ne \hat{\theta}$. And it follows from $\Phi X^T = 0$ that $\Phi E(\theta_1) = \Phi Y$ and $\Phi\hat{E} = \Phi Y$. Hence, $\Phi(E_1 - \hat{E}) = 0$. However, $\|E_1\|_0 \le N - N_1$ and $\|\hat{E}\|_0 \le N - N_1$ implies $\|E_1 - \hat{E}\|_0 \le \|E_1\|_0 + \|\hat{E}\|_0 \le 2(N - N_1)$, which is contradicted to the hypothesis $\mathrm{spark}(\Phi) > 2(N - N_1)$. □

Remark 2. A simple choice of Φ is $I - X^T(XX^T)^{-1}X$. Since all the entries except the last columns of X can be designed, we can make $\mathrm{spark}(\Phi)$ large. For example, when the input signals $\{u_k\}$ are sampled independently from random variables with Gaussian distribution, then $\mathrm{spark}(\Phi)$ is approximately $N - q - 2$ [21].

Let $\hat{\theta}_1$ be the solution to (11), then the parameters $c^+, b^+, c_1, \cdots, c_q$ are estimated by

$$\hat{c}^+ = (\hat{\theta}_1)_1,\ \hat{b}^+ = (\hat{\theta}_1)_{q+2},\ \hat{c}_i = \frac{(\hat{\theta}_1)_{i+1}}{\hat{c}^+}, i = 1, \cdots, q, \quad (14)$$

where $(\hat{\theta}_1)_j$ denotes the jth component of $\hat{\theta}_1$. Now we provide an estimation of d^+. Let the input signals u_k be i.i.d.

Gaussian random variables $\mathcal{N}(0, \sigma_u^2)$, then the outputs $v(k)$ of linear subsystem is a Gaussian stationary with distribution $\mathcal{N}(0, \sigma_v^2)$ [22], where $\sigma_v^2 = (1 + c_1^2 + \cdots + c_q^2)\sigma_u^2$. With this observation, we estimate d^+ according to the probability $\mathrm{P}(v_k > d^+)$, which is approximately equal to $\frac{|\hat{\mathcal{A}}_1|}{N+N_3}$, where $|A|$ denotes the cardinality of set A. Hence, an estimator $\hat{d}^+$ of d^+ is given follow by solving a generalization equation:

$$1 - G(\frac{\hat{d}^+}{\sigma_v}) = \frac{|\hat{\mathcal{A}}_1|}{N + N_3}, \tag{15}$$

where $G(x) \triangleq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} \, \mathrm{d}\, t$.

Remark 3. Notice that Gaussian distribution $G(x)$ is a strictly increasing function, thus, the solution $\hat{d}^+$ to equation (15) is unique. By Law of large number, $\frac{|\hat{\mathcal{A}}|_1}{N+N_3} = \frac{|\mathcal{A}_1|}{N+N_3} \to \mathrm{P}(v_k > d^+)$ a.s. as $N \to \infty$. Hence, $\hat{d}^+ \to d^+$ a.s. as $N \to \infty$, and the accuracy of estimation $\hat{d}^+$ heavily depends on the number of data points.

Corollary 1. *Under the condition of Theorem 1, the estimations given in (14) are exactly equal to their true values.*

Although we recover parameters by solving ℓ_0-norm optimization problem, it is intractable and NP hard. In the next subsection, we use ℓ_1-norm convex relaxation technique to recover parameters, which can be realized by a computational efficient algorithm.

3.3 ℓ_1-norm Convex Optimization Method

The ℓ_0-norm optimization problems (11) and (13) are combinatorial non-convex optimization and unsolvable using polynomial time algorithms. Inspired by compressed sensing theory [20], a common used method is replacing ℓ_0-norm by ℓ_1-norm. This technique makes the optimization convex, and the sparsity character of ℓ_0-norm is retained.

The following lemma states that under some conditions, ℓ_1-norm optimization problem is equivalent to the intractable ℓ_0-norm optimization problem.

Lemma 2 (see Theorem 7 of [23])**.** *If $\Phi x = y$ and*

$$\|x\|_0 < \frac{1}{2}(1 + \frac{1}{\mu(\Phi)}), \tag{16}$$

then the following two optimization problems are equivalent:

$$\min \|x\|_0 \quad \textit{subject to} \quad \Phi x = y \tag{17}$$

$$\min \|Wx\|_1 \quad \textit{subject to} \quad \Phi x = y \tag{18}$$

where $W \triangleq \mathrm{diag}(w(1), \cdots, w(N))$ is a diagonal matrix and the ith diagonal entry $w(i) \triangleq \|\phi_i\|_2$, ϕ_i is the ith column of Φ. Furthermore, the solution is unique.

Now we utilize this lemma to obtain a sufficient condition for recovering unknown parameters using ℓ_1-norm optimization method.

Theorem 2. *If there exists a matrix Φ satisfies $\Phi X^T = 0$ and*

$$N - N_1 < \frac{1}{2}(1 + \frac{1}{\mu(\Phi)}), \tag{19}$$

then $E(\theta_1)$ is the solution to ℓ_1-norm optimization problem:

$$\min_E \|WE\|_1 \textit{ subject to } \Phi E = \Phi Y, \tag{20}$$

where W is defined in Lemma 2. Furthermore, θ_1 is also recovered.

Proof. It follows directly from Lemma 2 that E is exactly recovered by solving (20), where x and y are replaced by E and ΦY respectively. The full column rank property of X^T guarantees the solution to $X^T\theta = Y - E(\theta_1)$ is unique, that is $\theta_1 = (XX^T)^{-1}X(Y - E(\theta_1))$. □

Remark 4. Determining the mutual coherence $\mu(\Phi)$ costs a great deal of computations [24], and is actually a NP hard problem. With the intuition of Lemma 1, we approximate $1 + \frac{1}{\mu(\Phi)}$ by spark(Φ). When Φ is selected as $I - X^T(XX^T)^{-1}X$, a rough approximation of $1 + \frac{1}{\mu(\Phi)}$ is $N - q - 2$.

The above approach can be directly applied to estimation of θ_2 and c^-, b^-. Recall the notations introduced in Section 3, since we recover θ_1 exactly, $\hat{\mathcal{A}}_1 = \mathcal{A}_1$, so $X_{\hat{\mathcal{A}}_1^c} = X_{\mathcal{A}_1^c}$ and $Y_{\hat{\mathcal{A}}_1^c} = Y_{\mathcal{A}_1^c}$. For simplification, denote $X_2 = X_{\mathcal{A}_1^c}$ and $Y_2 = X_{\mathcal{A}_1^c}$. Set $E_2 = Y_2 - X_2^T\theta_2$. Then, the following theorem concludes that E_2 can be recovered under some conditions.

Theorem 3. *If there exists a matrix Ψ satisfies $\Psi X_2^T = 0$ and*

$$N - N_1 - N_2 < \frac{1}{2}(1 + \frac{1}{\mu(\Psi)}), \tag{21}$$

then E_2 is the unique solution to ℓ_1-norm optimization problem:

$$\min \|ME\|_1 \textit{ subject to } \Psi E = \Psi Y_2, \tag{22}$$

where $M \triangleq diag(m(1), \cdots, m(N - N_1))$ is a diagonal matrix and the ith diagonal entry $m(i) \triangleq \|\psi_i\|_2$, ψ_i is the ith column of Ψ.

Proof. The proof is similar to the argument in Theorem 2, so we omit it here. □

Remark 5. Note that the condition (21) always holds under the assumption that $N_3 = 0$. In other words, in the noiseless measurements case, if we can recover θ_1, then θ_2 is recoverable as well. In the next section, we will consider the noisy measurements case, this assumption doesn't hold any longer.

3.4 Noisy Measurements

We turn to the problem that output y_k is corrupted by additive Gaussian noise ϵ_k, that is, the observations are $z_k = y_k + \epsilon_k$.

The assumption that $\mathcal{A}_3$ is known becomes invalid in this case. A simple method to estimate it is $\hat{\mathcal{A}}_3 = \{k : |y_k| \leq \delta\sigma^2\}$, where σ^2 is the variance of noise and $\delta \in (0, 1]$. If σ^2 is much smaller than the magnitude of most of y_k in $\mathcal{A}_1$ and $\mathcal{A}_2$, then $\hat{\mathcal{A}}_3$ contains a few indexes belong to $\mathcal{A}_1$ and $\mathcal{A}_2$, and most of indexes belong to $\mathcal{A}_3$ remain in $\hat{\mathcal{A}}_3$. To reduce the influence of noise on estimating parameters, we use a common approach called ℓ_2-norm regularization:

$$\min_{\theta, \mathcal{E}} \frac{1}{2}\|\mathcal{E}\|_2^2 + \gamma\|WE(\theta)\|_1, \text{ subject to } E(\theta) = Z - X^T\theta - \mathcal{E}, \tag{23}$$

where data matrices are defined as

$$X = [\phi_{i_1}, \phi_{i_2}, \cdots], Z = [z_{i_1}, z_{i_1}, \cdots]^T \tag{24}$$

for all $i_j \in \{1, \cdots, N\} \setminus \hat{\mathcal{A}}_3$ and W is a diagonal weighted matrix with positive diagonal entries. The first term of objective function make the magnitude of $\mathcal{E}$ be small and the second term encourage $E(\theta)$ to be sparse. Here we introduce a regularization parameter γ to balance the sparsity of $Y - X^T\theta_1$ and the magnitude of noise $\{\epsilon_k\}$. A discussion about how to select γ is presented in [25].

Remark 6. In the case where $N_1 \gg N_2$, the above convex optimization problem (23) is motivated to estimate θ_1, and the program for identifying θ_2 is similar to (23) except slight modification, where X and Z are replaced by $X_{\hat{\mathcal{A}}_1^c}$ and $Z_{\hat{\mathcal{A}}_1^c}$ respectively.

3.5 Enhance Sparsity via Re-weighted ℓ_1-norm Minimization

When the outputs y_k are corrupted with noise, $Z - X^T\theta_1$ is not sparse. It is still possible to estimate effectively $Z - X^T\theta_1$ via a technique called re-weighted ℓ_1-norm minimization [16].

For the noise-free case, a weighted ℓ_1-norm minimization problem is formulated as:

$$\min_{E} \|\bar{W}WE\|_1, \quad \text{subject to } \Phi E = \Phi Y, \tag{25}$$

where $\bar{W} = \text{diag}(\bar{w}(1), \cdots, \bar{w}(N)$ is a weighted matrix, and W is another weighted matrix defined in Lemma 2. The functions of $\bar{W}$ and W are different. The former one can be seen as some prior knowledge about the locations of supports. In this way, one assigns small weights to the components that are nonzero with high probability, otherwise, assign large weights. For example, if we known that $y(k)$ is generated by $\theta_1^T\phi(k)$, then $(E(\theta_1))_k$ must be 0, so we select a large weight $\bar{w}(k)$. The latter one is motivated to balance the difference of magnitudes of column vectors of Φ. For the noisy case, a similar method is

$$\min_{\theta,\mathcal{E}} \frac{1}{2}\|\mathcal{E}\|_2^2 + \gamma\|\bar{W}WE(\theta)\|_1$$
$$\text{subject to } E(\theta) = Z - X^T\theta - \mathcal{E}.$$

Even though there is no prior knowledge about the locations of supports, we can use an iterative algorithm, and reconstruct the weighted matrix $\bar{W}$ based on the estimation of previous step, which is shown in Algorithm 1.

To summarize the results of this section, we propose a reweighted l_1 minimization algorithm (see Algorithm 1) to estimate $c^+, b^+, d^+, c_1, \cdots, c_q$ with noisy measurements.

4 Simulations

In this section, we give numerical examples to demonstrate that our identification method recover unknown parameters in linear and nonlinear subsystem with overwhelming probability in the noise-free case. In addition, we also show that Algorithm 1 performs well in the noisy case. To solve convex optimization problem in this algorithm, we use CVX, a package for specifying and solving convex programs [26].

Let the nonlinear subsystem be described by

$$f(v) = \begin{cases} 0.4v - 0.6 & \text{if } v > 0.4 \\ 0 & \text{if } -0.6 \le v \le 0.4 \\ 0.85v + 0.7 & \text{if } v < -0.6 \end{cases}$$

Algorithm 1 Identification via re-weighted ℓ_1-norm minimization

Input: Sample data: $\{u(k), y(k)\}_{k=1}^N$, maximum iterative number l_{max}, variance σ^2, small positive numbers ϵ, δ, η.
Initialization: iterative counter $l = 0$; Estimate $\hat{\mathcal{A}}_3 \doteq \{k : |y_k| \le \delta\sigma^2\}$, $N_3 \doteq |\hat{\mathcal{A}}_3|$; construct X and Z as (24); $\Phi \doteq I - X^T(XX^T)^{-1}X$, $w(i) \doteq \|\phi(i)\|_2$, and $\phi(i)$ is the ith column of Φ, $\bar{w}_0(i) \doteq 1$ and $W \doteq \text{diag}(w(1), \cdots, w(N))$, $\bar{W}_0 \doteq \text{diag}(\bar{w}_0(1), \cdots, \bar{w}_0(N))$.
while $l < l_{max}$ **do**
Solve re-weighted ℓ_1-norm minimization problem:

$$\min_{\theta,\mathcal{E}} \frac{1}{2}\|\mathcal{E}\|_2^2 + \gamma\|\bar{W}_l WE(\theta)\|_1,$$
$$\text{subject to } E(\theta) = Z - X^T\theta - \mathcal{E}.$$

Update the weights: for each $i = 1, \cdots, N - N_3$,

$$\bar{w}_{l+1}(i) \doteq \frac{1}{|(E(\theta_l))_i| + \eta},$$

and

$$\bar{W}_{l+1} \doteq \text{diag}(\bar{w}_{l+1}(1), \cdots, \bar{w}_{l+1}(N - N_3)),$$

where η is a small positive number used to avoid the denominator being zero.
end while
Estimation: $\hat{\theta}_1 \doteq \theta_{l_{max}}$; $\hat{\mathcal{A}}_1 \doteq \{k : |(Z - X^T\hat{\theta}_1)_k| < \epsilon\}$;
Identification:

$$\hat{c}^+ \doteq (\hat{\theta}_1)_1, \ \hat{b}^+ \doteq (\hat{\theta}_1)_{q+2}, \ \hat{c}_i \doteq \frac{(\hat{\theta}_1)_{i+1}}{\hat{c}^+}, i = 1, \cdots, q,$$

and solve equation (15) for $\hat{d}^+$.

Table 1: Percentage of recovering θ_1 versus $N_1/(N - N_3)$ via Algorithm 1 in noise-free case

$\frac{N_1}{N-N_3}(\%)$	50	53	56	59	62
recover(%)	85	93	97	100	100

and the linear subsystem ($q = 4$) be described by

$$v_k = u_k + 0.81u_{k-1} + 0.61u_{k-2} - 0.2u_{k-3} - 0.45u_{k-4}.$$

The excitation input $\{u_k\}$ are independent identical Gaussian random variables with distribution $\mathcal{N}(0, 2)$

In the first experiment, we verify the ability of l_1 optimization method to recover θ_1 when $\epsilon_k = 0$. According to Theorem 2 and Remark 4, if N_1 is roughly larger than $N - \frac{1}{2}(N - q - 2) = 53$, Algorithm 1 may recover θ_1 with overwhelming probability. We design the experiment: let $N = 100$, fix N_1, and run Algorithm 1 for 100 times, then compute the percentage of recovering θ_1 successfully. Set $\eta = 0.1$, the simulation results are shown in Table 1, which are in accordance with our assertions.

In the second experiment, we test the performance of Algorithm 1 for the noisy measurements case. Let ϵ_k be Gaussian white noise with distribution $\mathcal{N}(0, 0.1^2)$. Set $\eta = 0.1$, $\delta = 0.5$, $l_{max} = 10$, $\gamma = 0.01$. Fix the percentage of N_1, we average the estimation values of c^+, b^+, d^+, c_i, $i = 1$, $\cdots$, q for 100 times by using Algorithm 1. The results are

Table 2: Estimation values of c^+, b^+, d^+, c_i, $i = 1, \cdots, q$ for different percentage N_1/N in noisy measurements case via Algorithm 1

$\frac{N_1}{N}(\%)$	55	58	60	65	True value
$\hat{c}^+$	0.36	0.38	0.39	0.39	0.40
$\hat{b}^+$	-0.46	-0.53	-0.54	-0.58	-0.60
$\hat{d}^+$	0.25	0.31	0.33	0.36	0.40
$\hat{c}_1$	0.82	0.82	0.80	0.81	0.80
$\hat{c}_2$	0.64	0.62	0.60	0.61	0.61
$\hat{c}_3$	-0.21	-0.20	-0.19	-0.20	-0.20
$\hat{c}_4$	-0.48	-0.47	-0.44	-0.46	-0.45

presented in Table 2. As seen, all the estimation values except $\hat{d}^+$ are closely to true values. This phenomenon is in accordance with Remark 3, where we state that the estimation accuracy of d^+ heavily depends on the number of data points, and $N = 100$ is small in our experiment.

5 Conclusions

In this paper, we discuss a new approach to the identification of Weiner system with nonlinearity being a piece-wise linear function. When the measurements are noise-free, we first show that sparse optimization method can recover the unknown parameters contained in linear and nonlinear subsystems under some conditions. Since solving ℓ_0-norm optimization is still intractable, we use ℓ_1-norm convex relaxation and re-weighted ℓ_1-norm minimization to tackle this NP hard problem. When the measurements are corrupted with noise, we use ℓ_2-norm regularization technique to deal with the noise. For further research, it is of interest to relax the conditions for recovering parameters, and consider some more general linear and nonlinear subsystems in the Wiener model.

References

[1] R. C. Emerson, M. J. Korenberg, and M. C. Citron, "Identification of complex-cell intensive nonlinearities in a cascade model of cat visual cortex," *Biological Cybernetics*, vol. 66, no. 4, pp. 291–300, 1992.

[2] F. Jurado, "A method for the identification of solid oxide fuel cells using a hammerstein model," *Journal of Power Sources*, vol. 154, no. 1, pp. 145 – 152, 2006.

[3] A. Kalafatis, N. Arifin, L. Wang, and W. Cluett, "A new approach to the identification of ph processes based on the wiener model," *Chemical Engineering Science*, vol. 50, no. 23, pp. 3693 – 3701, 1995.

[4] H. Chen, "Recursive identification for Wiener model with discontinuous piece-wise linear function," *Automatic Control, IEEE Transactions on*, vol. 51, no. 3, pp. 390–400, 2006.

[5] P. Celka, N. Bershad, and J. Vesin, "Stochastic gradient identification of polynomial wiener systems: analysis and application," *Signal Processing, IEEE Transactions on*, vol. 49, no. 2, pp. 301–313, 2001.

[6] Y. Huang, H. Chen, and H. Fang, "Identification of wiener systems with nonlinearity being piecewise-linear function," *Science in China Series F: Information Sciences*, vol. 51, no. 1, pp. 1–12, 2008.

[7] E. Bai, "Identification of linear systems with hard input nonlinearities of known structure," in *Block-oriented Nonlinear System Identification* (F. Giri and E. Bai, eds.), vol. 404 of *Lecture Notes in Control and Information Sciences*, pp. 259–270, Springer London, 2010.

[8] H. Chen, "Pathwise convergence of recursive identification algorithms for hammerstein systems," *Automatic Control, IEEE Transactions on*, vol. 49, no. 10, pp. 1641–1649, 2004.

[9] E. Bai, "Frequency domain identification of hammerstein models," *Automatic Control, IEEE Transactions on*, vol. 48, no. 4, pp. 530–542, 2003.

[10] W. Greblicki, "Stochastic approximation in nonparametric identification of hammerstein systems," *Automatic Control, IEEE Transactions on*, vol. 47, no. 11, pp. 1800–1810, 2002.

[11] Z. Lang, "A nonparametric polynomial identification algorithm for the hammerstein system," *Automatic Control, IEEE Transactions on*, vol. 42, no. 10, pp. 1435–1441, 1997.

[12] J. Vörös, "Parameter identification of wiener systems with discontinuous nonlinearities," *Systems Control Letters*, vol. 44, no. 5, pp. 363 – 372, 2001.

[13] Y. Huang and H. Chen, "Parameter identification of wiener systems with arma linear subsystem and discontinuous piecewise-linear function," *Acta Mathematicae Applicatae Sinica(in Chinese)*, vol. 31, no. 6, pp. 961–980, 2008.

[14] E. Elhamifar and R. Vidal, "Sparse subspace clustering," in *Computer Vision and Pattern Recognition, 2009. IEEE Conference on*, pp. 2790–2797.

[15] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 11.

[16] E. J. Candès, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted l_1-minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5-6, pp. 877–905, 2008.

[17] Y. Huang, H. Chen, and H. Fang, "Recursive parameter identification of wiener systems with discontinuous piecewise-linear memoryless block and observation noise," in *Chinese Control Conference, 2006*, pp. 433–437.

[18] Y. Eldar and G. Kutyniok, *Compressed Sensing: Theory and Applications*. Cambridge University Press, 2012.

[19] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via l_1 minimization," *Proceedings of the National Academy of Sciences*, vol. 100, no. 5, pp. 2197–2202, 2003.

[20] E. J. Candès and T. Tao, "Decoding by linear programming," *Information Theory, IEEE Transactions on*, vol. 51, no. 12, pp. 4203–4215, 2005.

[21] L. Bako, "Identification of switched linear systems via sparse optimization," *Automatica*, vol. 47, no. 4, pp. 668 – 677, 2011.

[22] M. Loeve, *Probability theory*. Springer-Verlag New York, 4th ed. ed., 1977.

[23] A. Bruckstein, D. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM Review*, vol. 51, no. 1, pp. 34–81, 2009.

[24] A. M. Tillmann and M. E. Pfetsch, "The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing," *accepted for publication in Information Theory, IEEE Transactions on*.

[25] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[26] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.0 beta." http://cvxr.com/cvx, Sept. 2013.