Optimization With Few Violated Constraints for Linear Bounded Error Parameter Estimation

Er-Wei Bai, *Senior Member, IEEE*, Hyonyong Cho, Roberto Tempo, *Fellow, IEEE*, and Yinyu Ye

***Abstract*—In the context of linear constrained optimization, we study in this paper the problem of finding an optimal solution satisfying all but k of the given n constraints. A solution is obtained by means of an algorithm of the complexity $\min\{O(n\cdot k^d), O(n\cdot d^{k+1})\}$, where d is the dimension of the problem. We then use these results to solve the problem of robust identification in the presence of outliers in the setting of bounded error parameter identification. Finally, we show that the estimate obtained converges to the true but unknown parameter in the presence of outliers.**

***Index Terms*—Constrainted optimization, parameter estimation, system identification, unknown but bounded error.**

I. Introduction and Motivation

MANY engineering analysis and design problems boil down to finding the minimum of some function subject to a given constraint set. The solution of the minimization problem relies on the parameters that form the constraints. In many cases, however, the values of these parameters are known only to a certain degree due to imperfect knowledge of the system, the environment and the measurements. For instance, in system identification, the constraints depend on the measurement data. A few erroneous or highly disturbed measurements may have a substantial influence on the solution. Therefore, a solution based on the nominal values is often not what we are looking for. In some cases, the solution obtained may be unreliable and far off from the desired answer. We now take the bounded error parameter identification problem [1], [2], [5], [6], [11] as an example. Consider a single-input–single-output (SISO) discrete-time system

$$y_i = \phi_i^T \theta + v_i, \qquad i = 1, 2, \ldots, n \tag{1.1}$$

where $y_i \in R$ is the system output, $\phi_i \in R^m$ the measured regressor, $\theta \in R^m$ the unknown parameter vector to be identified, and $v_i \in R$ the noise. In this setting, the noise is assumed to be bounded by some constant $\epsilon \geq 0$ [1], [2], [5], [6], [11], i.e.,

$$|v_i| \leq \epsilon \tag{1.2}$$

Manuscript received January 25, 1998; revised October 29, 1999, September 19, 2000, and January 30, 2002. Recommended by Associate Editor S. Hara. This work was supported in part by the National Science Foundatio under Grants ECS-9710297 and ECS-0098181, and in part by IRITI-CNR of Italy.

E.-W. Bai and H. Cho are with the Department of Electrical and Computer Engineering, University of Iowa, Iowa City, Iowa 52242 USA (e-mail: erwei@icaen.uiowa.edu; hcho@icaen.uiowa.edu).

R. Tempo is with the IRITI-CNR, Politecnico di Torino, Italy (e-mail: tempo@polito.it).

Y. Ye is with the Department of Management Science, University of Iowa, Iowa City, Iowa 52242 USA (e-mail: yyye@dollar.biz.uiowa.edu).

Publisher Item Identifier 10.1109/TAC.2002.800644.

for $i = 1, 2, \ldots, n$. Then, the membership set

$$\Omega = \bigcap_{i=1}^{n} \left\{ \hat{\theta} \in R^m : -\epsilon \leq y_i - \phi_i^T \hat{\theta} \leq \epsilon \right\} \tag{1.3}$$

is the set of all parameters that are consistent with the system (1.1), the observed input–output data y_i, ϕ_i and the assumed noise bound (1.2). In other words, every $\hat{\theta} \in \Omega$ could generate the observed input–output data for some noise sequence belonging to (1.2) and thus is a valid estimate of θ. Intuitively, the quality of the identification may be measured in terms of the "size" of the uncertainty represented by the diameter of the membership set

$$\text{dia}\,\Omega = \sup_{\theta_1, \theta_2 \in \Omega} ||\theta_1 - \theta_2||_2. \tag{1.4}$$

The identification result is useful only if the diameter of Ω is small, i.e., the resulting uncertainty is small. Clearly, the diameter of Ω depends on the actual noise bound ϵ which is usually unknown and is often replaced by its estimate $\hat{\epsilon}$. If the estimate $\hat{\epsilon}$ is much larger than the actual bound ϵ, the membership set Ω is inevitably large and conservative. On the other hand, if the assumed $\hat{\epsilon}$ is smaller than the actual bound ϵ, the resulting membership set Ω may be empty. To illustrate the problem and a way to fix it, consider a scalar example of (1.1) with $\phi_i \equiv 0.1$, $v_i = (-1)^i$, $n = 100$ and θ being any constant. With $\epsilon = 1$, the membership set is given by

$$\begin{aligned}\Omega =& \{\hat{\theta} : -\epsilon \leq y_i - 0.1\hat{\theta} \leq \epsilon,\ i = 1, 2, \ldots, 100\}. \\ =& \left\{ \hat{\theta} : \frac{-\epsilon + \max_{1 \leq i \leq 100} v_i}{0.1} + \theta \leq \hat{\theta} \right. \\ & \left. \leq \frac{\epsilon + \min_{1 \leq i \leq 100} v_i}{0.1} + \theta \right\}.\end{aligned}$$

Since $\max v_i = 1$ and $\min v_i = -1$, the membership set becomes a singleton and

$$\Omega = \{\theta\} \quad \text{dia}\,\Omega = 0.$$

This implies that a perfect estimate $\hat{\theta} = \theta$ is obtained.

Now, suppose we have a single bad measurement, or outlier, at $i = 20$ resulting in $v_{20} = 3$. There are two cases. The first one still considers the noise bound $\epsilon = 1$. Because $\max v_i = 3$, we immediately have that the set

$$\Omega = \{\hat{\theta} : 20 + \theta \leq \hat{\theta} \text{ and } \hat{\theta} \leq \theta\}$$

is empty. In the second case, let $\epsilon = 3$ be the actual noise bound. Then

$$\begin{aligned}\Omega = &\left\{\hat{\theta}\colon \frac{-3+\max_{1\le i\le 100} v_i}{0.1} + \theta \le \hat{\theta} \right.\\ &\left. \le \frac{3+\min_{1\le i\le 100} v_i}{0.1} + \theta\right\}\\ =&\{\hat{\theta}\colon \theta \le \hat{\theta} \le 20+\theta\}.\end{aligned}$$

Clearly, dia $\Omega = 20$ results in a large uncertainty in the parameter estimation. In this simple example, we see that the membership set method is sensitive to the outliers. We also observe that increasing the number of measurements does not solve the problem. To make things worse, we notice that even an accurate noise bound $\hat{\epsilon} = 3$ would still give rise to a very conservative Ω in the presence of outliers. A way to reduce the effects of outliers is to detect and remove outliers. To this end, let us take a closer look at the previous example. Define the set $I_{100} = [1, 2, \ldots, 100]$ and let the set $I_{100}/1$ denote the collection of all subsets of I_{100} with 99 elements. Further, let $s \subset I_{100}/1$ be the subset of $I_{100}/1$ that does not contain $i = 20$ and $s^* \subset I_{100}/1$ be any subset of $I_{100}/1$ that does contain $i = 20$. Furthermore, let $\hat{\epsilon}^*$, $\hat{\theta}^*$ and $[i_1^*, i_2^*, \ldots, i_{99}^*] \subset I_{100}/1$ be the triplet that solves the following minimization problem:

$$\min_{[i_1,i_2,\ldots,i_{99}]\in I_{100}/1} \min_{\hat{\epsilon}} \hat{\epsilon}$$
$$\text{subject to } -\hat{\epsilon} \le y_i - \phi_i\hat{\theta} \le \hat{\epsilon}, \quad i \in [i_1, i_2, \ldots, i_{99}].$$

Define the set

$$\Omega^* = \left\{\hat{\theta}\colon -\hat{\epsilon}^* \le y_i - \phi_i\hat{\theta} \le \hat{\epsilon}^*,\ i \in [i_1^*, i_2^*, \ldots, i_{99}^*]\right\}.$$

Note that the set Ω^* is the membership set obtained by removing one measurement and setting the noise bound to $\hat{\epsilon}^*$. It can be easily verified that

$$\begin{aligned}\min_{\hat{\theta}} \max_{i\in s\in I_{100}/1} |y_i - \phi_i\hat{\theta}| &\le \max_{i\in s\in I_{100}/1} |y_i - \phi_i\theta| \le 1\\ &\le \min_{\hat{\theta}} \max_{i\in s^*\in I_{100}/1} |y_i - \phi_i\hat{\theta}|.\end{aligned}$$

Hence, we have $\hat{\epsilon}^* = 1$ and this implies

$$\begin{aligned}[i_1^*, i_2^*, \ldots, i_{99}^*] =&[1, \ldots, 19, 21, \ldots, 100] \text{ and}\\ \Omega^* =&\{\hat{\theta}\colon \theta \le \hat{\theta} \le \theta\} = \{\theta\}.\end{aligned}$$

Therefore $\hat{\theta}$ coincides with the true parameter θ. In other words, the effect of the outliers $v_{20} = 3$ at $i = 20$ is eliminated.

The idea behind the above procedure can be explained easily. The objective is to find a parameter $\hat{\theta}$ and a subset $\bar{s} \subset I_{100}/1$ such that $\max_{i\in\bar{s}} |y_i - \phi_i\hat{\theta}|$ is minimized. The intuition is that $\max_{i\in\bar{s}} |y_i - \phi_i\hat{\theta}|$ is "large" if some of the outliers are present in the set $\bar{s}$ and is "small" if no outlier is present. Therefore, by looking for the optimal $\hat{\epsilon}$ and $\hat{\theta}$ that satisfy all but a small number of constraints effectively removes the outliers from measurement data.

Having motivated the need of optimization with few violated constraints, we now summarize the goals and the results of this paper.

- The outliers have a substantial influence on the method of bounded error parameter estimation [1]–[3], [8], [11]. How to minimize the effects of the outliers in the bounded error parameter identification has been an open problem for a long time. We remark that there is large body of research on outliers in the setting of stochastic identification but only scattered work in the bounded error parameter estimation setting. One of the first works reported is the outlier minimal number estimator [8], [14]. In these papers, it is shown that the outlier minimal number estimator is optimal in terms of the breakdown point. To make the algorithm work, however, the noise bound ϵ needs to be known *a priori*. How to find a good noise bound, especially in the presence of outliers, was not discussed in these papers. Moreover, the complexity of the algorithm of the outlier minimal number estimator could be very high. Some methods were proposed recently to improve its efficiency [7].
- In this paper, we propose an optimization approach with few violated constraints to deal with outliers. It is shown that the proposed method minimizes the effects of the outliers. A result of this approach is to provide a noise bound estimate. In addition, under some mild technical assumptions, we prove that the estimate $\hat{\theta}$ converges to the true parameter vector θ even in the presence of outliers.
- It is shown in this paper that the complexity of the proposed algorithm to calculate an optimal solution that satisfies all n but k constraints is bounded by

$$\min\left\{O(n\cdot k^d),\ O(n\cdot d^{k+1})\right\}$$

where d is the dimension of the problem to be defined later. In particular, in the setting of bounded error parameter estimation, $d = m + 1$ where m is the dimension of the unknown parameter vector θ. This complexity is comparable to $O(n)$ for small d, $k \ll n$. We note that a trivial way to solve the problem of optimization with few violated constraints is to find the minimum value for each collection of $n - k$ constraints and this results in a complexity $O(n \cdot n!/k!(n-k)!)$ which is much higher than $\min\left\{O(n\cdot k^d),\ O(n\cdot d^{k+1})\right\}$ for small d, $k \ll n$. The results reported here are a continuation of the work of [10], which shows that the complexity of the problem of optimization with few violated constraints is bounded by $O(n\cdot(k+1)^d)$. In this paper, we therefore improve the complexity from $O(n\cdot(k+1)^d)$ to $\min\left(O(n\cdot k^d),\ O(n\cdot d^{k+1})\right)$. This improvement could be substantial, e.g., $d = 20$ and $k = 1$ give $(k+1)^d = 2^{20} = 1\,048\,576$ and $d^{k+1} = 20^2 = 400$.

Finally, we point out that the problem posed in this paper is reminiscent of the least quantile of squares estimate in the time series literature [15]. Thus, the method proposed in this paper also solves efficiently the least quantile of squares estimation problem for small d and k.

We now end this section by giving an outline of the paper. Section II proposes an approach for robust estimation in the bounded error parameter estimation setting. To efficiently solve

the problem, we re-formulate it in the framework of optimization with few violated constraints. Then, an algorithm is developed with the complexity discussed above. Convergence results are provided in Section III. Section IV shows some numerical simulation results. Concluding remarks are provided in Section V.

II. Optimization With Few Violated Constraints for Bounded Error Parameter Estimation

In this section, we study the problem of robust identification in the presence of outliers within the framework of optimization with few violated constraints. Let y_i and $\phi_i, i = 1, \ldots, n$ denote the given input–output measurements. Define a new variable

$$x^T = \left(\hat{\theta}^T, x_{m+1}\right) = (\hat{\theta}^T, \hat{\epsilon}) \tag{2.1}$$

and the constraints

$$\left\{x \in R^{m+1}; -\hat{\epsilon} \le y_i - \phi_i^T\hat{\theta} \le \hat{\epsilon}\right\} = \left\{x \in R^{m+1}: -x_{m+1} \le y_i - \phi_i^T\hat{\theta} \le x_{m+1}\right\}$$

for $i = 1, \ldots, n$. Since both constraints $-x_{m+1} \le y_i - \phi_i^T\hat{\theta} \le x_{m+1}$ appear in pair, the condition $\hat{\epsilon} = x_{m+1} \ge 0$ is guaranteed. Otherwise, if $\hat{\epsilon} = x_{m+1} < 0$, no constraint $\left\{x \in R^{m+1}: -x_{m+1} \le y_i - \phi_i^T\hat{\theta} \le x_{m+1}\right\}$ can be satisfied. Now, the problem of robust identification in the presence of outliers is to find the backward lexicographically smallest point $(\hat{\theta}^T, \hat{\epsilon})^T = x \in R^{m+1}$ such that all n but at most k constraints $-x_{m+1} \le y_i - \phi_i^T\hat{\theta} \le x_{m+1}$ are satisfied. Then, $\hat{\theta}$ is the parameter estimate and $\hat{\epsilon}$ is the noise bound estimate. The backward lexicographically smallest point, or the lexicographically smallest point for short in this paper, means that the last coordinate is the most important. In other words, let $x = (x_1, \ldots, x_{m+1})^T$, $y = (y_1, \ldots, y_{m+1})^T$ be any two points in R^{m+1}. Then, x is lexicographically smaller than y, or $x < y$ if and only if

$$x_j < y_j \text{ for some } j \le m+1$$

and

$$x_k = y_k \text{ for all } k \text{ such that } j < k \le m+1.$$

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

Next, we define the problem of optimization with few violated constraints in an abstract framework.

Let H denote the set of n constraints in R^{m+1}. For any subset $G \subseteq H$, let $|G|$ denote the number of constraints in G, for example $|H| = n$ and let w be a function which maps every subset $G \subseteq H$ to the minimum value of some function, i.e., the value of $w(G)$ stands for the smallest value attainable for a certain cost function satisfying all the constraints in G. One example is $w(G) = \min_{x \in X_G} f(x)$ for some function $f(x)$: $R^{m+1} \to R$, where

$$X_G = \{x \in R^{m+1}: \text{all constraints of } G \text{ are satisfied}\}.$$

We now introduce six definitions which are standard in the literature of LP-type problems, e.g., see [10] and [16].

Definition 2.1: 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 2.2: For given pair of the constraint set H and the function w, the maximum cardinality (number) of constraints in a basis is called the dimension denoted by $d(H, w)$ or d for short.

We remark that for a linear cost function with linear constraints, the dimension d is exactly equal to $m+1$, the dimension of the parameter vector to be calculated [10].

Definition 2.3: We say that a constraint $h \in H$ violates a set G if $w(G \bigcup h) > w(G)$. For $G \subseteq H$, we denote by $V(G)$ all the constraints of H violating G.

Definition 2.4: Let $G \subseteq H$, then the level of G is defined as $|V(G)|$, i.e., the number of constraints violating G.

Definition 2.5: The minimization problem (H, w) is called an LP-type problem if the following condition is satisfied.

Condition 2.1: For all $F, G \subseteq H$

$$\left.\begin{array}{c} F \subseteq G \subseteq H \\ w(F) = w(G) \\ h \in H \end{array}\right\} \Longrightarrow \left\{\begin{array}{c} w(F \bigcup h) > w(F) \\ \Longleftrightarrow \\ w(G \bigcup h) > w(G) \end{array}\right. .$$

Definition 2.6: An LP-type problem is nondegenerate if $w(B) \neq w(B')$ for any two distinct bases $B \neq B'$ in H.

We now formally define the problem of optimization with few violated constraints.

Problem of Optimization With Few Violated Constraints (OWFVC): Consider a LP-type problem (H, w). For a given $0 \le k \le n$, find a basis that has the minimum value w and satisfies all n but at most k constraints.

We denote by $\mathcal{B}_k$ the set of all bases of level k, i.e., the collection of all the bases representing the sets of level k, and $\mathcal{B}_{\le k}$ for the set of bases of level at most k. In order to solve the OWFVC problem, we need to find 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 equal to or less than k in an efficient way when $d, k \ll n$ and then to select the one with the minimum value. The key is to show that the number of bases for $\mathcal{B}_{\le k}$ grows no faster than $\min\left\{O(k^d), O(d^{k+1})\right\}$ as k increases and moreover every basis for $\mathcal{B}_{j+1}$ can be generated from $\mathcal{B}_j$ in the sense defined in the following lemma.

Lemma 2.1: Consider a nondegenerate and feasible LP-type problem. Then, we have

1) Every basis for $\mathcal{B}_{j+1}$ can be generated from $\mathcal{B}_j$ in the sense that for each $B' \in \mathcal{B}_{j+1}$, there exists a basis $\{b_1, b_2, \ldots, b_q\} = B_1 \in \mathcal{B}_j$, $q \le d$ such that $B' = B(H/(V(B_1) \bigcup b_i))$, i.e., B' is a basis of $H/(V(B_1) \bigcup b_i)$ for some b_i, $1 \le i \le q$, where the sign $/$ means "deprived of."
2) Every basis of $\mathcal{B}_k$ can be reached from the basis $B(H)$ of H by a direct path in the sense that $\mathcal{B}_1$ can be generated from $\mathcal{B}_0$, which is $B(H)$, $\mathcal{B}_2$ can be generated from $\mathcal{B}_1, \ldots$, and $\mathcal{B}_k$ can be generated from $\mathcal{B}_{k-1}$.

Proof: First, we observe that 2) is a direct consequence of 1), and thus, we only need to show 1). Let $B' \in \mathcal{B}_{j+1}$. Write $G = H/V(B')$, then $B' = B(G) = B(H/V(B'))$. For every $h \in V(B')$, consider the value $w(G \bigcup h)$ and let $h_1 \in V(B')$ be an element giving the smallest of these values. In fact, such h_1 is unique. To see that, let $h_2 \in V(B')$ and $h_2 \neq h_1$ such that

$$w\left(G \bigcup h_1\right) = w\left(G \bigcup h_2\right).$$

Let $B_1 = B(G \bigcup h_1)$ and $B_2 = B(G \bigcup h_2)$. Since $h_1, h_2 \in V(B')$, we have $h_1 \in B_1$ and $h_2 \in B_2$. However, this implies that $B_1 \neq B_2$ and

$$w(B_1) = w\left(G \bigcup h_1\right) = w\left(G \bigcup h_2\right) = w(B_2)$$

a contradiction to the nondegeneracy assumption. Next, we show that

$$V(B_1) \bigcup h_1 = V(B').$$

Note that $G \subseteq G \bigcup h_1$, $B_1 = B(G \bigcup h_1)$ and $B' = B(G)$

$$\Longrightarrow V(B_1) \subseteq V(B').$$

Also, $h_1 \in V(B') \Longrightarrow V(B_1) \bigcup h_1 \subseteq V(B')$. We now need to show equality. To this end, suppose there exists some $h \neq h_1$ and $h \in V(B')$ but $h \notin V(B_1)$

$$\begin{aligned} \Longrightarrow w(B_1) &= w\left(B_1 \bigcup h\right) \\ \Longrightarrow w\left(G \bigcup h_1\right) &= w\left(G \bigcup h_1 \bigcup h\right). \end{aligned}$$

This contradicts $w(G \bigcup h) > w(G \bigcup h_1)$ for any $h \in V(B')$ by the choice and the uniqueness of $h_1 \in V(B')$. Accordingly

$$V(B_1) \bigcup h_1 = V(B'). \tag{2.2}$$

Combining the fact that $h_1 \in B_1 = \{b_1, b_2, \ldots, b_l\}$, $l \leq d$, it follows that:

$$\begin{aligned} B' &= B\left(\frac{H}{V(B')}\right) = B\left(\frac{H}{(V(B_1) \bigcup h_1)}\right) \\ &= B\left(\frac{H}{(V(B_1) \bigcup b_i)}\right) \end{aligned}$$

for some $b_i \in B_1$. Now, we have to show that $B_1 \in \mathcal{B}_j$, i.e., B_1 is a basis of level j. To this end, it is easy to see from (2.2) that $V(B')$ has exactly $j+1$ elements and $h_1 \notin V(B_1)$. This implies that $V(B_1)$ has exactly j elements or $B_1 \in \mathcal{B}_j$. This completes the proof.

Lemma 2.2: Consider a nondegenerate and feasible LP-type problem with $d > 1$. Then

$$\begin{aligned} |\mathcal{B}_0| &= 1 \quad |\mathcal{B}_{\leq 1}| \leq 1 + d, \text{ and} \\ |\mathcal{B}_{\leq k}| &\leq \min\left\{O(k^d),\, O(d^{k+1})\right\} \end{aligned} \tag{2.3}$$

for any $k \geq 2$, where $|\mathcal{B}_i|$ and $|\mathcal{B}_{\leq i}|$ denote the number of bases in $\mathcal{B}_i$ and $\mathcal{B}_{\leq i}$, respectively.

Proof: For $k = 0$, there is only one basis $B(H)$ with maximum d elements by the nondegeneracy assumption. When $k = 1$, from Lemma (2.1), every basis at level $k = 1$ has a predecessor in $B(H)$ which implies that there are at most d bases at level $k = 1$. Thus, $|\mathcal{B}_{\leq 1}| \leq 1 + d$. We now show the case $k \geq 2$. First, we show that $|\mathcal{B}_{\leq k}| \leq O(d^{k+1})$. In light of Lemma 2.1, there are at most d bases at level $k = 1$ and each basis has at most d elements. Similarly, at level $k = 2$, there are at most d^2 bases and at level $k = 3$, there are at most d^3 bases. By continuing this argument, we obtain

$$|\mathcal{B}_{\leq k}| \leq 1 + d + d^2 + d^3 + \cdots + d^k = \frac{d^{k+1} - 1}{d - 1} = O(d^{k+1}).$$

We now show for $k \geq 2$

$$|\mathcal{B}_{\leq k}| \leq O((k+1)^d).$$

Note $n = |H|$ and draw a sample $S \subseteq H$ by independently picking each element of H with probability $p = 1/k$ into S and probability $1 - p = 1 - 1/k$ not into S. Any given basis B in H is a basis of S if and only if

$$B \subseteq S \text{ and } S \bigcap V(B) = \emptyset$$

where $\emptyset$ denotes the empty set. In turn, this implies that the probability of B becomes a basis of S is

$$\text{Prob}\{B \text{ is a basis of } S\} = p^{|B|}(1-p)^{|V(B)|}.$$

Now, suppose there are N bases $B_1, B_2, \ldots, B_N$ of level $\leq k$, i.e., $|\mathcal{B}_{\leq k}| = N$. The expected number ρ of these bases which become a basis of S is given by

$$\begin{aligned} \rho =& N \cdot \text{Prob}\{\text{all } N \text{ bases are a basis of } S\} + (N-1) \\ & \cdot \text{Prob}\{N-1 \text{ bases are a basis of } S\} + \cdots + 1 \\ & \cdot \text{Prob}\{\text{one of the bases is a basis of } S\} \\ \geq & 1 \cdot \text{Prob}\{\text{one of the bases is a basis of } S\}. \end{aligned} \tag{2.4}$$

Observe that

$$\begin{aligned} &\text{Prob}\{\text{one of the bases is a basis of } S\} \\ &\quad = \text{Prob}\{B_1 \text{ is a basis of } S\} + \cdots \\ &\qquad + \text{Prob}\{B_N \text{ is a basis of } S\} \\ &\qquad - \text{Prob}\{\text{more than one of } B_i \text{ are a basis of } S\}. \end{aligned}$$

By nondegeneracy assumption, S can only have one basis

$$\Longrightarrow \text{Prob}\{\text{more than one of } B_i \text{ are a basis of } S\} = 0.$$

Thus, it follows from (2.4) that:

$$\begin{aligned} \rho \geq & \text{Prob}\{B_1 \text{ is a basis of } S\} + \cdots \\ & + \text{Prob}\{B_N \text{ is a basis of } S\} \\ = & p^{|B_1|}(1-p)^{|V(B_1)|} + p^{|B_2|}(1-p)^{|V(B_2)|} \\ & + \cdots + p^{|B_N|}(1-p)^{|V(B_N)|}. \end{aligned}$$

Because the dimension is d and B_i's are the bases for the level $\leq k$

$$\begin{aligned} |B_i| \leq & d \text{ and } |V(B_i)| \leq k \\ \Longrightarrow \rho \geq & N \cdot p^d (1-p)^k. \end{aligned}$$

On the other hand, S has only one basis

$$\rho = 1 \cdot \text{Prob}\{\text{one of the bases is a basis of } S\} + 0$$

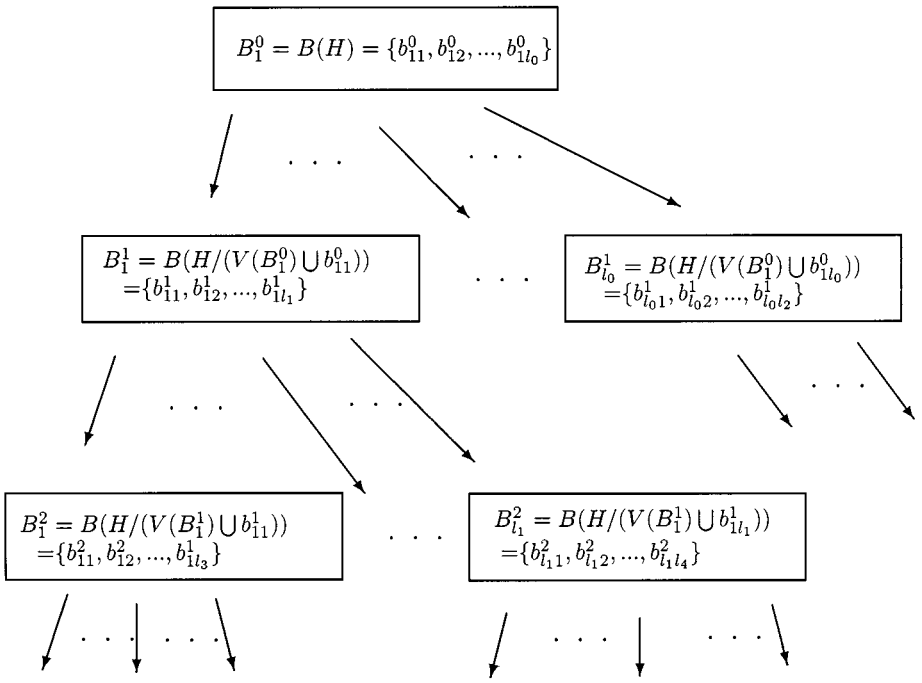

Fig. 1. The structure of the OWFVC algorithm.

$$\cdot \text{Prob}\{\text{none of } B_i \text{ is a basis of } S\} \le 1$$

$$\Longrightarrow 1 \ge Np^d(1-p)^k \text{ or}$$

$$N \le p^{-d}(1-p)^{-k} = k^d \left(\frac{1}{1-\frac{1}{k}}\right)^k \le 4k^d = O(k^d).$$

Thus,

$$|\mathcal{B}_{\le k}| = N \le \min(O(k^d), O(d^{k+1})).$$

This completes the proof.

Based on the aforementioned two lemmas, we now present the algorithm which is illustrated in Fig. 1 with some l_0, l_1, l_2, l_3 and $l_4 \le d$, to solve the problem of optimization with few violated constraints. The idea is to define a direct graph on the vertex set $\mathcal{B}_{\le k}$, i.e., to find all the bases of $\mathcal{B}_{j+1}$ from $\mathcal{B}_j$ and then to calculate each value w that requires solving $|B| \le d$ LP-type problem of the form (G, w).

Algorithm for the OWFVC problem

Let (H, w) be a nondegenerate and feasible LP-type problem. Given $n = |H|$, $d > 1$ and $k \ge 0$, find a basis that has the minimum value w and satisfies all n but at most $0 \le k \le n$ 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 $H/(V(B) \bigcup b)$ for every basis $B \in \mathcal{B}_{j-1}$ and each $b \in B$. Check if the achieved basis $B(H/(V(B) \bigcup b))$ coincides with any basis obtained before. If it is, then it is redundant and we remove it from the search path. Also, check if $V(B') = V(B) \bigcup \{b\}$, i.e., if the obtained basis 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 $\mathcal{B}_{\le k}$.

Before presenting the main result of this section, we remark that there are several algorithms for solving an LP-type problem with $k = 0$ and a fixed d in time $O(n)$ [16]. These algorithms differ slightly in the assumptions on the primitive operations available.

We now state the main result of this section.

Theorem 2.1: Let the LP-type problem (H, w) be nondegenerate and feasible. Let $d > 1$ be fixed. Then, the OWFVC problem can be solved by the previous algorithm in $O(n)$ time for $0 \le k \le 1$ and in $\min\{O(n \cdot k^d), O(n \cdot d^{k+1})\}$ time for any $k \ge 2$.

Proof: From the proofs of Lemmas 2.1 and 2.2, it is clear that the problem can be solved in time $O(d \cdot n) = O(n)$ for $0 \le k \le 1$ and in time $O(n \cdot d^{k+1})$ for $k \ge 2$. We now need to show that for $k \ge 2$, $O(n \cdot k^d)$ is also an upper bound. To this end, notice that there are at most k^d bases in $\mathcal{B}_{\le k}$. What we have to show is that the computation time for calculating redundant bases that coincide with previously obtained bases is linearly bounded by nk^d. Let B' be a redundant basis at level i that coincides with some bases obtained before. Then, its predecessor has to be a basis at level $i-1$ and is not redundant. The total number of such nonredundant predecessors is at most k^d and each has at most d successors. Thus, the maximum number

of redundant bases that need to be calculated is bounded by dk^d with calculation time $O(d\cdot n\cdot k^d)=O(n\cdot k^d)$. This completes the proof.

Remark 2.1: Theorem 2.1 shows that the proposed algorithm is very efficient for small d, $k \ll n$ which is the case in the bounded error parameter identification where n can be very large, k is the bound on the number of outliers and $d=m+1$, the number of the parameters.

Remark 2.2: Regarding the assumption that the problem is nondegenerate, we note that if the original problem is degenerate, then by using infinitesimal perturbation, a nondegenerate refinement can be formed. The solution of the refinement problem also solves the original problem. Interested readers may find more details in [10] and [13].

Remark 2.3: With respect to feasibility, we remark that the optimization problem is feasible if at least one solution exist, i.e., if the set defined by the constraint set H is not empty. In bounded error parameter estimation, the set

$$\begin{aligned} H &= \left\{x\in R^{m+1}; -\hat{\epsilon}\le y_i-\phi_i^T\hat{\theta}\le\hat{\epsilon}\right\}_1^n \\ &= \left\{x\in R^{m+1}: -x_{m+1}\le y_i-\phi_i^T\hat{\theta}\le x_{m+1}\right\}_1^n \end{aligned}$$

is always nonempty for $\hat{\epsilon}\ge\max|v_i|$. Therefore, the feasibility assumption is automatically satisfied as long as the noise v_i is bounded.

III. Robust Identification in the Presence of Outliers: Convergence Results

Consider the system (1.1) and let the noise be

$$v_i=v_i^g+v_i^b,\qquad i=1,2,\ldots,n \tag{3.1}$$

where v_i^g denotes the "good" disturbance and v_i^b the "bad" disturbance or outliers. Now, let $I_n=\{1,2,\ldots,n\}$ denote the set of time indices and $I_{i_k}=\{i_1,i_2,\ldots,i_k\}$ any subset of I_n containing k elements. Note that $|I_n|=n$ and $|I_{i_k}|=k$. Further, let I_g be the (good) subset of I_n such that $v_i^b=0$ if $i\in I_g$ and I_b be the (bad) subset of I_n such that $v_i^b\neq 0$ if $i\in I_b$.

We now state an assumption which is used in this section.

Assumption 3.1:

- The regressor ϕ_i is independent of v_i and is persistently exciting, i.e., there exists a positive integer l such that

$$0<\alpha I\le\frac{1}{l}\sum_{i=i_o+1}^{i_o+l}\phi_i\phi_i^T \tag{3.2}$$

  for all $i_o\ge 0$ and some $\alpha>0$.
- The "bad" noise v_i^b can be nonzero at most $k(n)>0$ times. Further $0<n-k(n)(l+1)\to\infty$ as $n\to\infty$.
- The "good" noise v_i^g is a sequence of independent random variables with some *unknown* distributions 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 $\rho>0$

$$\text{Prob}\{-\epsilon\le v_i^g\le-\epsilon+\rho\}\ge p(\rho)>0 \tag{3.3}$$

and

$$\text{Prob}\{\epsilon-\rho\le v_i^g\le\epsilon\}\ge p(\rho)>0 \tag{3.4}$$

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

Clearly, using Assumption 3.1, $|I_g|+|I_b|=n$ and

$$|I_b|\le k,\quad |I_g|\ge n-k.$$

We now make a few remarks regarding Assumption 3.1. Equation (3.2) is the standard condition of persistent excitation. Equations (3.3) and (3.4) regarding the tightness of the noise bound have been already used in bounded error parameter estimation context [4], [17]. Both tightness and persistent excitation conditions are required to establish the convergence result even in the absence of outliers [4], [17]. The essence is that the good noise is tightly distributed in an interval $[-\epsilon,\epsilon]$ and the outliers can happen at any time with any magnitude, but not very frequently. In fact, the maximum occurrence is bounded by $k(n)$. The tightness assumptions on the good part of the noise is only needed to establish convergence results but the algorithm developed in this paper can of course be used when the assumption is not satisfied.

In general, it is very difficult to define outliers or bad data that depend on the assumed system structure as well as on the assumptions of the unknown noise. In this paper, we assume that the system structure is linear and known, and bad data is due to measurement error. The upper bound k on the outliers is also important. In reality, the exact number of outliers is unknown. In many applications, it is reasonable to assume that the number of bad data does not exceed a certain percentage of the total data points n. For instance, if the bad data does not exceed 1% of the total data, then $k=0.1n$ is obtained.

As discussed in the Introduction, a robust way to estimate the noise bound and the unknown parameter in the presence of outliers is to find an optimal pair that satisfy all n but a small number $k(n)$ of observed input–output data. Thus, the problem of robust identification in the presence of outliers can be formally stated as follows. *Consider the system* (1.1) *and Assumption 3.1. Find the minimum* $\hat{\epsilon}>0$ *and the corresponding* $\hat{\theta}$ *so that all but at most* $k(n)$ *constraints* $\left\{-\hat{\epsilon}\le y_i-\phi_i^T\hat{\theta}\le\hat{\epsilon}\right\}$ *are satisfied.*

In order to solve the robust identification problem, observe that by removing at most k (not necessarily bad) constraints, say at time $I_{i_k}=\{i_1,i_2,\ldots,i_k\}$, what remains in the good set is $I_g/(I_g\bigcap I_{i_k})$ and in the bad set is $I_b/(I_b\bigcap I_{i_k})$. Obviously

$$\left|I_g/(I_g\bigcap I_{i_k})\right|\ge n-k-k=n-2k\quad \left|I_b/(I_b\bigcap I_{i_k})\right|\le k.$$

Next, let $(\hat{\epsilon},\hat{\theta},\hat{I}_{i_k})$ be any triplet with $\hat{I}_{i_k}\subset I_n$ and $|\hat{I}_{i_k}|\le k(n)$ so that

$$-\hat{\epsilon}\le y_i-\phi_i^T\hat{\theta}\le\hat{\epsilon},\quad i\in I_n/\hat{I}_{i_k}. \tag{3.5}$$

In other words, $\hat{\epsilon}$ and $\hat{\theta}$ are the estimates of ϵ and θ, respectively, that satisfy all n but at most $k(n)$ constraints. Let $(\hat{\epsilon}_n,\hat{\theta}_n,I_{i_k})$ be any such triplet that achieves the minimum value of $\hat{\epsilon}$, i.e.,

$$0<\hat{\epsilon}_n\le\hat{\epsilon}$$

for all possible $\hat{\epsilon}$ satisfying (3.5).

Now, let

$$x^T = (\hat{\theta}^T, \hat{\epsilon})$$
$$H = \left\{ x \in R^{m+1} : -x_{m+1} \le y_i - \phi_i^T \hat{\theta} \le x_{m+1} \right\}_1^n$$

denote the augmented variable and the constraint set respectively. Further, for each subset $G \subset H$, let $w(G)$ denote the lexicographically smallest point in R^{m+1} satisfying all the constraints in G. Then, the robust identification problem is exactly the OWFVC problem and the algorithm can be summarized.

Robust Identification Algorithm in the Presence of Outliers:

Consider the system (1.1) under Assumption 3.1.

Step 1) Collect data, and define x and the constraint set H as in (2.1).

Step 2) Apply the OWFVC algorithm to find a triplet $(\hat{\epsilon}_n, \hat{\theta}_n, I_{i_k})$ that achieves the minimum value of $\hat{\epsilon}$, i.e., $0 < \hat{\epsilon}_n \le \hat{\epsilon}$ for all possible $\hat{\epsilon}$ satisfying

$$-\hat{\epsilon} \le y_i - \phi_i^T \hat{\theta}_n \le \hat{\epsilon}, \quad i \in I_n / \hat{I}_{i_k}.$$

$\hat{\epsilon}_n$ and $\hat{\theta}_n$ are the estimates of ϵ and θ.

We now present a convergence result showing the robustness of the previous algorithm.

Theorem 3.1: Let $\hat{\epsilon}_n$ and $\hat{\theta}_n$ be obtained by applying the above algorithm. Suppose the conditions of Theorem 2.1 and Assumption 3.1 are satisfied. Then, $d = m + 1$ and moreover, as $n \to \infty$, we have

$$\hat{\epsilon}_n \to \epsilon, \quad \text{dia}\, \Omega \to 0, \text{ and } \hat{\theta}_n \to \theta$$

with probability one.

Proof: $d = m + 1$ is obvious. Now

$$-\hat{\epsilon}_n \le y_i - \phi_i^T \hat{\theta}_n \le \hat{\epsilon}_n$$

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

$$\hat{\epsilon}^g = \min_{\hat{\theta}} \max_{i \in I_g / (I_g \bigcap I_{i_k})} \left| y_i - \phi_i^T \hat{\theta} \right|$$

and

$$\hat{\epsilon}^b = \min_{\hat{\theta}} \max_{i \in I_b / (I_b \bigcap I_{i_k})} \left| y_i - \phi_i^T \hat{\theta} \right|$$

respectively.

By Lemma 6.1, $\hat{\epsilon}^g \to \epsilon$ with probability one as $n \to \infty$. Combining this with the fact that $\hat{\epsilon}^g \le \hat{\epsilon}_n \le \epsilon$, we have that $\hat{\epsilon}_n \to \epsilon$ with probability one.

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

$$\Omega^g = \left\{ \hat{\theta} : -\hat{\epsilon}_n \le y_i - \phi_i^T \hat{\theta} \le \hat{\epsilon}_n,\ i \in I_g / (I_g \bigcap I_{i_k}) \right\}$$
$$\Omega^b = \left\{ \hat{\theta} : -\hat{\epsilon}_n \le y_i - \phi_i^T \hat{\theta} \le \hat{\epsilon}_n,\ i \in I_b / (I_b \bigcap I_{i_k}) \right\}.$$

Since $\hat{\theta}_n \in \Omega_{I_n / I_{i_k}}$, 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.

Theorem 3.1 says that an accurate estimate $\hat{\theta}_n$ can be robustly obtained in the presence of outliers. We remark that the only condition on outliers is that its occurrence is bounded by $k(n)$.

IV. Discussion and Simulation Results

In this section, we provide simulation results and some discussion on the implementation of robust bounded error parameter estimation by means of the OWFVC algorithm developed in Section II. We first make some remarks concerning the OWFVC algorithm.

Remark 4.1: 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 $(m + 1)$ constraints $\{-x_{m+1} \le y_i - \phi_i \hat{\theta} \le x_{m+1}\}$, $i = i_1, \ldots, i_{m+1}$ which intersect at a point that has the lexicographically smallest value. These $(m+1)$ constraints constitute a basis for H. 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 $\{-\hat{\epsilon} \le y_i - \phi_i \hat{\theta} \le \hat{\epsilon}\}$ are satisfied. With $x^T = (\hat{\theta}^T, \hat{\epsilon})$, this is clearly a linear programming problem. Denote such a basis by $B(H) = \{b_1, \ldots, b_{m+1}\}$, with $b_i \in H$. Next, we can calculate the bases at $k = 1$ level, i.e., find bases for $H / (V(B) \bigcup b_i)$, $i = 1, 2, \ldots, m + 1$, where $V(B)$ are the constraints in H violating $B(H)$. This is again a linear programming problem. In this sense, the OWFVC 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 a reason why we can achieve a low complexity for the OWFVC algorithm stated in Section II.

Remark 4.2: The reason why probabilistic assumptions on the noise in the bounded error parameter estimation setting are used is to extend the well defined notion of convergence. In fact, probabilistic assumptions on the noise are not necessary. The critical point is that the noise visits the bound $\pm\epsilon$ "often." With this "often" assumption, convergence is guaranteed in a deterministic sense.

We now simulate a fourth-order FIR system

$$y_i = (u_i, u_{i-1}, u_{i-2}, u_{i-3})\theta + v_i = \phi_i^T \theta + v_i$$

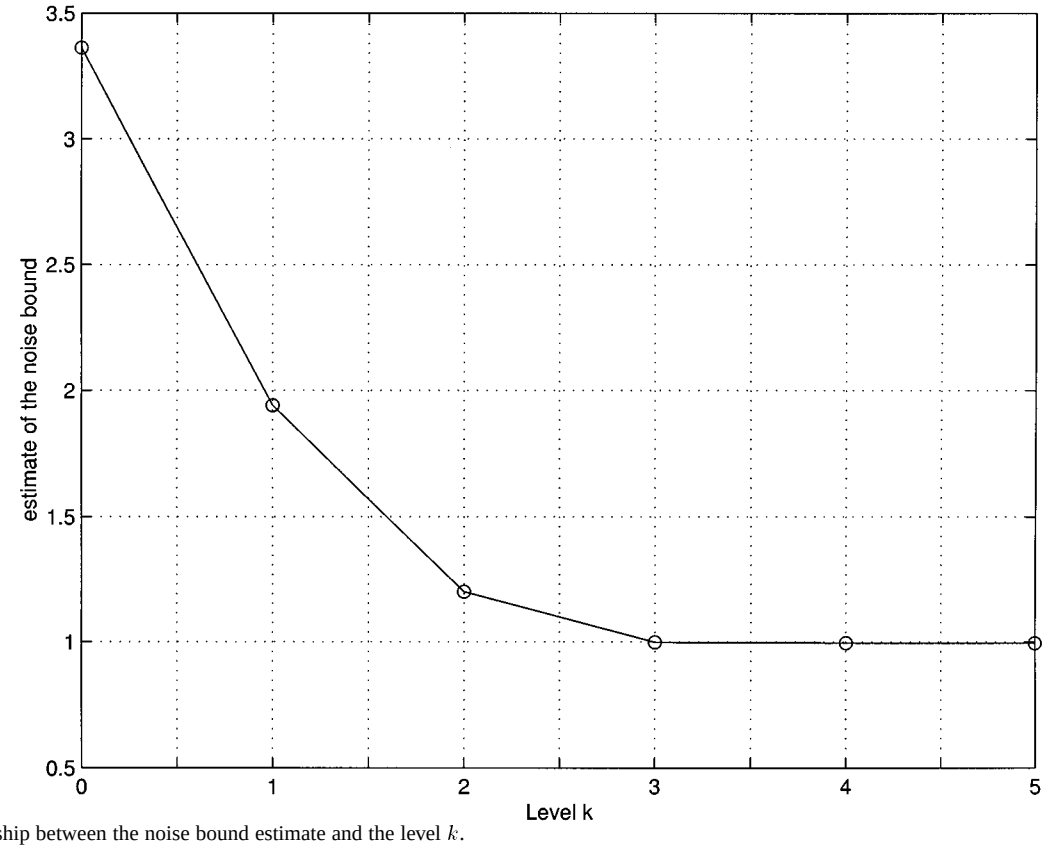

Fig. 2. Relationship between the noise bound estimate and the level k.

where the true parameter vector $\theta = (1, 0.8, -0.5, 1.2)^T$ and u_i is an i.i.d. random variable uniformly distributed in $[-1, 1]$, $i = 1, \ldots, 600$. The noise v_i is also an i.i.d. random variable in $[-1, 1]$. For simulation purpose, we added three outliers to the noise data

$$v_{50} = 3 \quad v_{250} = 1.5 \quad v_{450} = 5.$$

Clearly,three outliers account for 0.5% of the total data.

Note that the actual but unknown error bound is $\epsilon = 1$. Fig. 2 shows that an accurate estimate $\hat{\epsilon}_n$ can be obtained by allowing three constraints be violated. In real applications, the actual number of the outliers $k = 3$ may be unknown and can be estimated from the observed data. From Fig. 2, we see that there are large changes for $k = 0, 1, 2, 3$ and virtually no changes for $k = 3, 4, 5$. Therefore, we may conclude that $k = 3$ is the estimate of the upper bound on the number of outliers. Fig. 3 shows the parameter estimation error $||\hat{\theta}_n - \theta||_2 = ||\hat{\theta}_n - (1, 0.8, -0.5, 1.2)^T||_2$ and the noise bound estimation error $|\hat{\epsilon}_n - \epsilon| = |\hat{\epsilon}_n - 1|$ for the violation level $k = 0, 1, 2, 3, 4$. We see from the figure that when the level k increases, the estimation errors decrease. In particular, if $k \geq 3$, which is the actual number of outliers, the estimates are almost identical to the unknown parameters $(1, 0.8, -0.5, 1.2)^T$ and the unknown noise bound is equal to one. We conclude that the effect of outliers is efficiently eliminated as expected.

V. Concluding Remarks

In any identification setting, it is necessary to protect the estimates from "bad" data or outliers. This is usually done by changing the identification setting, i.e., changing the identification criteria depending on a priori knowledge of the noise. In this paper, in the bounded error parameter identification context, it is shown that a robust estimate can be obtained without modifying the identification setting. The method was developed within the framework of optimization with few violated constraints. We believe that the results derived in this paper are not limited to system identification but also applicable to many other applications in engineering analysis and design.

Appendix

The following lemma is needed to prove Theorem 3.1.

A. Lemma 6.1

Consider the system (1.1) with Assumption (3.1). Let

$$\hat{\theta} = \arg\min_{\bar{\theta}} \max_{i \in (I_g / (I_g \bigcap I_{i_k}))} \left| y_i - \phi_i^T \bar{\theta} \right|$$

where $I_{i_k} = \{i_1, \ldots, i_k\}$ is an arbitrary subset of I_n containing k elements. Define

$$\hat{\epsilon} = \max_{i \in (I_g / (I_g \bigcap I_{i_k}))} \left| y_i - \phi_i^T \hat{\theta} \right| \text{ and } \Omega_{I_g / (I_g \bigcap I_{i_k})} = \bigcap_{i \in (I_g / (I_g \bigcap I_{i_k}))} \left\{ \hat{\theta} : -\epsilon \leq y_i - \phi_i^T \hat{\theta} \leq \epsilon \right\}.$$

Then, if ϕ_i is persistently exciting as defined in (3.2), we have

$$\hat{\epsilon} \to \epsilon, \quad \text{dia}\, \Omega_{I_g / (I_g \bigcap I_{i_k})} \to 0 \quad \Omega_{I_g / (I_g \bigcap I_{i_k})} \to \{\theta\}$$

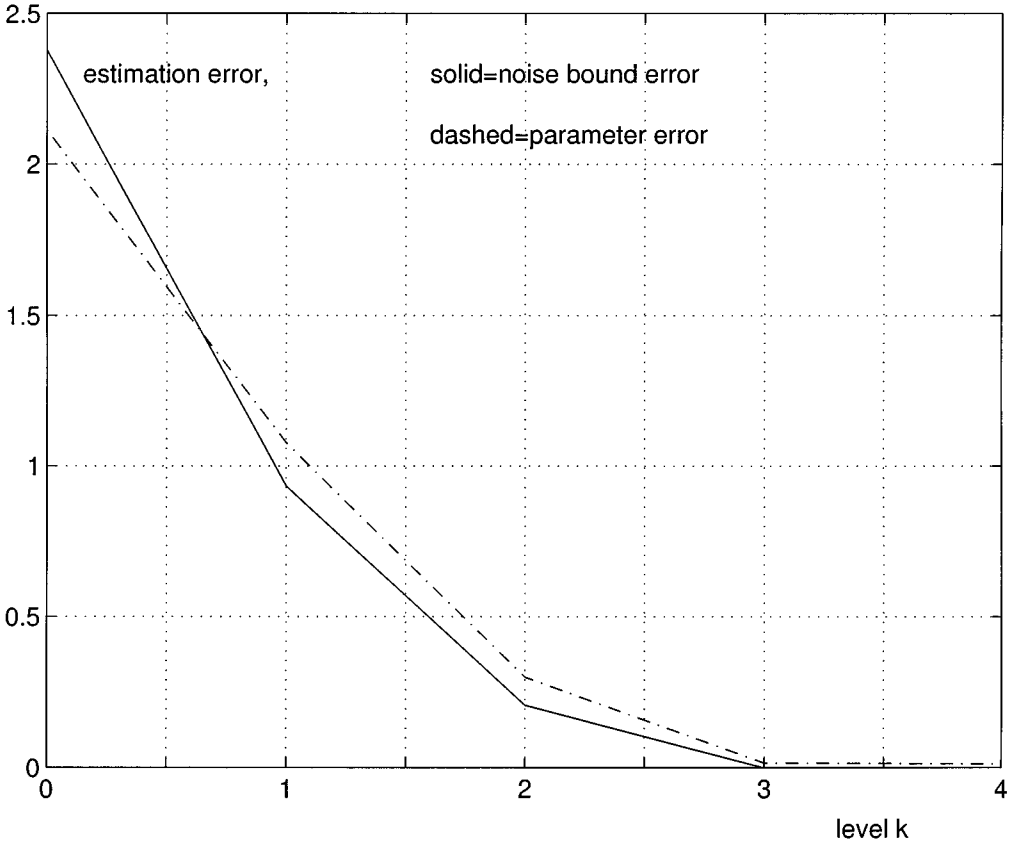

Fig. 3. Estimation errors versus the level k.

with probability one as $n \to \infty$ provided $n - k(n)(l+1) \to \infty$ as $n \to \infty$.

Proof: Note that $\theta \in \Omega_{I_g/(I_g \bigcap I_{i_k})}$ for every $I_g/(I_g \bigcap I_{i_k})$. Now

$$\begin{aligned}\text{dia}\,\Omega_{I_g/(I_g \bigcap I_{i_k})} &= \sup_{\theta_1,\theta_2 \in \Omega_{I_g/(I_g \bigcap I_{i_k})}} \|\theta_1 - \theta_2\|_2 \\ &\le \sup_{\theta_1 \in \Omega_{I_g/(I_g \bigcap I_{i_k})}} \|\theta_1 - \theta\|_2 \\ &\quad + \sup_{\theta_2 \in \Omega_{I_g/(I_g \bigcap I_{i_k})}} \|\theta_2 - \theta\|_2 \\ &= 2 \sup_{\hat{\theta} \in \Omega_{I_g/(I_g \bigcap I_{i_k})}} \|\hat{\theta} - \theta\|_2.\end{aligned}$$

To show dia $\Omega_{I_g/(I_g \bigcap I_{i_k})} \to 0$ with probability one, it suffices to show that for any arbitrary but fixed $\hat{\theta}$, if $\hat{\theta} \in \Omega_{I_g/(I_g \bigcap I_{i_k})}$ as $n \to \infty$, we then have, with probability one, that $\|\tilde{\theta}\| = \|\hat{\theta} - \theta\| \to 0$. For large n, let q be an integer such that

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

Clearly,

$$q - k \ge \frac{n-k}{l} - k - 1 = \frac{1}{l}(n - k(l+1) - l) \to \infty$$

as $n \to \infty$. By the hypothesis, ϕ_i is persistently exciting. Thus, within any window $[(i-1)l+1, (i-1)l+l]$, there always exists at least one $i^0 \in [(i-1)l+1, (i-1)l+l]$ such that for some constant $\alpha > 0$

$$\left|\phi_{i^0}^T \tilde{\theta}\right| = \left|\phi_{i^0}^T (\theta - \hat{\theta})\right| \ge \alpha \|\tilde{\theta}\|.$$

Let $i^0 \in [(i-1)l+1, (i-1)l+l]$, $i = 1, 2, \ldots, q$ be a such time sequence. Now, $|I_{i_k}| = k(n)$, there are at most k windows $[(i-1)l+1, (i-1)l+l]$ that overlap with I_{i_k}. Therefore, the set $I_g/(I_g \bigcap I_{i_k})$ contains at least $q - k$ windows that do not overlap with I_{i_k}. Let the corresponding time subsequence of i^0 be denoted by $i_1^0, i_2^0, \ldots, i_{q-k}^0$. Now, $\hat{\theta} \in \Omega_{I_g/(I_g \bigcap I_{i_k})}$, it follows that at each i_j^0:

$$\epsilon \ge \left|y_{i_j^0} - \phi_{i_j^0}^T \hat{\theta}\right| = \left|\phi_{i_j^0}^T \tilde{\theta} + v_{i_j^0}\right| \tag{6.1}$$

because $i_j^0 \notin I_b \Longrightarrow v_{i_j^0}^b = 0$. Now, from the assumption that $v_{i_j^0}$ approaches both bounds ϵ and $-\epsilon$ with nonzero probability, it follows that at each i_j^0 and for any small $\rho > 0$, we have with nonzero probability:

$$\epsilon \ge \left|\phi_{i_j^0}^T \tilde{\theta} + v_{i_j^0}\right| = \left|\phi_{i_j^0}^T \tilde{\theta}\right| + \left|v_{i_j^0}\right| \ge \alpha \|\tilde{\theta}\| + \left|v_{i_j^0}\right|$$

or

$$\epsilon - \left|v_{i_j^0}\right| \ge \alpha \|\tilde{\theta}\|$$

and

$$\epsilon - \left|v_{i_j^0}\right| \le \rho.$$

In other words, at each i_j^0 and for any small $\rho > 0$, there exists a nonzero probability $p_3(\rho) > 0$ such that

$$\text{Prob}\left\{||\tilde{\theta}|| \leq \frac{1}{\alpha}\rho\right\} \geq p_3(\rho)$$

or

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

Since $v_{i_j^0}$'s, $j = 1, 2, \ldots, q-k$, are independent, it follows that:

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

Thus, as $n \to \infty$

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

Furthermore, for each $\rho > 0$

$$\sum_{q-k=1}^{\infty} (1 - p_3(\rho))^{q-k} \leq \frac{1}{p_3(\rho)} < \infty.$$

This implies by the Borel–Cantelli's lemma [9] that with probability one

$$||\tilde{\theta}|| \to 0$$

as $n \to \infty$. Accordingly

$$\text{dia}\, \Omega_{I_g/(I_g \bigcap I_{i_k})} \leq 2||\tilde{\theta}|| \to 0$$

with probability one as $n \to \infty$. Finally, $\hat{\epsilon} \to \epsilon$ is a direct consequence.

REFERENCES

[1] "Special issue on bounded-error estimation," *Int. J. Adap. Control Signal Processing*, vol. 8, no. 1, 1994.

[2] "Special issue on bounded-error estimation," *Int. J. Adap. Control Signal Processing*, pt. II, vol. 9, no. 1, 1995.

[3] E. W. Bai and H. Cho, "Minimization with few violated constraints and its application in set-membership identification," in *Proc. IFAC World Congress*, vol. H, Bejing, China, 1999, pp. 343–348.

[4] E. W. Bai, H. Cho, and R. Tempo, "Convergence properties of the membership set," *Automatica*, vol. 34, pp. 1245–1249, 1998.

[5] E. W. Bai, Y. Ye, and R. Tempo, "Bounded error parameter estimation: A sequential analytic center approach," *IEEE Trans. Automat. Contr.*, vol. 44, pp. 1107–1117, June 1999.

[6] J. Chen and G. Gu, *Control Oriented System Identification: An H_∞ Approach*. New York: Wiley.

[7] M. Kieffer, J. Jaulin, E. Walter, and D. Meizel, "Nonlinear identification based on unreliable priors and data with application to robot localization," in *Robustness in Identification and Control*, A. Garulli, A. Tesi, and A. Vicino, Eds. London, U.K: Springer-Verlag, 1999, vol. 245, Lecture Notes in Control and Information Science, pp. 190–203.

[8] H. Lahanier, E. Walter, and R. Gomeni, "OMNE: A new robust membership set estimator for the parameter of nonlinear models," *J. Pharm. Biopharm.*, vol. 15, pp. 203–219, 1987.

[9] L. Ljung, *System Identification: Theory for the Users*. Upper Saddle River, NJ: Prentice-Hall, 1987.

[10] J. Matousek, "On geometric optimization with few violated constraints," *Discrete Comput. Geo.*, vol. 14, pp. 365–384, 1995.

[11] M. Milanese, J. Norton, H. Piet-Lahanier, and E. Walter, Eds., *Bounding Approaches to System Identification*. New York: Plenum, 1996.

[12] M. Milanese and A. Vicino, "Optimal estimation theory for dynamic system with set membership uncertainty: An overview," *Automatica*, vol. 27, pp. 997–1009, 1991.

[13] G. K. Murty, *Linear Programming*. New York: Wiley, 1983.

[14] L. Pronzato and E. Walter, "Robustness to outliers of bounded error restimator and consequences on experiment design," in *Bounding Approaches to System Identification*, M. Milanese, J. Norton, H. Piet-Lahanier, and E. Walter, Eds. New York: Plenum.

[15] P. Rousseeuw and A. Leroy, *Robust Regression and Outlier Detection*. New York: Wiley, 1987.

[16] M. Sharir and E. Welzl, "A combinatorial bound for linear programming and related problems," in *Lecture Notes in Computer Science*. Berlin, Germany: Springer-Verlag, 1992, vol. 577, pp. 569–579.

[17] S. M. Veres and J. P. Norton, "Structure selection for bounded parameter models: Consistency condition and selection criterion," *IEEE Trans. Automat. Contr.*, vol. 36, pp. 474–481, Apr. 1991.

Er-Wei Bai (M'90–SM'00) was educated in Fudan University, Shanghai Jiaotong University, both in Shangai, China, and the University of California, Berkeley.

He is Professor of Electrical Engineering at the University of Iowa, Iowa City, where he teaches and conducts research in the area of identification and signal processing.

Dr. Bai serves the IEEE Control Systems Society (CSS) and the International Federation of Automatic Control (IFAC) in various capacities.

Hyonyong Cho received the B.S. degree from Seoul National University, Korea, the M.S. degree from the Korea Advanced Institute of Science and Technology, Korea, both in electrical engineering, and the Ph.D. degree in electrical and computer engineering from the University of Iowa, Iowa City, in 1998.

He joined Korea Telecom Research Center in 1986, and was a member of the technical staff until 1999. He is currently a Visiting Researcher at the University of Iowa, with interest in parameter estimation and signal detection.

Roberto Tempo (M'90–SM'98–F'00) was born in Cuorgne, Italy, in 1956. He graduated in electrical engineering from Politecnico di Torino, Italy, in 1980.

From 1981 to 1983, he was with the Dipartimento di Automatica e Informatica, Politecnico di Torino. Italy. In 1984, he joined the National Research Council (CNR) of Italy at the research institute IRITI, Torino, where he has been a Director of Research of Systems and Computer Engineering since 1991, and is currently an elected member of the Scientific Council. He has held visiting and research positions at the University of Illinois at Urbana-Champaign, German Aerospace Research Organization, Oberpfaffenhofen, and Columbia University, New York. His research activities are mainly focused on robustness analysis and control of uncertain systems and identification of complex systems subject to bounded errors. He has been an Associate Editor of *Systems and Control Letters*, and is currently an Editor of *Automatica*. He is Vice-President for Conference Activities of the Control Systems Society, and is also a member of the European Union Control Association Council.

Dr. Tempo received the "Outstanding Paper Prize Award" from the International Federation of Automatic Control (IFAC) for a paper published in *Automatica* in 1993. He has been an Associate Editor of the IEEE TRANSACTIONS ON AUTOMATIC CONTROL.

Yinyu Ye received the Ph.D. degree from Stanford University, Stanford, CA, in 1988.

After a short postdoctoral program at Cornell University, Ithaca, NY, he joined the faculty of the Department of Management Sciences, the University of Iowa, Iowa City, in 1988, where he is currently Henry B. Tippie Research Professor of Management Sciences. He spent the summer semester of 1991 at Rice University, Houston, TX, fall semester 1993 at Cornell University, and the semester program of 1998 at the University of California, Berkeley. His general research interests lie in the areas of optimization, complexity theory, algorithm design and analysis, and applications of Mathematical Programming, Operations Research and System engineering.