Journal of Global Optimization **23:** 195–211, 2002.

An Optimization Based Robust Identification Algorithm in the Presence of Outliers

ER-WEI BAI
Department of Electrical and Computer Engineering, University of Iowa, Iowa City, IA 52242, USA (E-mail: erwei@icaen.uiowa.edu)

Abstract. In this paper, the robust identification problem is formulated in the framework of optimization with few violated constraints and an efficient numerical algorithm is presented with a low complexity. Moreover, it is shown that the proposed method requires minimum a priori information and is convergent in the presence of outliers.

Key words: Optimization, LP-type algorithm, System identification

1. Introduction

In parameter identification, it is well known that a single or few highly disturbed values of the measurement data, referred to as the outliers, have a substantial influence on the estimates. Thus, it is necessary to protect the estimates from these bad data. There is a large amount of research work along this direction, see the excellent books [4, 10, 12] and, in particular, their applications to system identification [7, 14]. The easiest way to improve the robustness of the estimates is probably to adopt a robust cost function, e.g., by using l_1 norm instead of l_2 norm because absolute deviations rather than squared ones are less sensitive to outliers. The M-estimator is another example in this direction [12]. Unfortunately, those simple remedies do not solve the problem in most cases. To this end, a minmax distribution approach to robust identification was initiated in [5] and summarized in [6]. The idea is to find a minimax distribution in terms of minimum sensitivity in the given class of distributions. This approach is especially useful for some optimal estimators such as Bayes and Maximum Likelihood estimators. Because of high sensitivity to deviations of the distributions, however, these optimal estimators may cease to work and become unstable in the presence of outliers. Influence statistics is also applied to rank and identify outliers. The most popular one is probably the Cook's distance [4, 10] for identifying how influential a particular observation is based on the least squares estimates. However, the method seems to be limited and is not automatic requiring good judgment by the designer. In addition to the computational complexity, it can be ineffective if two or more outliers are close. Another approach is to treat outliers as abrupt changes of the signal [1]. Then, the problem reduces to detecting these changes. The basic idea is multiple hypotheses testing. A model or several models are constructed based on a priori statistical information. Hypotheses testing are performed after receiving input-output measurements. This approach also needs extensively a priori statistical information. There are also several practical ways to robustify the estimates in the presence of outliers [7, 14]. For instance, the changes due to outliers can be drastic in the data record especially in the residual plot and thus outliers may be picked up by eye inspection of the residual plot. Though these practical ways require much less a priori knowledge, they need human interaction and hence reduce the efficiency of the estimates. Further, these practical ways have less theoretical foundations and may not work in some cases. For example, as pointed out on p. 75 of [12], the residuals from the least squares can not be used for picking up outliers because outliers can possess very small least squares residuals as the least squares fit is pulled too much in the direction of these deviating points.

There exists a conceptually effective way to deal with the outliers. Suppose we have n measurements within which a small number k are outliers. To detect such k outliers, we may calculate an estimate by deleting k observations from n measurements. There are $n!/(n-k)!k!$ such combinations. After trying all $n!/(n-k)!k!$ combinations, we have the 'best' estimate and the corresponding k measurements are likely to be the outliers. This method is intuitively simply and effective. However, the computational complexity is exponential (combinatorial) in n, the total number of data points and hence is infeasible in a practical setting.

In this paper, we propose a new approach to deal with outliers: optimization with few violated constraints approach. Contrary to existing estimators, the proposed approach assumes a minimum amount of a priori statistical information on the noises, has a lower computational complexity and does not need human interaction. If the upper bound $k(n)$ on the number of ourliers out of n measurements is known, our scheme guarantees convergence with probability one in the presence of outliers. The computational complexity of our algorithm is bounded by

$$\min\{O(n \cdot k^{m+2}),\, O(n \cdot (m+2)^{k+1})\},$$

where m is the number of unknown parameters. This complexity is comparable to $O(n)$ for small $m, k << n$. The work reported in this paper is a continuation of our previous work [2] where an algorithm for optimization with few violated constraints was developed, analyzed and applied to identification in the setting of membership set identification. In this paper, we will apply the algorithm developed in [2] to stochastic identification. Because noise assumptions in the setting of membership set identification are quite different from that in the setting of stochastic identification, the results reported in this paper are non-trivial extensions of [2].

We end this section by outlining the paper. Section 2 describes the problem and illustrates the idea of our approach by a simple example. The system description is provided in Section 3, and Section 4 presents the algorithm to deal with outliers along with its complexity analysis. Convergence results are obtained in Section 5. Finally, some remarks are provided in Section 6.

2. A motivating example and problem statement

To clearly state the problem and to visualize the idea behind the main result of the paper, we consider a simple example of a scalar static equation

$$y_i = \theta + v_i, \quad i = 1, 2, \dots, n.$$

The goal is to estimate the unknown constant θ from the output observations $y_1, y_2, \dots, y_n$ corrupted by a noise sequence $v_1, v_2, \dots, v_n$. Let v_i's be independent, identically distributed and $p(v_i)$ be the probability density function of v_i. Then, the Maximum Likelihood Estimate (MLE) is given by

$$\theta_n = \arg\min_{\hat{\theta}} \left(-\sum_{i=1}^{n} \log p(y_i - \hat{\theta}) \right).$$

Now, further assume that v_i is the rectangular distribution in the interval $[-a, a]$ for some $a > 0$, i.e.,

$$p(v_i) = \begin{cases} \frac{1}{2a} & v_i \in [-a, a] \\ 0 & v_i \notin [-a, a] \end{cases}.$$

Then, the MLE θ_n can be played by any $\hat{\theta}$ satisfying inequalities

$$-a \leqslant y_i - \hat{\theta} \leqslant a, \quad i = 1, 2, \dots, n$$

which can be obtained by [11]

$$\begin{aligned} \theta_n &= \arg\min_{\hat{\theta}} \max_{1 \leqslant i \leqslant n} |y_i - \hat{\theta}| \\ &= \arg\min_{\hat{\theta}} \max \left\{ |\max_{1 \leqslant i \leqslant n} y_i - \hat{\theta}|, |\min_{1 \leqslant i \leqslant n} y_i - \hat{\theta}| \right\} = \frac{1}{2} \left(\max_{1 \leqslant i \leqslant n} y_i + \min_{1 \leqslant i \leqslant n} y_i \right) \\ &= \theta + \frac{1}{2} \left(\max_{1 \leqslant i \leqslant n} v_i + \min_{1 \leqslant i \leqslant n} v_i \right). \end{aligned} \tag{2.1}$$

It can be verified that $\theta_n \to \theta$ with probability one as $n \to \infty$. If, however, the noise distribution is approximately rectangular with some very small but non-zero tails outside the interval $[-a, a]$ or due to some fault we had a few bad measurements of y_i, the MLE estimate θ_n of (2.1) becomes unstable because it responds strongly to outliers. For instance, let $v_{i^*} = 3a$ be a single outlier at some $1 \leqslant i^* \leqslant n$, the MLE estimate of (2.1) becomes

$$\theta_n \to \theta + \frac{1}{2}(3a - a) = \theta + a.$$

We see that a single outlier changes the estimate completely, independent of n.

To reduce the effect of the outliers, let us take a closer look at the estimate (2.1). Let the set $I_n = [1, 2, \dots, n]$ and the set $I_n/1$ denote the collection of all subsets of I_n with $(n-1)$ elements. Further, let $s \subset I_n/1$ be any subset of $I_n/1$ that does not contain i^* and $s^* \subset I_n/1$ be any subset of $I_n/1$ that does contain i^*. It can be easily verified that for any $n \geqslant 2$

$$\min_{\hat{\theta}} \max_{i \in s \subset I_n/1} |y_i - \hat{\theta}| \leqslant \max_{i \in s \subset I_n/1} |y_i - \theta| \leqslant a \leqslant \min_{\hat{\theta}} \max_{i \in s^* \subset I_n/1} |y_i - \hat{\theta}|.$$

In other words, if we know that there is an outlier at some unknown i^*, by modifying the estimate (2.1) as

$$\theta_n = \arg \min_{\hat{\theta}} \min_{\bar{s} \in I_n/1} \max_{i \in \bar{s}} |y_i - \hat{\theta}| \tag{2.2}$$

any subset $s^* \subset I_n/1$ containing i^* can be identified and eliminated. In turn, the effects of the outliers can be removed. In fact, if there is indeed only one outlier, it is intuitively clear that the estimate given by (2.2) satisfying

$$\theta_n \to \theta$$

with probability one as $n \to \infty$. This is the basic idea behind our main results: to find a $\hat{\theta}$ and a subset $\bar{s} \subset I_n/1$ such that $\max_{i \in \bar{s}} |y_i - \hat{\theta}|$ is minimized. The intuition is that $\max_{i \in \bar{s}} |y_i - \hat{\theta}|$ is large if some of outliers are present in the set $\bar{s}$ and is small if no outlier is present. Therefore, by looking for a subset in $I_n/1$ and a $\hat{\theta}$ that minimize $\max_{i \in \bar{s}} |y_i - \hat{\theta}|$ effectively removes outliers from measurement data. In the next section, we will extend this simple idea to a very general setting of identification for unknown numbers of outliers. Moreover, the noise sequence is no longer assumed to be i.i.d. but the output of some unknown linear system driven by white noise.

3. System description

Consider a stable single input-single output discrete time system

$$y_i = \bar{\phi}_i^T \bar{\theta} + \bar{v}_i, \quad i = 1, 2, \ldots, n \tag{3.1}$$

where $y_i \in R$ is the output, $\bar{\theta} \in R^m$ the unknown parameter vector to be identified and $\bar{\phi}_i \in R^m$ the deterministic and measurable regressor. The noise

$$\bar{v}_i = \bar{v}_i^g + \bar{v}_i^b = \sum_{k=0}^{N-1} \alpha_k \eta_{i-k} + \bar{v}_i^b \in R \tag{3.2}$$

consists of two parts $\bar{v}_i^g$ and $\bar{v}_i^b$. The part $\bar{v}_i^b$ represents the bad measurement data or the outliers. It is assumed that $\bar{v}_i^b$ can be non-zero at any time with any magnitude. However, the number of times that $\bar{v}_i^b$ is non-zero is upper bounded by a small number

$$k(n) < n. \tag{3.3}$$

This upper bound k reflects the fact that outliers can happen at any time with a large magnitude, but they do not occur frequently.

The 'normal' part of the noise is denoted by $\bar{v}_i^g$ that is the output of some unknown linear time invariant system $\bar{v}_i^g = \sum_{k=0}^{N-1} \alpha_k \eta_{i-k}$ driven by a white noise η_i. The linear system part is represented by a FIR model with unknown impulse response α_k. For not trivializing the problem, we assume that not all $\alpha_k = 0$, otherwise $\bar{v}_i^g \equiv 0$. The order $N-1$ of the FIR model is also unknown and can be arbitrarily large but fixed. No minimum phase assumption is assumed on this FIR system. The white noise η_i is an independent, identically distributed bounded random sequence in the interval

$$\eta_i \in [\underline{a}, \bar{a}]$$

for some unknown $-\infty < \underline{a} < \bar{a} < \infty$. Again for not trivializing the problem, we assume that $\underline{a} < \bar{a}$. Otherwise, $\underline{a} = \bar{a}$ implies that η_i has a delta distribution which means η_i is a constant with probability one. The support interval $[\underline{a}, \bar{a}]$ does not necessarily contain the origin but is assumed to be tight in the sense that for any arbitrarily small $\rho > 0$,

$$\text{Prob}\{\underline{a} \leqslant \eta_i \leqslant \underline{a} + \rho\} \geqslant p_1(\rho) > 0 \tag{3.4}$$

and

$$\text{Prob}\{\bar{a} - \rho \leqslant \eta_i \leqslant \bar{a}\} \geqslant p_1(\rho) > 0 \tag{3.5}$$

for all $i \in [1, 2, \dots, n]$ and some $p_1(\rho) > 0$. Note that this tightness assumption does not add any restriction to the unknown bounds $\underline{a}$ and $\bar{a}$ at all. If $[\underline{a}, \bar{a}]$ is not tight, then there always exists some $[\underline{a}', \bar{a}'] \subset [\underline{a}, \bar{a}]$ such that $[\underline{a}', \bar{a}']$ is tight.

The system (3.1) represents a large class of discrete time systems. For instance, FIR systems as well as IIR systems modeled by some bases like Laguerre and/or Kautz functions. Also, the assumed setups of $\bar{v}_i^g$ and $\bar{v}_i^b$ allow a large class of noises including those with unbounded supports. For instance, consider a Gaussian distribution on η_i with unknown mean value μ and variance σ^2. Set $[\underline{a}, \bar{a}] = [\mu - 3\sigma, \mu + 3\sigma]$. Then, the probability of $\eta_i \notin [\mu - 3\sigma, \mu + 3\sigma]$ is very small and accordingly the effects of η_i outside $[\underline{a}, \bar{a}]$ can be considered as outliers summarized in the part of $\bar{v}_i^b$.

For the system (3.1) and a given measurement data y_i and $\bar{\phi}_i$ $i = 1, 2, \dots, n$, define the augmented system

$$\begin{aligned} y_i &= \bar{\phi}_i^T \bar{\theta} + \bar{v}_i = (\bar{\phi}_i^T, 1) \begin{pmatrix} \bar{\theta} \\ \bar{\theta}_{m+1} \end{pmatrix} + \bar{v}_i - \bar{\theta}_{m+1} \\ &= \phi_i^T \theta + v_i, \quad i = 1, 2, \dots, n \end{aligned} \tag{3.6}$$

with

$$\bar{\theta}_{m+1} = \frac{1}{2} \left(\sum_{0}^{N-1} \max\{\alpha_j \underline{a}, \alpha_j \bar{a}\} + \sum_{0}^{N-1} \min\{\alpha_j \underline{a}, \alpha_j \bar{a}\} \right)$$

and

$$\phi_i^T = (\bar{\phi}_i^T, 1),\ \theta^T = (\bar{\theta}^T, \bar{\theta}_{m+1}), v_i = \bar{v}_i - \bar{\theta}_{m+1}.$$

Note that $\bar{\theta}_{m+1}$ is unknown and will be determined later by the identification algorithm. y_i and ϕ_i on the other hand are available.

For a given data observation points $i \in I_n = [1, 2, \dots, n]$ and the upper bound $k(n)$ on the outliers, let the set

$$I_n/k(n) = \{[i_1, i_2, \dots, i_{n-k}]\ ;\ i_j \in [1, 2, \dots, n],\ i_j \neq i_l \text{ if } j \neq l\} \tag{3.7}$$

denote the collections of the subsets of $[1, 2, \dots, n]$ with $n-k$ distinct elements. Now consider an optimization problem: Consider the augmented system (3.6), to find a triplet $\hat{\epsilon}^*(n)$, $\hat{\theta}^*(n)$ and a subset $[i_1^*, i_2^*, \dots, i_{n-k}^*] \in I_n/k(n)$ such that the triplet solves the following minimization problem

$$\begin{aligned} &\min_{[i_1,\dots,i_{n-k}]\subset I_n/k(n)} \quad \min \hat{\epsilon}(n) \\ &\quad\text{subject to} \quad -\hat{\epsilon}(n) \leqslant y_i - \phi_i^T\hat{\theta}(n) \leqslant \hat{\epsilon}(n), \quad i \in [i_1, \dots, i_{n-k}]. \end{aligned} \tag{3.8}$$

Denote the solution $\hat{\theta}^*(n)$ by

$$\hat{\theta}^*(n) = \begin{pmatrix} \hat{\theta}_1^*(n) \\ \vdots \\ \hat{\theta}_m^*(n) \\ \hat{\theta}_{m+1}^*(n) \end{pmatrix}.$$

The meaning of the above minimization is to find a $\hat{\theta}$ and the subset $[i_1, \dots, i_{n-k}] \subset I_n/k(n)$ such that $\max_{i\in[i_1,\dots,i_{n-k}]} |y_i - \phi_i^T\hat{\theta}|$ is minimized. The intuition is again that $\max_{i\in[i_1,\dots,i_{n-k}]} |y_i - \phi_i^T\hat{\theta}|$ is large if some of outliers $\bar{v}_i^b$'s are present in the set $[i_1, \dots, i_{n-k}]$ and is small if no outlier is present. Therefore, by looking for a subset in $I_n/k(n)$ and a $\hat{\theta}$ that minimize $\max_{i\in[i_1,\dots,i_{n-k}]} |y_i - \phi_i^T\hat{\theta}|$ effectively removes outliers from measurement data.

4. Algorithm and Complexity Analysis

From the previous discussion, we see that the proposed robust identification scheme is basically to solve the optimization problem of (3.8). This can be done in the framework of LP-type optimization developed in [8] and further discussed in [2].

Consider the system (3.6). For a given input-output measurements y_i, ϕ_i, $i = 1, 2, \dots, n$, define

$$x^T = (x_1, ..., x_{m+1}, x_{m+2}) = (\hat{\theta}^T, \hat{\theta}_{m+2}),$$

and

$$\{h_i\} = \{x \in R^{m+2} : -x_{m+2} \leqslant y_i - \phi_i^T\hat{\theta} \leqslant x_{m+2}\}.$$

Let H be the constraint set

$$H = \{h_1, h_2, \dots, h_n\}$$

and $w(G)$ denote the lexicographically smallest point $x \in R^{m+2}$ that satisfies all the constraints in $G \subseteq H$. Here, the lexicographically smallest point means that the last coordinate is the most important. In other words, let $x = (x_1, ..., x_{m+2})^T$, $y = (y_1, ..., y_{m+2})^T$ be any two points in R^{m+2}. Then, x is lexicographically smaller than y, or $x < y$ if and only if

$$x_j < y_j \quad \text{for some } j \leqslant m + 2$$

and

$$x_k = y_k \quad \text{for all } k \text{ such that } j < k \leqslant m + 2.$$

Note that whether $x_i < y_i$ or $x_i > y_i$ for $i < j$ is irrelevant.

We now introduce a few definitions.

DEFINITION 4.1. A subset $B \subseteq H$ is called a basis if $w(G) < w(B)$ for all proper subsets $G \subset B$. A basis for a subset $G \subseteq H$, denoted by $B(G)$, is a basis $B \subseteq G$ with $w(B) = w(G)$.

DEFINITION 4.2. We say that a constraint $h \in H$ violates a set G if $w(G + h) > w(G)$. For $G \subseteq H$, we denote by $V(G)$ all the constraints of H violating G.

DEFINITION 4.3. Let $G \subseteq H$, then the level of G is defined as $|V(G)|$, i.e., the number of constraints violating G.

DEFINITION 4.4. The optimization problem (3.8) is non-degenerate if $w(B) \neq w(B')$ for any two distinct bases B and B' in H.

Clearly, for a given $k < n$, the optimization problem of (3.8) is equivalent to finding a basis G and the corresponding $w(G)$ that satisfies all n but at most k constraints and is the minimum. We denote by $\mathcal{B}_k$ the set of all bases in H of level k, i.e., the collection of all the bases representing the sets of level k, and $\mathcal{B}_{\leqslant k}$ for the set of bases of level at most k. What we need to find is a basis with the smallest value among all bases of level at most k. A trivial way to solve this problem is to find the minimum value for each collection of $n - k$ constraints. However, there are $n!/k!(n - k)!$ possible combinations and thus the computational complexity is high. The way we propose is to search all the bases of level k in an efficient way discussed in [2] and then to select the one with the minimum value.

4.1. ALGORITHM FOR SOLVING THE OPTIMIZATION PROBLEM (3.8)

Let (H, w) be non-degenerate. Given $n = |H|$ and $k \geqslant 0$, to find a basis in H that has the minimum value w and satisfies all n but at most $0 \leqslant k \leqslant n$ constraints.

Step 1. Determine the unique basis $\mathcal{B}_0 = B(H)$ and set $j = 1$.

Step 2. For each $\mathcal{B}_{j-1}$, determine all its neighbors in $\mathcal{B}_j$ by finding the basis of $H - V(B) - h$ for every basis $B \in \mathcal{B}_{j-1}$ and each $h \in B$. Check if the achieved basis $B(H - V(B) - h)$ coincides with any basis obtained before. If it is, then it is redundant and we remove it from the search path. Also, check if $V(B') = V(B) \bigcup \{h\}$, i.e., if the obtained is in $\mathcal{B}_j$. If the basis obtained is not redundant, go to Step 3.

Step 3. If $j = k$, go to Step 4. Otherwise, set $j = j + 1$ and go to Step 2.

Step 4. Find the basis with the smallest value of w among all $\mathcal{B}_{\leqslant k}$.

We remark that at each step of the algorithm, one needs to find a basis which has the minimum value of w. Finding a basis with the minimum w is exactly a linear programming problem. For instance, at level $k = 0$, finding a basis for H is to find a set of $(m + 2)$ constraints $\{-x_{m+2} \leqslant y_i - \phi_i\hat{\theta} \leqslant x_{m+2}\}$, $i = i_1, ..., i_{m+2}$ which intersect at a point that has the lexicographically smallest value. These $(m + 2)$ constraints constitute a basis for H. Or equivalently, finding a basis for H at $k = 0$ is to find a minimal $\hat{\epsilon} > 0$ and the corresponding $\hat{\theta}$ so that all n constraints $\{-\hat{\epsilon} \leqslant y_i - \phi_i\hat{\theta} \leqslant \hat{\epsilon}\}$ are satisfied. This is clearly a linear programming problem. Denote such a basis by $B(H) = \{b_1, ..., b_{m+1}\}$, with $b_i \in H$. Next, we can calculate the bases at $k = 1$ level, i.e., find bases for $H - V(B) - b_i)$, $i = 1, 2, ..., m + 2$, where $V(B)$ are the constraints in H violating $B(H)$. This is again a linear programming problem. In this sense, the algorithm requires to apply linear programming algorithms repeatedly for each level k. We notice that computational complexity of a linear programming problem is, in general, polynomial in n. In our setting, however, the dimension of x is fixed and this implies that the computational complexity of each linear programming is linearly bounded by the number of constraints n. This is one reason why we can achieve a low complexity stated in the following theorem.

THEOREM 4.1. *Let the problem* (H, w) *be non-degenerate. Then, the optimization problem of (3.8) can be solved by the above algorithm in* $O(n)$ *time for* $0 \leqslant k \leqslant 1$ *and in* $\min\{O(nk^{m+2}), O(n(m + 2)^{k+1})\}$ *time for any* $k \geqslant 2$.

Proof. The proof is similar to what provided in [2] and thus is omitted.

The result of Theorem 4.1 shows that the proposed algorithm is very efficient for small $m, k << n$. This is the case when the number of outliers is bounded by a small k. Note that in identification n can be very large and m is the dimension of the parameter to be identified and small.

We now consider a numerical example from [9]

$$y(z) = 0.51\frac{z^{-1}}{1 + 0.8z^{-1}}u(z) + 0.35\frac{z^{-1}(1 - 1.2z^{-1})}{(1 + 0.8z^{-1})^2}u(z) + v(z)$$

and

$$v(z) = (0.6 - 0.3z^{-1} + 0.1z^{-2})\eta(z).$$

This discrete system is in fact the discretized model of the continuous transfer function $g(s) = 1/(10s+1)(s+1)$ sampled with sampling period 1 (second). The unusual regressors are motivated by the Laguerre polynomials and the pole -0.8 was chosen between the true system poles, see [9] for details. The time domain model of the above system is given by

$$y_i = \phi_i^T \begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix} + v_i = (x_i, w_i)\begin{pmatrix}0.51\\ 0.35\end{pmatrix} + v_i$$

where

$$x_i = -0.8x_{i-1} + u_{i-1}, \quad w_i = -1.6w_{i-1} - 0.64w_{i-2} + u_{i-1} - 1.2u_{i-2},$$

$$v_i = 0.6\eta_i - 0.3\eta_{i-1} + 0.1\eta_{i-2}.$$

For simulation, $n = 600$, input u_i is a random sequence uniformly in $[-1, 1]$ and η_i is a random sequence uniformly in $[-0.3, 0.7]$. Let the maximum number of outlier be 6 (1% of 600) and the actual number be 5 at

$$v(50) = 3,\ v(150) = -4,\ v(250) = 5,\ v(450) = -6,\ v(550) = 7.$$

The top diagram of Figure 1 shows the estimates $\hat{\theta}$ of $(0.51, 0.35)^T$ by allowing removing $k = 0, 1, 2, 3, 4, 5, 6$ constraints respectively and the bottom one shows the parameter estimation errors $\|\hat{\theta} - \begin{pmatrix}0.51\\ 0.35\end{pmatrix}\|$ for $k = 0, 1, 2, 3, 4, 5, 6$ respectively. As expected, the parameter estimation errors at $k = 5, 6$ are almost zero.

5. Convergence results

In this section, we will show that the estimate obtained by the optimization of (3.8) converges to the true but unknown θ under some conditions. To this end, recall the definition $\bar{v}^g$. Then, we have

LEMMA 5.1. *$\bar{v}_i^g$ is a random variable with the support*

$$[\underline{\epsilon}, \bar{\epsilon}] = \left[\sum_{k=0}^{N-1} \underline{f}_k, \sum_{k=0}^{N-1} \bar{f}_k\right],$$

where

$$\bar{f}_k = \max\{\alpha_k\underline{a}, \alpha_k\bar{a}\}, \quad \underline{f}_k = \min\{\alpha_k\underline{a}, \alpha_k\bar{a}\}.$$

Moreover, the support $[\underline{\epsilon}, \bar{\epsilon}]$ of $\bar{v}_i^g$ is tight, i.e., for any arbitrarily small $\rho > 0$,

$$Prob\{\underline{\epsilon} \leqslant \bar{v}_i^g \leqslant \underline{\epsilon} + \rho\} \geqslant p_2(\rho) > 0$$

$$Prob\{\bar{\epsilon} - \rho \leqslant \bar{v}_i^g \leqslant \bar{\epsilon}\} \geqslant p_2(\rho) > 0$$

for all $i \in [1, 2, \ldots, n]$ and some $p_2(\rho) > 0$.

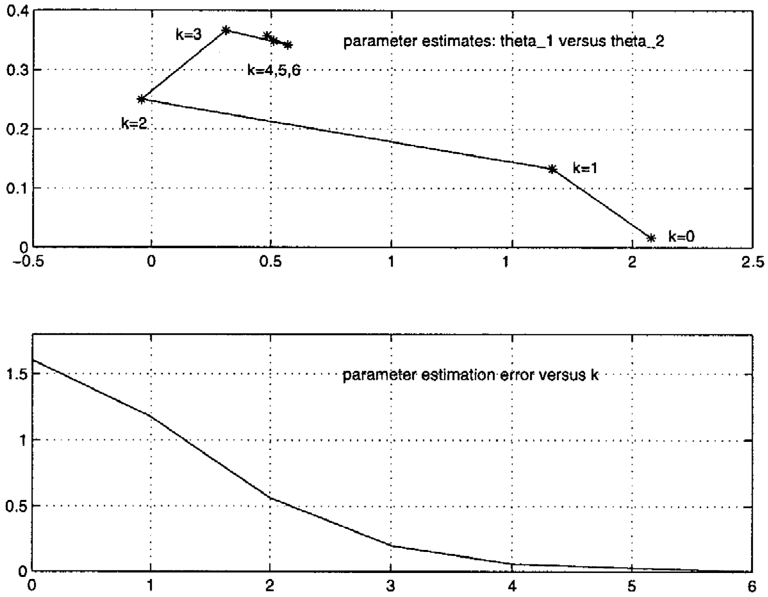

Figure 1.

Proof. Note that $\eta_i \in [\underline{a}, \bar{a}]$. This implies

$$\max_i \bar{v}_i^g = \max_i \sum_{k=1}^{N-1} \alpha_k \eta_{i-k} \leqslant \sum_{k=1}^{N-1} \bar{f}_k = \bar{\epsilon}$$

and

$$\min_i \bar{v}_i^g = \min_i \sum_{k=1}^{N-1} \alpha_k \eta_{i-k} \geqslant \sum_{k=1}^{N-1} \underline{f}_k = \underline{\epsilon}.$$

Now, $\eta_i, \ldots, \eta_{i-N+1}$ are independent and can be arbitrarily close to the bounds $\underline{a}$ and $\bar{a}$ with non-zero probability. This implies that $\bar{v}_i^g$ can be arbitrarily close to the bounds $\underline{\epsilon}$ and $\bar{\epsilon}$ with non-zero probability. In other words, the bound $[\underline{\epsilon}, \bar{\epsilon}]$ is tight.

LEMMA 5.2. *Define* $v_i^g = \bar{v}_i^g - \underline{\epsilon} + \bar{\epsilon}/2$. *Then,*

1. *With* $\epsilon = \bar{\epsilon} - \underline{\epsilon}/2$, v_i^g *is a random variable with the support* $[-\epsilon, \epsilon]$ *and moreover the support* $[-\epsilon, \epsilon]$ *is tight for each* $i \in [1, 2, \ldots, n]$.
2. v_i^g *assumes both positive and negative signs with non-zero probability for each* $i \in [1, 2, \ldots, n]$.
3. v_i^g *and* v_{i+j}^g *are independent as long as* $j \geqslant N$.

Proof.

$$\max_i v_i^g = \max_i \bar{v}_i^g - \frac{\bar{\epsilon} + \underline{\epsilon}}{2} = \bar{\epsilon} - \frac{\bar{\epsilon} + \underline{\epsilon}}{2} = \epsilon,$$
$$\min_i v_i^g = \min_i \bar{v}_i^g - \frac{\bar{\epsilon} + \underline{\epsilon}}{2} = \underline{\epsilon} - \frac{\bar{\epsilon} + \underline{\epsilon}}{2} = -\epsilon.$$

Moreover, the tightness of the bound $[\underline{\epsilon}, \bar{\epsilon}]$ on $\bar{v}_i^g$ implies the tightness of the bound $[-\epsilon, \epsilon]$ on v_i^g. Since $-\epsilon < 0 < \epsilon$, v_i^g obviously assumes both positive and negative signs with non-zero probability. This completes the proofs of the first and the second parts. To show (3), notice that

$$v_i^g = \bar{v}_i^g - \frac{\underline{\epsilon} + \bar{\epsilon}}{2} = \sum_{k=0}^{N-1} \alpha_k \eta_{i-k} - \frac{\underline{\epsilon} + \bar{\epsilon}}{2},$$

$$v_{i+j}^g = \bar{v}_{i+j}^g - \frac{\underline{\epsilon} + \bar{\epsilon}}{2} = \sum_{k=0}^{N-1} \alpha_k \eta_{i+j-k} - \frac{\underline{\epsilon} + \bar{\epsilon}}{2}.$$

$\eta_i, \dots, \eta_{i-N+1}$ and $\eta_{i+j}, \dots, \eta_{i+j_N+1}$ are independent as long as $i+j-N+1 > i$ or $j \geqslant N$. Thus, v_i^g and v_{i+j}^g are independent provided $j \geqslant N$.

Now, the system (3.1) can be re-written as

$$\begin{aligned} y_i &= \bar{\phi}_i^T \bar{\theta} + \bar{v}_i^g + \bar{v}_i^b = \bar{\phi}_i^T \bar{\theta} + \bar{v}_i^g - \frac{\underline{\epsilon} + \bar{\epsilon}}{2} + \frac{\underline{\epsilon} + \bar{\epsilon}}{2} + \bar{v}_i^b \\ &= (\bar{\phi}_i^T, 1) \begin{pmatrix} \bar{\theta} \\ \bar{\theta}_{m+1} \end{pmatrix} + v_i^g + \bar{v}_i^b \\ &= \phi_i^T \theta + v_i^g + \bar{v}_i^b \end{aligned} \tag{5.1}$$

where $\bar{\theta}_{m+1} = \underline{\epsilon} + \bar{\epsilon}/2$, but unknown. v_i^g represents the 'good' noise part distributed in the interval $[-\epsilon, \epsilon]$ for some unknown $\epsilon > 0$ and $\bar{v}_i^b$ denotes the outliers. Note that if $\bar{v}_i^b$ were identically zero and ϵ were known, then the so-called membership set

$$\Omega_{I_n} = \bigcap_{i=1}^{n} \{\hat{\theta} \in R^{m+1} \ : \ -\epsilon \leqslant y_i - \phi_i^T \hat{\theta} \leqslant \epsilon\} \tag{5.2}$$

is the set of all possible parameter estimates that are consistent with the equation (5.1), the observed input-output data y_i, $\phi_i^T = (\bar{\phi}_i^T, 1)$ and the noise bound ϵ. The problem is that ϵ is unknown and outliers $\bar{v}_i^b$'s are not identically zero. Moreover, v_i^g and v_j^g could be dependent. To this end, note $I_n = [1, 2, \dots, n]$ and define

$$J_k = [i_1, i_2, \dots, i_k] \subset I_n.$$

Denote $|I_n| = n$ and $|J_k| = k$ the number of elements in the sets of I_n and J_k respectively. Also, let I_g be the (good) subset of I_n that $\bar{v}_i^b = 0$ if $i \in I_g$ and I_b be the (bad) subset of I_n that $\bar{v}_i^b \neq 0$ if $i \in I_b$. Clearly, under the system description

$$I_g \bigcap I_b = \emptyset, \ I_g \bigcup I_b = I_n$$

where $\emptyset$ indicates the empty set. Furthermore, the maximum occurrence of the outliers is bounded by k and thus, we have

$$|I_b| \leqslant k, \ |I_g| \geqslant n - k, \ |I_g| + |I_b| = n.$$

By removing k (not necessarily bad) points, say $J_k = [i_1, i_2, ..., i_k]$ from $I_n = [1, 2, \dots, n]$, what left in the good set is $I_g/(I_g \bigcap J_k)$ and in the bad set is $I_b/(I_b \bigcap J_k)$. Obviously,

$$|I_g/(I_g \bigcap J_k)| \geqslant n - k - k = n - 2k \ \text{ and } \ |I_b/(I_b \bigcap J_k)| \leqslant k.$$

By using the notation I_n/k as in (3.7), we obtain

$$I_n/J_k \in I_n/k.$$

For each subset $s \subset I_n$, the corresponding membership set can be defined

$$\Omega_s = \bigcap_{i \in s} \{\hat{\theta} \in R^{m+1} \ : \ -\epsilon \leqslant y_i - \phi_i^T \hat{\theta} \leqslant \epsilon\}. \tag{5.3}$$

It is important to remark that if $(I_n/J_k) \bigcap I_b = \emptyset$, i.e., none of the outlier happens in the set I_n/J_k, Ω_{I_n/J_k} is not empty and at least the true but unknown parameter vector θ belongs to the set Ω_{I_n/J_k}.

Now for each subset $s \subset I_n$, define the diameter of the membership set Ω_s

$$\text{dia } \Omega_s = \sup_{\theta_1, \theta_2 \in \Omega_s} \|\theta_1 - \theta_2\|_2.$$

We now state an assumption before presenting the key lemma.

ASSUMPTION 5.1.

- *The regressor ϕ_i is persistently exciting, i.e., there exists some $\alpha > 0$ and integer $l_0 > 0$ so that*

$$0 < \alpha^2 I \leqslant \frac{1}{l_0} \sum_{i=i_0+1}^{i_0+l_0} \phi_i \phi_i^T \tag{5.4}$$

  for all $i_0 \geqslant 0$.
- *Let $l = max(N, l_0)$. Then, for large n, the upper bound $k(n)$ on the outlier satisfies*

$$2k(l+1) \leqslant \xi n$$

  for some constant $0 \leqslant \xi < 1$.

- *The 'good' noise v_i^g is tightly bounded in the interval $[-\epsilon, \epsilon]$ for some unknown $\epsilon > 0$, i.e., there is a positive probability $p(\rho) > 0$ such that for any small enough $\rho > 0$*

$$\text{Prob}\{-\epsilon \leqslant v_i^g \leqslant -\epsilon + \rho\} \geqslant p(\rho) > 0 \tag{5.5}$$

and

$$\text{Prob}\{\epsilon - \rho \leqslant v_i^g \leqslant \epsilon\} \geqslant p(\rho) > 0 \tag{5.6}$$

for each $i \in (I_g/I_g \bigcap J_k)$, where J_k is an arbitrary subset of I_n containing k elements.

Assumption (5.4) is the sufficient richness condition on the input. In fact, (5.4) requires enough spectral context in the input. For details, see [3] and [14].

LEMMA 5.3. *Consider the system (5.1) under Assumption 5.1. Then,*

$$dia\ \Omega_{I_g/(I_g \cap J_k)} \to 0$$

with probability one as $n \to \infty$ and consequently $\Omega_{I_g/(I_g \cap J_k)} \to \{\theta\}$ and $\hat{\epsilon}(n) \to \epsilon$.

Proof. Note that $\theta \in \Omega_{I_g/(I_g \cap J_k)}$ for each n and

$$\begin{aligned} \text{dia}\ \Omega_{I_g/(I_g \cap J_k)} &= \sup_{\theta_1, \theta_2 \in \Omega_{I_g/(I_g \cap J_k)}} \|\theta_1 - \theta_2\|_2 \\ &\leqslant \sup_{\theta_1 \in \Omega_{I_g/(I_g \cap J_k)}} \|\theta_1 - \theta\|_2 + \sup_{\theta_2 \in \Omega_{I_g/(I_g \cap J_k)}} \|\theta_2 - \theta\|_2 \\ &= 2 \sup_{\hat{\theta} \in \Omega_{I_g/(I_g \cap J_k)}} \|\hat{\theta} - \theta\|_2. \end{aligned}$$

To show dia $\Omega_{I_g/(I_g \cap J_k)} \to 0$ with probability one, it suffices to show that for an arbitrary but fixed $\hat{\theta}(n)$, if $\hat{\theta}(n) \in \Omega_{I_g/(I_g \cap J_k)}$, then $\|\tilde{\theta}(n)\| = \|\hat{\theta}(n) - \theta\| \to 0$ as $n \to \infty$. Note $l = max(N, l_0)$, it follows that

$$n - 2k(1 + l) \geqslant n - \xi n \to \infty, \quad as\ \ n \to \infty.$$

Hence, for large n, let q be an integer such that

$$2ql \leqslant n - 2k \leqslant 2(q + 1)l.$$

By the hypothesis that ϕ_i is persistently exciting, thus within any window $[(i - 1)2l + 1, (i - 1)2l + l_0]$, there always exists at least one $i^0 \in [(i - 1)2l + 1, (i - 1)2l + l_0]$ such that

$$|\phi_{i^0}^T \tilde{\theta}(n)| \geqslant \alpha \|\tilde{\theta}(n)\|.$$

Let $i^0 \in [(i-1)2l+1, (i-1)2l+l_0]$, $i = 1, 2, \dots, q$ be a such time sequence. Now, $|J_k| = k(n)$ for each n, there are at most k windows $[(i-1)2l+1, (i-1)2l+l_0]$ that overlap with J_k. Therefore, the set $I_g/(I_g \cap J_k)$ contains at least $q-k$ windows that do not overlap with J_k. Let the corresponding time sub-sequence of i^0 be denoted by $i^0_1, i^0_2, \dots, i^0_{q-k}$. Suppose $\hat{\theta}(n) \in \Omega_{I_g/(I_g \cap J_k)}$, it follows that at each i^0_j,

$$\begin{aligned}\epsilon &\geqslant |y_{i^0_j} - \phi^T_{i^0_j}\hat{\theta}(n)| = |\phi^T_{i^0_j}\tilde{\theta}(n) + v^g_{i^0_j} + \bar{v}^b_{i^0_j}| \\ &= |\phi^T_{i^0_j}\tilde{\theta}(n) + v^g_{i^0_j}|\end{aligned} \tag{5.7}$$

because $i^0_j \notin I_b \Longrightarrow \bar{v}^b_{i^0_j} = 0$. Now, from the assumption that $v^g_{i^0_j}$ approaches the both bounds ϵ and $-\epsilon$ with a non-zero probability, it follows that with non-zero probability at each i^0_j,

$$\epsilon \geqslant |\phi^T_{i^0_j}\tilde{\theta}(n) + v^g_{i^0_j}| = |\phi^T_{i^0_j}\tilde{\theta}(n)| + |v^g_{i^0_j}| \geqslant \alpha\|\tilde{\theta}(n)\| + |v^g_{i^0_j}| \tag{5.8}$$

and

$$\epsilon - |v^g_{i^0_j}| > \rho.$$

In other words, at each i^0_j and any small $\rho >$, there exists a $p_3(\rho) > 0$ such that

$$\text{Prob}\left\{\|\tilde{\theta}(n)\| \leqslant \frac{1}{\alpha}\rho\right\} \geqslant p_3(\rho)$$

or

$$\text{Prob}\left\{\|\tilde{\theta}(n)\| > \frac{1}{\alpha}\rho\right\} \leqslant 1 - p_3(\rho).$$

Further v^g_i and v^g_j are independent as long as $|i-j| \geqslant l(\geqslant N)$ and consequently

$$\text{Prob}\left\{\|\tilde{\theta}(n)\| > \frac{1}{\alpha}\rho\right\} \leqslant (1 - p_3(\rho))^{q-k}.$$

As $2ql \leqslant n - 2k \leqslant 2(q+1)l$, we obtain as $n \to \infty$

$$\infty \leftarrow \frac{1}{2l}(n - 2k(l+1) - 2l) = \frac{n - 2l - 2k}{2l} - \frac{2kl}{2l} \leqslant q - k.$$

Clearly, for any $\rho > 0$,

$$\text{Prob}\left\{\|\tilde{\theta}(n)\| > \frac{1}{\alpha}\rho\right\} \to 0$$

as $n \to \infty$. Furthermore, for each $\rho > 0$,

$$q - k \geqslant \frac{1}{2l}(n - 2k(l+1)) - 1 \geqslant \frac{1-\xi}{2l}n - 1$$

and this implies

$$(1-p_3(\rho))^{q-k} \leqslant \frac{1}{1-p_3(\rho)}\left[(1-p_3(\rho))^{\frac{1-\xi}{2l}}\right]^n.$$

Therefore, for every small $\rho > 0$

$$\sum_{n=1}^{\infty} \mathrm{Prob}\{\|\tilde{\theta}(n)\| > \frac{1}{\alpha}\rho\} \leqslant \sum_{n=1}^{\infty} \frac{1}{1-p_3(\rho)}\left[(1-p_3(\rho))^{\frac{1-\xi}{2l}}\right]^n < \infty.$$

This implies by the Borel-Cantelli's Lemma that with probability one

$$\|\tilde{\theta}(n)\| \to 0$$

as $n \to \infty$. Accordingly,

$$\mathrm{dia}\ \Omega_{I_g/(I_g \cap J_k)} \leqslant 2\|\tilde{\theta}(n)\| \to 0$$

with probability one as $n \to \infty$. This completes the proof.

Now, we are in a position to show main convergence result.

THEOREM 5.1. *Consider the system (3.1) and its augmented system (3.6) under Assumption 5.1. Then, the estimate obtained by the optimization with few violated constraints (3.8) satisfies*

$$\hat{\theta}(n) \to \theta$$

with probability one as $n \to \infty$.

Proof.

$$-\hat{\epsilon}(n) \leqslant y_i - \phi_i^T \hat{\theta}(n) \leqslant \hat{\epsilon}(n)$$

for all $i \in I_n/J_k = (I_g/I_g \bigcap J_k) \bigcup (I_b/I_b \bigcap J_k)$. Let $\hat{\epsilon}^g$ and $\hat{\epsilon}^b$ be the minimum values such that

$$\hat{\epsilon}^g = \min_{\hat{\theta}} \max_{i \in I_g/I_g \bigcap J_k} |y_i - \phi_i^T \hat{\theta}|,$$

and

$$\hat{\epsilon}^b = \min_{\hat{\theta}} \max_{i \in I_b/I_b \bigcap J_k} |y_i - \phi_i^T \hat{\theta}|$$

respectively.

By the key lemma 5.3, $\hat{\epsilon}^g \to \epsilon$ with probability one as $n \to \infty$. Combining the fact that $\hat{\epsilon}^g \leqslant \hat{\epsilon}(n) \leqslant \epsilon$, we have that $\hat{\epsilon}(n) \to \epsilon$ with probability one.

Now, we show that $\hat{\theta}(n) \to \theta$. To this end, let Ω_{I_n/J_k} denote the membership set after removing k constraints, i.e., $\Omega_{I_n/J_k} = \Omega^g \bigcap \Omega^b$ with

$$\Omega^g = \left\{\hat{\theta} : -\hat{\epsilon}(n) \leqslant y_i - \phi_i^T \hat{\theta} \leqslant \hat{\epsilon}(n),\ i \in I_g/I_g \bigcap J_k\right\},$$

$$\Omega^b = \left\{\hat{\theta} : -\hat{\epsilon}(n) \leqslant y_i - \phi_i^T \hat{\theta} \leqslant \hat{\epsilon}(n),\ i \in I_b/I_b \bigcap J_k.\right\}.$$

Since $\hat{\theta}(n) \in \Omega_{I_n/J_k}$, we have $\hat{\theta}(n) \in \Omega^g$ and $\hat{\theta}(n) \in \Omega^b$. Now from the fact that the set Ω^g converges to a singleton $\{\theta\}$ with probability one, it follows that $\hat{\theta}(n) \to \theta$ with probability one. This completes the proof.

6. Concluding remarks

In this paper, an optimization with few violated constraints approach is proposed to parameter identification in the presence of outliers. The approach requires a minimum statistical knowledge of the unknown noise and is convergent. It is also important to note that no minimum phase condition is assumed on the filter which models the noise. Instead, what is needed for the proposed approach is the availability of the upper bound $k(n)$ on the number of outliers.

Another point worth mentioning is that the non-degeneracy condition is assumed for the complexity analysis. The non-degeneracy condition is standard in the linear programming literature. If the original problem is degenerate, the degeneracy can be removed by infinitesimal perturbation. The solution to the refinement problem also solves the original problem. Interested readers can find more details in [8].

Acknowledgement

This work was supported in part by the funds of NSF.

References

1. M. Basseville and A. Benveniste (1980), *Detection of Abrupt Changes in Signals and Dynamical Systems*, Springer, Berlin.
2. Bai, E. W., H. Cho, R.Tempo and Y. Ye (1999). Minimization with few violated constraints and its application in set-membership identification, *Proc of IFAC World Congress*, Beijing, China, Vol. H, 343–348, also to appear in *IEEE Trans. on AC*
3. Bai, E.W. and S. Sastry (1985), Persistency of excitation, sufficient richness and parameter convergence in discrete time adaptive control, *System and Control Letters*, 153–163.
4. R.D. Cook and S. Weisberg (1982), Residuals and Fluence in Regression, Chapman and Hall, NY.
5. Huber, P.J. (1973), Robust regression: asymptotics, conjectures and Monte Carlo, *Annals of Statistics* 1, 799–821.

6. Huber, P.J. (1981), Robust Statistics, Wiley, NY.
7. Ljung, L. (1987), System Identification: Theory for the User, Prentice-Hall, Englewood Cliffs, NJ.
8. Matousek, J. (1995), On geometric optimization with few violated constraints, *Discrete and Computational Geometry* 14, 365–384.
9. Ninness, B. and G. Goodwin (1995), Rapprochement between bounded error and stochastic estimation theory, *Int. J. of Adaptive Control and Signal Processing* 9, 107–132.
10. Neter, J., M. Kutner, C. Nachtsheim and W. Wasserman (1996), Applied Linear Regression Models, Irwin Inc., Chicago, IL
11. Poljak, B.T. and JA. Z. Tsypkin (1980), Robust identification, *Automatica* 53, 53–63.
12. Rousseeuw, P and A. Leroy (1987), Robust Regression and Outlier Detection, John Wiley, NY.
13. Sharir, M. and E. Welzl (1992), A combinatorial bound for linear programming and related problems, in *Lecture Notes in Computer Science* 577, 569–579, Springer, Berlin.
14. Soderstrom, S. and P. Stoica (1989), System Identification, Prentice-Hall, Englewood Cliffs, NJ.