10

Worst-Case Identification Using Quantized Observations

In this chapter, the parameter identification problem under unknown-but-bounded disturbances and quantized output sensors is discussed. In Chapter 9, an input sequence in (9.5) was used to generate observation equations in which only one parameter appears, reducing the problem to the identification of gain systems. A more general input design method is introduced in this chapter to achieve parameter decoupling that transforms a multi-parameter model into a single-parameter model. The input sequence with the shortest length that accomplishes parameter decoupling is sought. Identification algorithms are introduced, and their convergence, convergence rates, and time complexity for achieving a predefined estimation accuracy are investigated.

Section 10.1 formulates the problem and derives a lower bound on identification errors. Section 10.2 studies the input design that achieves parameter decoupling. Parameter decoupling reduces the original identification problem to a one-parameter identification problem. Section 10.3 presents algorithms for identifying one parameter with quantized observations. Time-complexity issues are further studied in Section 10.4. Finally, Section 10.5 illustrates the identification algorithms and their convergence properties by several examples.

L.Y. Wang et al., *System Identification with Quantized Observations*, Systems & Control: Foundations & Applications, DOI 10.1007/978-0-8176-4956-2_10,

10.1 Worst-Case Identification with Quantized Observations

Consider an FIR system

$$y_k = \sum_{i=0}^{n_0-1} a_i u_{k-i} + d_k, \quad k = k_0, k_0+1, \ldots, \tag{10.1}$$

where $\{d_k\}$ is a sequence of disturbances, $\{a_i\}$ are unknown system parameters, and the input $u_k \in \mathbb{U} = \{u_k : 0 < u_k \le u_{\max}, k = k_0, k_0+1, \ldots\}$. The output y_k is measured by a quantized sensor with m_0 thresholds $C_1 < C_2 < \cdots < C_{m_0}$. Namely, the sensor $s = \mathcal{S}(y)$ is represented by the indicator function

$$s_k = \mathcal{S}(y_k) = \sum_{i=1}^{m_0} i I_{\{y_k \in (C_i, C_{i+1}]\}}, \tag{10.2}$$

where $i = 1, \ldots, m_0$ with $C_0 = -\infty$ and $C_{m_0+1} = \infty$. Although a sum is presented in (10.2), at any time k, only one term is nonzero. Hence, $s_k = j$, $j = 0, 1, \ldots, m_0$, implies that $y_k \in (C_j, C_{j+1}]$.

Define $\theta = [a_0, \ldots, a_{n_0-1}]'$. Then the system input–output relationship becomes

$$y_k = \phi_k' \theta + d_k, \tag{10.3}$$

where $\phi_k = [u_k, u_{k-1}, \ldots, u_{k-n_0+1}]'$. The system will be studied under the following assumptions.

(A10.1) For a fixed $p \ge 1$,

(i) the sequence of disturbances $d = \{d_k : k \ge 0\}$ is bounded by $\|d\|_\infty \le \delta$;

(ii) the prior information on θ is given by $\Omega_0 = \text{Ball}_p(\theta_0, e_0) \subset \mathbb{R}^{n_0}$ for some known $\theta_0 \in \mathbb{R}^{n_0}$ and $e_0 > 0$.

For a selected input sequence u_k, let $s = \{s_k, k = k_0, \ldots, k_0+N-1\}$ be the observed output. Define

$$\Omega_N(k_0, u, s) = \Big\{\theta : \quad s_k = \sum_{i=0}^{m_1} i I_{\{\phi_k'\theta + d_k \in (C_i, C_{i+1}]\}} \\ \text{for some } \ |d_k| \le \delta,\ k = k_0, \ldots, k_0+N-1\Big\},$$

and the optimal worst-case uncertainty after N steps of observations as

$$e_N = \inf_{\|u\|_\infty \le u_{\max}} \sup_{k_0} \sup_{s} \text{Rad}_p(\Omega_N(k_0, u, s) \cap \text{Ball}_p(\theta_0, e_0)).$$

Proposition 10.1. *Assuming Assumption* (A10.1)*, for* (10.3)*, the uncertainty set* $\mathrm{Ball}_1(0,(C_1-\delta)/u_{\max})$ *is not identifiable.*

Proof. For any $\theta \in \mathrm{Ball}_p(0,(C_1-\delta)/u_{\max})$, the output

$$\begin{aligned} y_k &= \phi_k'\theta + d_k \le \|\phi(k)\|_\infty \|\theta\|_1 + \delta \\ &\le u_{\max}\frac{C_1-\delta}{u_{\max}} + \delta = C_1. \end{aligned}$$

It follows that $s_k = 1, \forall k$. Hence, the observations could not provide further information to reduce uncertainty. □

10.2 Input Design for Parameter Decoupling

In order to simplify the problem, an input design method was introduced in [111] to decouple system parameters for identification. We are seeking the shortest input sequence lengths to decouple (10.1) into n_0 single-parameter observation equations.

Definition 10.2. An input sequence $\{u_i, i = k_0, \dots, k_0+N_0-1\}$ is said to be n_0*-parameter-decoupling* if, for each $j = 0,\dots,n_0-1$, there exists u_{k_j} such that $y_k = a_j u_{k_j} + d_k$ for some $k_0 \le k \le k_0+N_0-1$ and without considering inputs before time k_0. In other words, the N_0 observation equations contain at least one single-parameter observation equation for each parameter a_j. The input sequence is called the shortest n_0-parameter-decoupling sequence if N_0 is minimal.

Example 10.3. The shortest n_0-parameter-decoupling input segment is not unique. For example, when $n_0 = 3$ and $k_0 = 0$, input $u = \{u_1, 0, 0, u_4, 0, u_6, 0\}$, we have

$$y_5 = a_1 u_4, \quad y_8 = a_2 u_6, \quad y_4 = a_3 u_1.$$

That is, u is a three-parameter-decoupling input segment. By exhaustive testing, we can verify that u is shortest. It can be easily checked that $\{0, u_2^*, 0, u_4^*, 0, 0, u_7^*\}$ is also three-parameter decoupling.

Since parameter decoupling is independent of the actual values of u_{k_j}, for simplicity we will always use $u_{k_j} = 1$ to represent the nonzero input value at k_j.

Definition 10.4. A vector is called a $\{0,1\}$-vector if its components are 1 or 0.

Definition 10.5. A $\{0,1\}$-vector is said to contain the jth row of the $n_0 \times n_0$ identity matrix if the jth row of the $n_0 \times n_0$ identity matrix is a block of it. In this case, the 1 in the block is said to map to the jth row of the $n_0 \times n_0$ identity matrix.

Definition 10.6. A {0,1}-vector is said to be complete if

(i) it contains every row of the $n_0 \times n_0$ identity matrix;

(ii) each 1 maps to at most one row of the $n_0 \times n_0$ identity matrix.

Two complete {0,1}-vectors are equivalent if they have the same length.

Definition 10.7. For a given $\{0,1\}$-vector $b = [b_1, b_2, \ldots, b_N]'$ and $c = [c_1, c_2, \ldots, c_N]'$, if $c_i = b_{N-i+1}$ for $i = 1, 2, \ldots, N$, then b and c are said to be converse to each other.

Definition 10.8. Assume a {0,1}-vector $b = (b_1, b_2, \ldots, b_N)$ is complete, and b_l maps to the first row of the $n_0 \times n_0$ identity matrix. If $c = (c_1, \ldots, c_N)$ satisfies $c_i = b_{l+i-1}$ for $i = 1, \ldots, N - n_0 + 1$ and $c_{N-l+j+2} = b_j$ for $j = 1, \ldots, l-1$, then the transfer from b to c is called initial-1.

Proposition 10.9. *For a complete* $\{0,1\}$*-vector* $b = (b_1, \ldots, b_N)$*, after converse and/or initial-*1 *transfers, the new vector is equivalent to* b*.*

Proof. Assume that $b_{k_i} (i = 1, \ldots, n_0)$ maps to the ith row of the $n_0 \times n_0$ identity matrix.

Converse: Denote $c = (b_N, \ldots, b_1)$ as the vector that is converse to b. By definition, b_{N-k_i+1} is the component of c that maps to the ith row of the $n_0 \times n_0$ identify matrix, so c has property (i) in Definition 10.6. Since $k_i \neq k_j$ for $i \neq j$, we have $N - k_i + 1 \neq N - k_j + 1$. Hence, c has property (ii). In addition, b and c have the same length. So b is equivalent to c.

It is similar to prove for the initial-1 transfer. □

Definition 10.10. Two 1's in a {0,1}-vector are called neighbors if there's no 1 between them.

Proposition 10.11. *For a complete* $\{0,1\}$*-vector, if there are more than* (n_0-1) *0s between two neighboring 1s, then keeping only* (n_0-1) *0s between them will not change its completeness.*

Lemma 10.12. *The shortest length of complete* $\{0,1\}$*-vectors is*

$$\nu(n_0) = \begin{cases} \frac{{n_0}^2+2n_0-1}{2}, & \textit{if } n_0 \textit{ is odd,} \\ \frac{{n_0}^2+2n_0-2}{2}, & \textit{if } n_0 \textit{ is even.} \end{cases} \tag{10.4}$$

Proof. By Propositions 10.11 and 10.9, any {0,1}-vector is equivalent to the one with first $\nu(n_0)$ components:

$$\overbrace{1,0,\ldots,0,1}^{n_0+1},\overbrace{0,1,0,\ldots,0,1}^{n_0+1},\ldots,\overbrace{0,\ldots,0,1,\underbrace{0,\ldots,0}_{n_0-l},1}^{n_0+1},\overbrace{0,\ldots,0}^{l-1}, \tag{10.5}$$

where $l = \frac{n_0+1}{2}$ (or $\frac{n_0}{2}$) when l is odd (or even). For $k \le n_0$, let

$$l_k = \begin{cases} \frac{k+1}{2}, & \text{if } k \text{ is odd,} \\ n_0 - \frac{k}{2} + 1, & \text{if } k \text{ is even.} \end{cases}$$

Then, the kth 1 in (10.5) maps to the l_kth row of the $n_0 \times n_0$ identity matrix.

Hence, the first to the $\nu(n_0)$th components of the vector in (10.5) is complete. □

Theorem 10.13. *For system* (10.1)*, the length of the shortest* n_0*-parameter-decoupling input segment is* $\nu(n_0)$*. Furthermore,*

(i) *if* n_0 *is even,* $u_k = 0$ *for all* k *except* $k = k_0 + i(n_0+2), k_0 + (i+1)(n_0+1) - 1$*, or, for all* k *except* $k = k_0 + \nu(n_0) - i(n_0+2) - 1, k_0 + \nu(n_0) - (i+1)(n_0+1)$*, where* $i = 0, 1, \ldots, \frac{n_0}{2} - 1$*;*

(ii) *if* n_0 *is odd,* $u_k = 0$ *for all except* $k = k_0 + i(n_0+2) + 1, k_0 + (i+1)(n_0+1)$*, and* $k_0 + \frac{n_0^2+n_0}{2}$*; or, all except* $k_0 + \nu(n_0) - i(n_0+2), k_0 + \nu(n_0) - (i+1)(n_0+1) + 1$*, and* $k_0 + \frac{n_0-1}{2}$*, where* $i = 0, 1, \ldots, \frac{n_0-3}{2}$*.*

Proof. Suppose $u = [u_{k_0}, u_{k_0+1}, \ldots, u_{k_0+N-1}]'$ is a shortest n_0-parameter-decoupling input segment. By definition, there exists k_1 such that

$$u_{k_1} \neq 0 \text{ and } u_k = 0 \quad \text{for} \quad k = k_1 - n_0 + 1, \ldots, k_1 - 1. \tag{10.6}$$

So we have $y_{k_1} = a_0 u_{k_1}$, and hence a_0 is decoupled.

Consider the nonzero components of u as 1, u becomes a $\{0,1\}$-vector. Then, (10.6) can be considered as the n_0th row of the $n_0 \times n_0$ identity matrix. Namely, the $\{0,1\}$-vector u must satisfy condition (i) of Lemma 10.12. Furthermore, by the definition of the shortest parameter-decoupling input vector, $k_i \neq k_j$ for $i \neq j$, So the $\{0,1\}$-vector u is required with condition (ii) of Lemma 10.12. Lemma 10.12 confirms the first part of Theorem 10.13. The second part follows the proof of Lemma 10.12. □

10.3 Identification of Single-Parameter Systems

In this section, the identification of single-parameter systems is studied. The conditions of identification using quantized sensors are given, and the effect of threshold values to identification is discussed.

Consider the single-parameter system

$$y_k = a u_k + d_k, \quad k = k_0, k_1 + 2, \ldots, \tag{10.7}$$

where $a \in [\underline{a}_0, \overline{a}_0]$ and $\underline{a}_0 > 0$, $d = \{d_{k_0}, d_{k_0+1}, \ldots\}$ is the sequence of disturbances satisfying $\|d\|_\infty \le \delta$, u_k is the input, and the output y_k is measured by the quantized sensor (10.2).

Remark 10.14. Proposition 10.1 confirms that the uncertainty is irreducible when $|a| < (C_1 - \delta)/u_{\max}$, but in order to investigate the relationship between identification and all threshold values, we assume $|a| > (C_{m_0} - \delta)/u_{\max}$. In this case, input $u_0 = \frac{C_{m_0}+\delta}{C_1-\delta}u_{\max}$. If $s(0) = 0$, $y(0) \le C_1$, which indicates $a > 0$; else, $s(0) \ne 0$ indicates $a < 0$. Thus, the sign of a is known. Without loss of generality, we assume $\underline{a}_0 > (C_{m_0} - \delta)/u_{\max}$.

For simplification, the following symbols will be used in this chapter:

1. $\underline{a}_k, \overline{a}_k$: the uncertainty upper and lower bounds of a at time k;
2. $e_k = \overline{a}_k - \underline{a}_k$;
3. $L_k = e_k/e_{k-1}$;
4. $r_k = \underline{a}_k/\overline{a}_k$;
5. $\Delta C_l = C_{l+1} - C_l$, $l = 0, \ldots, m_0 - 1$;
6. $\max \Delta C_l = \max_{1\le l\le m_0-1} \Delta C_l$ and $\min \Delta C_l = \min_{1\le l\le m_0-1} \Delta C_l$;
7. $R = \frac{C_1 - \max \Delta C_l - \delta}{C_{m_0} + \max \Delta C_l + \delta}$;
8. $P_k(i,j) = \frac{(C_i+\delta)\overline{a}_{k-1} - (C_j-\delta)\underline{a}_{k-1}}{(C_i+C_j)e_{k-1}}$, $i \le j$;
9. $Q_k(i,j) = \frac{\overline{a}_{k-1}(\max_{i\le l\le j-1}\Delta C_l + 2\delta)}{(C_j + \max_{i\le l\le j-1}\Delta C_l + \delta)e_{k-1}}$, $i \le j$;
10. $\vee\{x_1, x_2\} = \max\{x_1, x_2\}$ and $\wedge\{x_1, x_2\} = \min\{x_1, x_2\}$;
11. $L_k(i,j) = \wedge\{\vee\{P_k(i,j), Q_k(i,j)\}, 1\}$;
12. $\widetilde{L}_k(i) = \wedge\{P_k(i,i), 1\}$.

10.3.1 General Quantization

A quantized sensor is binary when $m_0 = 1$. By Chapter 9, when the decoupled observations (10.7) are measured by a binary sensor with threshold C_1, in the worst case, the minimum of L_k is $\widetilde{L}_k(1)$ and the optimal input is $u_k = \frac{2C_1}{\overline{a}_{k-1} + \underline{a}_{k-1}}$. Subsequently, we will consider $m_0 \ge 2$.

Theorem 10.15. *For* (10.7)*, when* $e_{k-1} < (\min \Delta C_l + 2)/u_{\max}$*, the minimum of* L_k *is* $L_k = \widetilde{L}_k(m_0)$ *in the worst-case sense.*

Proof. By (10.7), we have $a = (y_k - d_k)/u_k$. The necessary condition of $L_k < 1$ is that for all $u_k \in \mathbb{U}$, so there exists some threshold C_i such that

$$\underline{a}_{k-1} \le \frac{C_i - \delta}{u_k} \le \frac{C_i + \delta}{u_k} \le \overline{a}_{k-1}. \tag{10.8}$$

Suppose C_{i_0} satisfies (10.8), since

$$e_{k-1} < (\min \Delta C_l + 2\delta)/u_{\max} \le (C_{i_0} - C_{i_0-1} + 2\delta)/u_{\max},$$

we have

$$C_{i_0-1} - \delta < C_{i_0} + \delta - u_{\max}(\overline{a}_{k-1} - \underline{a}_{k-1}).$$

By (10.8), $C_{i_0} + \delta < \overline{a}_{k-1} u_{\max}$, so

$$\frac{C_{i_0-1} - \delta}{u_k} < \underline{a}_{k-1}.$$

Since $C_1 < C_2 < \cdots < C_{m_0}$, for $i < i_0$, C_i does not satisfy (10.8). Similarly, for $i > i_0$, C_i does not satisfy (10.8). Namely, for a given $u_k \in \mathbb{U}$, there exists one threshold at most, which satisfies (10.8). By Chapter 9, in a worst-case sense, the minimum of L_k is $\widetilde{L}_{i_0}(k)$ when only C_{i_0} satisfies (10.8), and the optimal input is $u_k = 2C_{i_0}/(\overline{a}_{k-1} + \underline{a}_{k-1})$.

Furthermore, since $\widetilde{L}_i(k)$ is monotonically decreasing for i and $C_1 < C_2 < \cdots < C_{m_0}$, the minimum of L_k is $\widetilde{L}_1(k)$ and the optimal input is $u_k = 2C_1/(\overline{a}_{k-1} + \underline{a}_{k-1})$. □

In order to design input with quantized sensors to make L_k less than $\widetilde{L}_k(C_1)$, we now describe how to identify systems with quantized observations.

Theorem 10.16. *Assume $\delta < \min \Delta C_l/2$, $u_{\max} \ge (C_{m_0} - \delta)/\underline{a}_0$, and $e_{k-1} \ge (C_{m_0} - C_1 + 2\delta)/u_{\max}$. Then*

$$L_k = L_k(1, m_0). \tag{10.9}$$

Furthermore,

(i) *if $r_{k-1} \le R$, then $L_k = P_k(1, m)$ and the optimal input is $u_k = \frac{C_1 + C_{m_0}}{\overline{a}_{k-1} + \underline{a}_{k-1}}$;*

(ii) *if $r_{k-1} > R$, then $L_k = Q_k(1, m)$ and the optimal input is $u_k = \frac{C_{m_0} + \max \Delta C_+ \delta}{\overline{a}_{k-1}}$.*

Proof. Since $e_{k-1} \ge (C_{m_0-1} - C_1 + 2\delta)/u_{\max}$, there exists $u_k \in \mathbb{U}$ such that

$$\frac{C_1 + \delta}{u_k} \ge \underline{a}_{k-1}, \quad \frac{C_{m_0} + \delta}{u_k} \le \overline{a}_k,$$

which together with $\delta < \min \Delta C_l/2$ gives

$$C_{l+1} - \delta \ge C_l + \delta, \quad \text{for} \quad l = 1, 2, \ldots, m_0 - 1.$$

Hence, there exists $u_k \in \mathbb{U}$ such that

$$\underline{a}_{k-1} \le \frac{C_1 - \delta}{u_k} \le \frac{C_1 + \delta}{u_k} \le \cdots \le \frac{C_{m_0} - \delta}{u_k} \le \frac{C_{m_0} + \delta}{u_k} \le \overline{a}_{k-1}. \tag{10.10}$$

By (10.4), $a = (y_k - d_k)/u_k$. For the input in (10.10), consider s_k: If $s_k = m_0$, then $a > (C_{m_0} - \delta)/u_k$, and hence,

$$\underline{a}_k = \vee\{(C_{m_0} - \delta)/u_k, \underline{a}_{k-1}\} = (C_{m_0} - \delta)/u_k, \quad \overline{a}_k = \overline{a}_{k-1}.$$

Denote $\gamma_k = u_k e_{k-1}$. Then, we have

$$L_k = (\overline{a}_{k-1} u_k - (C_{m_0} - \delta))/\gamma_k.$$

If $s_k = j$, $j = 1, \ldots, m_0 - 1$, then $(C_j - \delta)/u_k < a \le (C_{j+1} + \delta)/u_k$, and hence,

$$\underline{a}_k = (C_j - \delta)/u_k, \quad \overline{a}_k = (C_{j+1} + \delta)/u_k, \quad L_k = (\Delta C_j + 2\delta)/\gamma_k.$$

If $s_k = 0$, then $a \le (C_1 - \delta)u_k$, and hence,

$$\underline{a}_k = \underline{a}_{k-1}, \quad \overline{a}_k = (C_1 + \delta)/u_k, \quad L_k = (C_1 + \delta - \underline{a}_{k-1} u_k)/\gamma_k.$$

So, in the worst case, we have

$$\begin{cases} L_k \ge (\overline{a}_{k-1} u_k - (C_{m_0} - \delta))/\gamma_k, \\ L_k \ge (\Delta C_j + 2\delta)/\gamma_k, \\ L_k \ge (C_1 + \delta - \underline{a}_{k-1} u_k)/\gamma_k. \end{cases} \tag{10.11}$$

Since $u_k > 0$, (10.11) is equivalent to

$$\begin{cases} u_k & \le \frac{C_{m_0} - \delta}{\overline{a}_{k-1} - L_k e_{k-1}}, \\ u_k & \ge \frac{\Delta C_j + 2\delta}{L_k e_{k-1}}, \\ u_k & \ge \frac{C_1 + \delta}{\underline{a}_{k-1} + L_k e_{k-1}}, \end{cases}$$

which means that there exists $u_k \in \mathbb{U}$ satisfying (10.10) such that

$$L_k \ge P_k(1, m_0), \quad L_k \ge Q_k(1, m_0), \tag{10.12}$$

and the equalities in (10.12) are achieved in the case of

$$u_k = \frac{C_{m_0} + C_1}{\overline{a}_{k-1} + \underline{a}_{k-1}} \quad \text{and} \quad u_k = \frac{C_{m_0} + \max\{C_i - C_{i+1}\} + d}{\overline{a}_{k-1}},$$

respectively. So, (10.9) is true.

Furthermore, when $r_{k-1} \le R$, we have $P_k(1, m_0) \ge Q_k(1, m_0)$. Let $u_k = \frac{C_1 + C_{m_0}}{\overline{a}_{k-1} - \underline{a}_{k-1}}$. Then, by

$$u_k = \frac{C_1 + C_{m_0}}{\overline{a}_{k-1} + \underline{a}_{k-1}} \le \frac{C_{m_0} - \delta + C_1 + \delta}{2\underline{a}_{k-1}} \le \frac{C_{m_0} - \delta}{\underline{a}_{k-1}} \le u_{\max},$$

$u_k \in \mathbb{U}$. From (10.11), one can get $L_k = P_k(1, m_0)$.

Similarly, when $r_{k-1} > R$, let $u_k = \frac{C_{m_0} + \max \Delta C_l + \delta}{\overline{a}_{k-1}}$. Then, $u_k \in \mathbb{U}$ and $L_k = Q_k(1, m_0)$. □

Theorem 10.16 shows that when $r_{k-1} \leq R$, which means that the uncertainty bound is large, $\overline{a}_{k-1} - (C_{m_0} - \delta)/u_k$ and $(C_1 + \delta)/u_k - \underline{a}_{k-1}$ are larger than $(\max \Delta C_l + 2\delta)/u_k$. So, L_k is determined by C_1 and C_{m_0}. When $\underline{a}_{k-1}/\overline{a}_{k-1} > R$, which means the uncertainty bound is small, $\overline{a}_{k-1} - (C_{m_0} - \delta)/u_k$ and $(C_1 + \delta)/u_k - \underline{a}_{k-1}$ are less than $(\max \Delta C_l + 2\delta)/u_k$. Then L_k is determined by the two neighboring thresholds with the maximum space between them. As a result, L_k derived from $m_0 - 1$ thresholds may be less than the one derived from m_0 thresholds.

Example 10.17. Let $C_1 = 85$, $C_2 = 94$, $C_3 = 97$, $C_4 = 100$, and $\delta = 1$. Compare L_k derived from $\{C_1, C_2, C_3, C_4\}$ and $\{C_2, C_3, C_4\}$.

1. For $\{C_1, C_2, C_3, C_4\}$, if $r_0 \leq \frac{C_1 - (C_2 - C_1) - \delta}{C_4 + (C_2 - C_1) + \delta} = \frac{15}{22}$, then

$$L_1 = P_1(1, 4) = \frac{86\overline{a}_0 - 99\underline{a}_0}{185 e_0}.$$

If $r_0 > \frac{15}{22}$, then

$$L_1 = Q_1(1, 4) = \frac{\overline{a}_0}{10 e_0}.$$

2. For $\{C_2, C_3, C_4\}$, if $r_0 \leq \frac{C_2 - (C_3 - C_2) - \delta}{C_4 + (C_3 - C_2) + \delta} = \frac{45}{52}$, then

$$L_1 = P_1(2, 4) = \frac{95\overline{a}_0 - 99\underline{a}_0}{194 e_0}.$$

If $r_0 > \frac{45}{52}$, then

$$L_1 = Q_1(2, 3) = \frac{\overline{a}_0 (C_4 - C_3 + 2\delta)}{(2C_4 - C_3 + \delta) e_0} = \frac{5\overline{a}_0}{104 e_0}.$$

The curves of L_1 with r_0 are plotted in Figure 10.1. It is shown that we have $L_1(1, 4) \leq L_1(2, 4)$ as $r_0 \leq 0.77$ and $L_1(1, 4) > L_1(2, 4)$ as $r_0 > 0.77$. The reason is that the space between C_1 and C_2 is larger than others, so $\max_{1 \leq l \leq 3} \Delta C_l > \max_{2 \leq l \leq 3} \Delta C_l$, which causes the result in Figure 10.1 by (10.9).

Before studying the general case, we start from the case $m_0 = 2$.

Corollary 10.18. *For system* (10.7) *with* $m_0 = 2$ *and* $u_k \in \mathbb{U}$, *in the worst case, the minimum value of* L_k *is*

$$L_k = \wedge\{\vee\{P_k(1, 2), Q_k(1, 2)\}, \widetilde{L}_k(1)\}. \tag{10.13}$$

Furthermore,

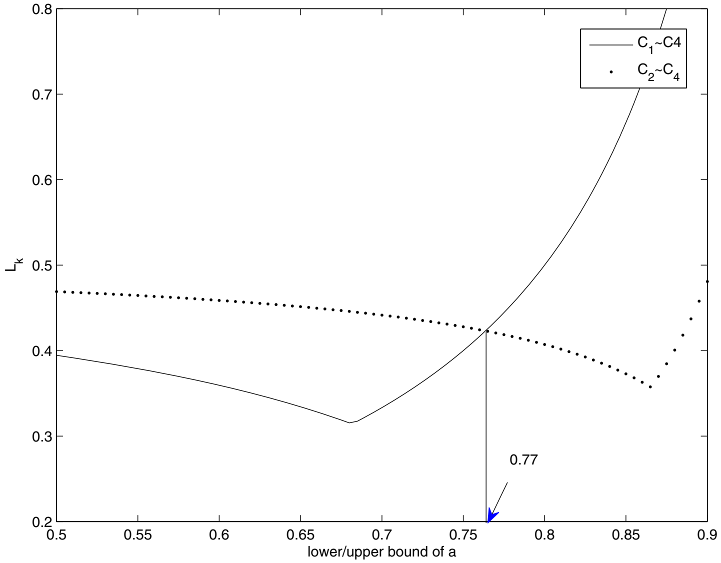

FIGURE 10.1. L_k derived from $C_1 \sim C_4$ and $C_2 \sim C_4$

(i) *if*

$$r_{k-1} \leq \frac{2C_1 - C_2 - \delta}{2C_2 - C_1 + \delta},$$

then $L_k = P_1(1,2)$ *and the optimal input is*

$$u_k = \frac{C_1 + C_2}{\overline{a}_{k-1} + \underline{a}_{k-1}};$$

(ii) *if*

$$\frac{2C_1 - C_2 - \delta}{2C_2 - C_1 + \delta} \leq r_{k-1} \leq \frac{C_1 - \delta}{2C_2 - C_1 + \delta},$$

then $L_k = Q_k(1,2)$ *and the optimal input is*

$$u_k = \frac{2C_2 - C_1 + \delta}{\overline{a}_{k-1}};$$

(iii) *if*

$$\frac{C_1 - \delta}{2C_2 - C_1 + \delta} < r_{k-1} \leq \frac{C_2 - \delta}{C_2 + \delta},$$

then $L_k = \widetilde{L}_k(1)$ *and the optimal input is*

$$u_k = \frac{2C_2}{\overline{a}_{k-1} + \underline{a}_{k-1}};$$

(iv) *if*

$$r_{k-1} > \frac{C_2 - \delta}{C_2 + \delta},$$

then $L_k = 1$ *for any* $u_k \in \mathbb{U}$.

Proof. Since $m_0 = 2$, we can construct inputs with one or two thresholds. By Chapter 9, the minimum L_k with a single threshold is $L_k = \widetilde{L}_k(2)$. In addition to Theorem 10.16, (10.13) is true.

For (i), we have $P_k(1,2) \geq Q_k(1,2)$, which together with $C_1 < C_2$ gives

$$P_k(1,2) = \frac{(C_1 + \delta)\overline{a}_{k-1} - (C_2 - \delta)\underline{a}_{k-1}}{(C_1 + C_2)e_{k-1}} \leq \widetilde{L}_k(2).$$

Hence, $L_k = \widetilde{L}_k(C_2)$.

For (ii), we have $P_k(1,2) \leq Q_k(1,2)$, which together with $r_{k-1} \leq \frac{C_1-\delta}{2C_2-C_1+\delta}$ gives $Q_k(1,2) \leq \widetilde{L}_k(2)$. Hence, $L_k = Q_k(1,2)$. By Theorem 10.16, one can get the optimal input of (i)–(iii).

For (iii), we have $L_k = \widetilde{L}_k(m_0)$. By Chapter 9, the optimal input is $u_k = \frac{2C_2}{\overline{a}_{k-1}+\underline{a}_{k-1}}$.

For (iv), by Chapter 9 we have $L_k = 1$ for any inputs. □

10.3.2 Uniform Quantization

Consider that L_k is affected by C_1, C_{m_0}, and $\max \Delta C_l$, and $Q_k(1, m_0)$ decreases with decreasing $\max \Delta C_l$. In order to minimize $\max \Delta C_l$, we assume

$$C_2 - C_1 = C_3 - C_2 = \cdots = C_{m_0} - C_{m_0-1} := \Delta C.$$

Lemma 10.19. *Consider* (10.7). *Assume that* $\delta < \Delta C/2$, $u_{\max} \geq (C_{m_0} - \delta)/\underline{a}_0$, *and* $e_{k-1} \geq (C_{m_0} - C_1 + 2\delta)/u_{\max}$. *For* $2 \leq i \leq j \leq m_0$, *denote*

$$\widetilde{L}_k(i,j) = \vee\{P_k(i,j), Q_k(i-1,i)\}.$$

In the worst case, we have

$$\widetilde{L}_k(1, m_0) \leq \wedge\{\widetilde{L}_k(1, m_0 - 1), \widetilde{L}_k(C_2, C_{m_0})\}. \tag{10.14}$$

Proof. By $C_1 < C_2$, we have $P_k(1, m_0) \leq Q_k(1, m_0 - 1)$ and $\widetilde{L}_k(1, m_0) \leq \widetilde{L}_k(1, m_0 - 1)$. To prove (10.14), we need only show that $\widetilde{L}_k(1, m_0) \leq \widetilde{L}_k(2, m_0)$.

For $r_{k-1} \leq \frac{C_1-\Delta C-\delta}{C_i+\Delta C+\delta}$, $i = m_0 - 1, m_0$, we have $P_k(1,i) \leq Q_k(i-1,i)$, and hence, $\widetilde{L}_k(1,i) = P_k(1,i)$. Similarly, for $r_{k-1} > \frac{C_1-\Delta C-\delta}{C_i+\Delta C+\delta}$, we have $\widetilde{L}_k(1,i) = Q_k(i-1,i)$.

Considering that $\widetilde{L}_k(1,i)$ is piecewise about $\underline{a}_{k-1}/\overline{a}_{k-1}$, we study the following three cases:

(i) When $r_{k-1} \le \frac{C_1-\Delta C-\delta}{C_{m_0}+\Delta C+\delta}$, we have

$$\widetilde{L}_k(1,i) = P_k(1,i), \quad i = m_0-1,\ m_0.$$

Noting that

$$\begin{aligned} P_k(1,i) &= \frac{(C_1+\delta)\overline{a}_{k-1} - (C_i-\delta)\underline{a}_{k-1}}{(C_1+C_i)e_{k-1}} \\ &= \frac{(C_1+\delta)(\overline{a}_{k-1}+\underline{a}_{k-1})}{(C_1+C_i)e_{k-1}} - \frac{\underline{a}_{k-1}}{e_{k-1}}, \end{aligned}$$

and $C_{m_0} \ge C_2$, we have $P_k(1,m_0) \le P_k(2,m_0)$, and hence, $\widetilde{L}_k(1,m_0) \le \widetilde{L}_k(C_2,C_{m_0})$.

(ii) When $r_{k-1} \ge \frac{C_1-\Delta C-\delta}{C_{m_0-1}+\Delta C+\delta}$, we have

$$\widetilde{L}_k(1,i) = Q_k(i,i+1), \quad i = m_0-1, m_0.$$

Noticing that

$$Q_k(i,i+1) = \frac{\overline{a}_{k-1}(\Delta C+2\delta)}{(C_i+\Delta C+\delta)e_{k-1}} = \frac{\overline{a}_{k-1}(\Delta C+2\delta)}{(C_i+\Delta C+\delta)e_{k-1}}$$

and $C_{m_0-1} < C_{m_0}$, we have $Q_k(m_0-2,m_0-1) \le Q_k(m_0-1,m_0)$, and hence, $\widetilde{L}_k(1,m_0) \le \widetilde{L}_k(1,m_0-1)$.

(iii) When $\frac{C_1-\Delta C-\delta}{C_{m_0}+\Delta C+\delta} < r_{k-1} < \frac{C_1-\Delta C-\delta}{C_{m_0-1}+\Delta C+\delta}$, we have

$$\widetilde{L}_k(1,m_0) = Q_k(m_0-1,m_0) \quad \text{and} \quad \widetilde{L}_k(1,C_{m_0-1}) = P_k(1,m_0-1).$$

By (ii), we have $r_{k-1} = \frac{C_1-\Delta C-\delta}{C_2+\Delta C+\delta}$, and hence, $\widetilde{L}_k(1,m_0) \le \widetilde{L}_k(1,m_0-1)$. Notice that $\widetilde{L}_k(1,m_0) = Q_k(m_0-1,m_0)$ increases about $\underline{a}_{k-1}/\overline{a}_{k-1}$. Then $\widetilde{L}_k(1,m_0-1) = Q_k(m_0-1,m_0)$ decreases about r_{k-1}. Thus, for case (iii), we have $\widetilde{L}_k(1,m_0) \le \widetilde{L}_k(1,m_0-1)$. To summarize, (10.14) is true. □

Theorem 10.20. *Consider* (10.7)*. Assume* $\delta < \Delta C/2$ *and* $e_{k-1} \ge (C_{m_0} - C_1 + 2\delta)/u_{\max}$*. Then for* $u_k \in \mathbb{U}$*, in the worst case, the minimum* L_k *is that*

$$L_k = \wedge\{\vee\{P_k(1,m_0), Q_k(m_0-1,m_0)\}, \widetilde{L}_k(m_0)\}. \tag{10.15}$$

Furthermore,

(i) *for* $r_{k-1} \le \frac{C_{m_0}-m_0\Delta C-\delta}{C_{m_0}+\Delta C+\delta}$*,* $L_k = P_k(1,m_0)$ *and the optimal input is* $u_k = \frac{C_1+C_{m_0}}{\overline{a}_{k-1}+\underline{a}_{k-1}}$*;*

(ii) *for* $\frac{C_{m_0}-m_0\Delta C-\delta}{C_{m_0}+\Delta C+\delta} < r_{k-1} \leq \frac{C_{m_0-1}-\delta}{C_{m_0}+\Delta C+\delta}$, $L_k = Q_k(m_0-1, m_0)$ *and the optimal input is* $u_k = \frac{C_{m_0}+\Delta C+\delta}{\overline{a}_{k-1}}$*;*

(iii) *for* $\frac{C_{m_0-1}-\delta}{C_{m_0}+\Delta C+\delta} < r_{k-1} \leq \frac{C_{m_0}-\delta}{C_{m_0}+\delta}$, $L_k = \widetilde{L}_k(m_0)$ *and the optimal input is* $u_k = \frac{2C_{m_0}}{\overline{a}_{k-1}+\underline{a}_{k-1}}$*;*

(iv) *for* $r_{k-1} > \frac{C_{m_0}-\delta}{C_{m_0}+\delta}$, $L_k = 1$ *for any* $u_k \in \mathbb{U}$.

Proof. By Corollary 10.18, (10.15) is true for $m_0 = 2$. Assume that for $i \geq 2$, (10.15) is true for $m_0 = i$. Then for $m_0 = i+1$, we have

$$L_k = \wedge\{\widetilde{L}_k(1, i+1), \widetilde{L}_k(1, i), \widetilde{L}_k(2, i+1), \widetilde{L}_k(i+1)\}.$$

By Lemma 10.19, we have $\widetilde{L}_k(1, i+1) \leq \wedge\{\widetilde{L}_k(1, i), \widetilde{L}_k(2, i+1)\}$, which implies that (10.15) is true for $m = i+1$. Thus, by induction we have (10.15).

The rest can be obtained by comparing $P_k(1, m_0)$, $Q_k(m_0-1, m_0)$, and $\widetilde{L}_k(m_0)$. □

By Theorem 10.20, $L_k < \widetilde{L}_k(m_0)$ is equivalent to $\widetilde{L}_k(1, m_0) < \widetilde{L}_k(m_0)$, or equivalently, $r_{k-1} < \frac{C_{m_0-1}-\delta}{C_{m_0}+\Delta C+\delta}$.

Theorem 10.21. *Consider* (10.7)*. Assume* $\delta < \Delta C/2$ *and denote*

$$\mathcal{K} = \{k : e_{k-1} \geq (C_{m_0} - C_1 + 2\delta)/u_{\max}\}.$$

For each k*, choose the optimal* u_k *to minimize* L_k*; then in the worst case, there exists at most one* $k \in \mathcal{K}$ *such that* $L_k = Q_k(m_0-1, m_0)$*.*

Proof. Since $\underline{a}_k$ and $\overline{a}_k$ are increasing and decreasing with respect to k, by Theorem 10.20, for

$$\frac{C_{m_0} - m_0\Delta C - \delta}{C_{m_0} + \Delta C + \delta} < r_{k-1} \leq \frac{C_{m_0-1} - \delta}{C_{m_0} + \Delta C + \delta},$$

the minimum of L_k is $L_k = Q_k(m_0-1, m_0)$. Hence, if there exist two k's in $\mathcal{K}$ such that $L_k = Q_k(m_0-1, m_0)$, then they must be neighbors. Denote them as k_0, k_0+1, respectively. Then,

$$\frac{L_{k_0+1}}{L_{k_0}} = \frac{\overline{a}_{k_0}}{\overline{a}_{k_0-1}} \frac{e_{k_0-1}}{e_{k_0}} = \frac{\overline{a}_{k_0}}{\overline{a}_{k_0-1}} \frac{1}{L_{k_0}},$$

or equivalently,

$$L_{k_0+1} = \frac{\overline{a}_{k_0}}{\overline{a}_{k_0-1}}.$$

Since $\overline{a}_{k_0} = \overline{a}_{k_0+1}$ in the worst case, we have $L_{k_0+1} = 1$, which contradicts the fact that $L_{k_0+1} < \widetilde{L}_{k_0+1}(m_0) < 1$ for

$$\frac{C_{m_0} - m_0 \Delta C - \delta}{C_{m_0} + \Delta C + \delta} < \frac{\underline{a}_{k_0}}{\overline{a}_{k_0}} \leq \frac{C_{m_0-1} - \delta}{C_{m_0} + \Delta C + \delta}.$$

□

By Theorem 10.21, we can figure out how many k's there exist such that $L_k < \widetilde{L}_k(m_0)$.

Theorem 10.22. *Consider* (10.7). *Assume $\delta < \Delta C/2$ and denote $\mathcal{K} = \{k : e_{k-1} \geq (C_{m_0} - C_1 + 2\delta)/u_{\max}\}$. In the worst case, choose the optimal input to minimize L_k. Then, the total number K of the elements $k \in \mathcal{K}$ such that $L_k < \widetilde{L}_k(m_0)$ satisfies*

$$\frac{\ln \alpha_4 - \ln(1-\alpha_4)}{\ln \alpha_3} \leq K \leq \frac{\ln \alpha_2 - \ln(1-\alpha_2)}{\ln \alpha_1} + 1, \tag{10.16}$$

where

$$\begin{aligned} \alpha_1 &= \frac{C_1 + \delta}{C_1 + C_{m_0}}, \quad \alpha_2 = \frac{(C_{m_0} - C_1 - 2\delta)\underline{a}_0}{(C_{m_0} - \delta)}, \\ \alpha_3 &= \frac{C_{m_0} - \delta}{C_1 + C_{m_0}}, \quad \alpha_4 = \frac{(C_{m_0} - C_1 - 2\delta)\overline{a}_0}{(C_{m_0} + \delta)}. \end{aligned}$$

Proof. By Theorem 10.20, $L_k = \wedge\{\vee\{Q_k(1, m_0), Q_k(m_0{-}1, m_0)\}, \widetilde{L}_k(m_0)\}$ for $k \in \mathcal{K}$ and $L_k < \widetilde{L}_k(m_0)$ only for $\frac{\underline{a}_{k-1}}{\overline{a}_{k-1}} < \frac{C_{m_0-1}-\delta}{C_{m_0}+\Delta C+\delta}$.

Furthermore, for $r_{k-1} \leq \frac{C_{m_0} - m_0\Delta C - \delta}{C_{m_0}+\Delta C+\delta}$, we have

$$L_k = P_k(1, m_0), \tag{10.17}$$

and for $\frac{C_{m_0} - m_0\Delta C - \delta}{C_{m_0}+\Delta C+\delta} < r_{k-1} \leq \frac{C_2-\delta}{C_{m_0}+\Delta C+\delta}$,

$$L_k = Q_k(m_0 - 1, m_0). \tag{10.18}$$

Thus, by Theorem 10.21, there exists at most one k satisfying (10.18).

From (10.17), one obtains

$$e_k = \alpha_1 e_{k-1} - \frac{(C_{m_0} - C_1 - 2\delta)}{(C_1 + C_{m_0})}\underline{a}_{k-1} \tag{10.19}$$

and

$$e_k = \alpha_3 e_{k-1} - \frac{(C_{m_0} - C_{m_0} - 2\delta)}{(C_1 + C_{m_0})}\overline{a}_{k-1}. \tag{10.20}$$

By (10.19) and the fact that $\underline{a}_i$ decreases about i, we have

$$\begin{aligned} e_k &= \alpha_1^k e_0 - \frac{(C_{m_0} - C_1 - 2)}{(C_1 + C_{m_0})} \sum_{i=0}^{k-1} \alpha_1^i \underline{a}_{k-i-1} \\ &\leq \alpha_1^k e_0 - \alpha_2(1 - \alpha_1^k) \\ &= (1 - \alpha_2)\alpha_1^k - \alpha_2. \end{aligned}$$

By $e_k > 0$, or equivalently, $(1-\alpha_2)\alpha_1^k - \alpha_2 > 0$, we have

$$k \leq \frac{\ln \dfrac{\alpha_2}{1-\alpha_2}}{\ln \alpha_1}.$$

Similarly,

$$k \geq \frac{\ln \alpha_4 - \ln(1-\alpha_4)}{\ln \alpha_3}.$$

Thus, by (10.18), we have (10.16). □

Remark 10.23. In Theorem 10.22, K is estimated by using the prior information of parameters. Since the bound of the unknown parameters decreases, the estimate of K is more accurate if updated parameter bounds are used.

10.4 Time Complexity

Section 10.3.2 shows that for uniform quantization with m_0 thresholds, in the worst case, the minimum L_k is

$$L_k = \wedge\{\vee\{P_k(1,m_0), Q_k(m_0-1,m_0)\}, \widetilde{L}_k(m_0)\}.$$

Considering that $\widetilde{L}_k(m_0)$ is about the largest threshold, and

$$r_{k-1} < \frac{C_{m_0-1} - \delta}{C_{m_0} + \Delta C + \delta},$$

we have $L_k < \widetilde{L}_k(m_0)$. Furthermore, by Theorem 10.21, in the worst case, there exists at most one $k \in \mathcal{K} = \{k : e_{k-1} \geq (C_{m_0} - C_1 + 2\delta)/u_{\max}\}$ such that $L_k = Q_k(m_0-1, m_0)$. So, we will study the effect of $P_k(1,m_0)$ on the parameter estimation and its time complexity.

(A10.2) $C_2 - C_1 = C_3 - C_2 = \cdots = C_{m_0} - C_{m_0-1} := \Delta C$, $\delta < \Delta C/2$, and $e > (C_{m_0} - C_1 + 2\delta)/u_{\max}$.

Lemma 10.24. *Consider* (10.7). *Let*

$$\sigma = \frac{m_0 \Delta C + 2\delta}{C_{m_0} + \Delta C - \delta}\overline{a}_0.$$

For $e \in (\sigma, e_0)$, $N(e)$ is the time complexity of the parameter's unknown bound from e_0 to e. Choose optimal inputs in each time k. Then,

$$N(e) \leq \frac{\ln(\alpha_2 + e) - \ln[(1-\alpha_2)e_0]}{\ln \alpha_1}, \tag{10.21}$$

where

$$\alpha_1 = \frac{C_1 + \delta}{C_1 + C_{m_0}}, \quad \alpha_2 = \frac{(C_{m_0} - C_1 - 2\delta)\underline{a}_0}{(C_{m_0} - \delta)}.$$

Proof. Since $e_{k-1} > e > \sigma$ and $\overline{a}_0 \geq \overline{a}_{k-1}$,

$$r_{k-1} \leq \frac{C_{m_0} - m_0\Delta C - \delta}{C_{m_0} + \Delta C + \delta}.$$

By Theorem 10.20, the minimum of L_k is $P_k(1, m_0)$. Hence,

$$\begin{aligned} e_k &= \alpha_1 e_{k-1} - \frac{C_{m_0} - C_1 - 2\delta}{C_1 + C_{m_0}} \underline{a}_{k-1} \\ &\leq \alpha_1^k e_0 - \frac{(C_{m_0} - C_1 - 2\delta)\underline{a}_0}{C_1 + C_{m_0}} \sum_{i=0}^{k-1} \alpha_1^i \\ &= (1-\alpha_2)\alpha_1^k e_0 - \alpha_2. \end{aligned}$$

In the worst case, a necessary and sufficient condition of $e_k \leq e$ is

$$(1-\alpha_2)\alpha_1^k e_0 - \alpha_2 \leq e,$$

or equivalently,

$$k \geq \frac{\ln \dfrac{\alpha_2 + e}{(1-\alpha_2)e_0}}{\ln \alpha_1}.$$

Thus, (10.21) is true. □

Theorem 10.25. *For* (10.3)*, let*

$$\begin{aligned} \overline{\theta}_0 &= \max_{1\leq i\leq n_0} \overline{a}_0(i), & \underline{\theta}_0 &= \min_{1\leq i\leq n_0} \underline{\theta}_0(i), \\ \sigma &= \frac{m_0\Delta C + 2\delta}{C_{m_0} + \Delta C - \delta}\overline{a}_0, & e_0 &= \mathrm{Rad}_p(\Omega_0). \end{aligned}$$

Then, for any given $e \in (\sigma, e_0)$, we have

$$N(e) \leq \nu(n_0)\frac{\ln(\alpha_2 + e/n_0^{1/p}) - \ln[(1-\alpha_2)e_0]}{\ln \alpha_1}, \tag{10.22}$$

where

$$\alpha_1 = \frac{C_1 + \delta}{C_1 + C_{m_0}} \quad \textit{and} \quad \alpha_2 = \frac{(C_{m_0} - C_1 - 2\delta)\underline{\theta}_0}{(C_{m_0} - \delta)}.$$

Proof. For the inputs constructed in Theorem 10.13, after $N = \nu(n_0)N_1$ steps, each parameter bound can be updated N_1 times. In addition to Lemma 10.24, we have

$$\begin{aligned} \mathrm{Rad}_p(\Omega_N) &\leq n_0^{1/p}\mathrm{Rad}_\infty(\Omega_N) \\ &\leq n_0^{1/p}[(1-\alpha_2)\alpha_1^{N/\nu(n_0)}\mathrm{Rad}_\infty(\Omega_0) - \alpha_2] \\ &\leq n_0^{1/p}[(1-\alpha_2)\alpha_1^{N/\nu(n_0)}\mathrm{Rad}_p(\Omega_0) - \alpha_2] \\ &= n_0^{1/p}[(1-\alpha_2)\alpha_1^{N/\nu(n_0)}e_0 - \alpha_2]. \end{aligned}$$

So, (10.22) is true. □

Replace the l^p norm in Theorem 10.25 by the l^∞ norm. Then the following theorem can be derived.

Corollary 10.26. *For* (10.3), *let* $\overline{\theta}_0 = \max_{1\le i\le n_0} \overline{a}_i$*,* $\underline{\theta}_0 = \min_{1\le i\le n_0} \underline{a}_{is}$*,* $\sigma = \frac{m_0\Delta C+2\delta}{C_{m_0}+\Delta C-\delta}\overline{\theta}_0$*, and* $e_0 = \mathrm{Rad}_\infty(\Omega_0)$*. For any given* $e \in (\sigma_0, e)$*, we have*

$$N(e) \le \nu(n_0)\frac{\ln(\alpha_2+e)-\ln[(1-\alpha_2)e_0]}{\ln\alpha_1},$$

where α_1 *and* α_2 *were introduced in Theorem* 10.25*.*

10.5 Examples

Two examples are used to demonstrate the methods developed in this chapter. In Example 10.27, the parameters are estimated, and then the input is designed to track a target output. In Example 10.28, the difference between the estimation of quantized observations ($m_0 > 2$) and binary observations is discussed.

Example 10.27 (Tracking problem). For given target output y^*, design inputs such that $y_k = y^*$. It is worth mentioning that the y_k is measured by a quantized sensor with thresholds C_i, $i = 1, 2, \ldots, m_0$. For the classical tracking problem, the target output y^* must be equal to some threshold, and for $y^* \ne C_i$, $i = 1, 2, \ldots, m_0$, the tracking problem cannot be solved by classical methods. However, by using the results developed in this chapter, the unknown parameter can be estimated first, then the inputs can be designed to track y^*.

Consider

$$y_k = a_1u_k + a_2u_{k-2} + a_3u_{k-3} + d_k,\ \ k = 3, 4, \ldots,$$

where $d \le 1$; the real a_1, a_2, a_3 are 12, 10, 5, but unknown; the prior information is: $a_i \in [1, 70]$, $i = 1, 2, 3$, $u_{\max} = 30$, $y^* = 70$, and y_k is measured by a two-threshold sensor with $C_1 = 70$, $C_2 = 80$.

By Theorem 10.13, let

$$u = \{u_0, 0, 0, u_3, 0, u_5, 0, u_7, 0, u_9, 0, 0, u_8, \ldots\};$$

then

$$\begin{aligned}
y_{13i+3} &= a_3u_{13i}, & y_{13i+12} &= a_3u_{13i+9}, \quad i = 1, 2, \ldots,\\
y_{13i+4} &= a_1u_{13i+3}, & y_{13i+13} &= a_1u_{13i+12}, \quad i = 1, 2, \ldots,\\
y_{13i+7} &= a_2u_{13i+5}, & y_{13i+9} &= a_2u_{13i+7} \quad i = 1, 2, \ldots
\end{aligned}$$

Hence, the parameters are decoupled and then estimated. The estimate is aimed at reducing the unknown bound of each parameter to be less than 1 (see Figure 10.2).

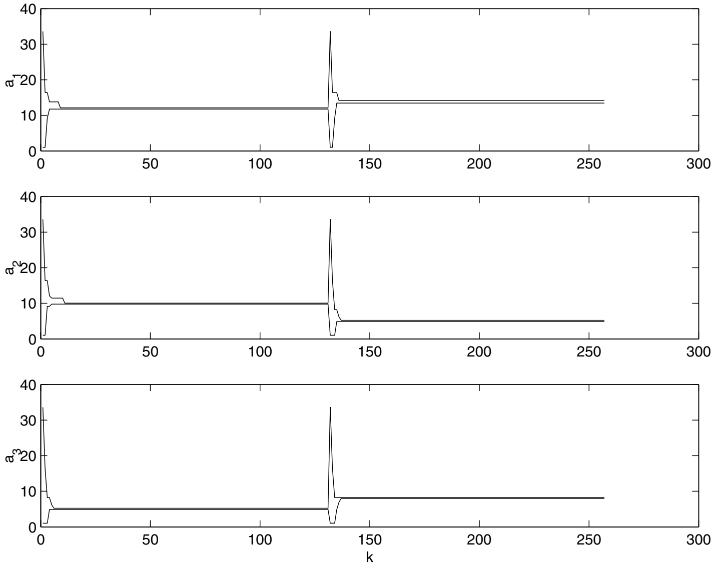

FIGURE 10.2. Reducing unknown parameter bound

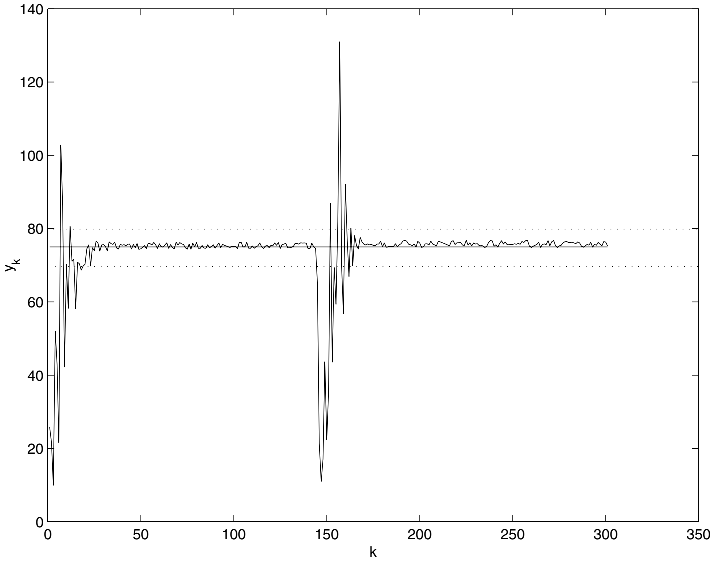

FIGURE 10.3. Tracking target output y^*

TABLE 10.1. Two-threshold sensor

Item	$k=0$	$k=1$	$k=2$	$k=3$	$k=4$
$\underline{a}_k$	1	1	1	8.26	11.35
$\overline{a}_k$	60	30.03	15.28	15.28	12.19
e_k	59	29.03	14.28	7.01	0.8
L_k	1	0.49	0.49	0.49	0.12

TABLE 10.2. Binary sensor

Item	$k=0$	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$
$\underline{a}_k$	1	1	1	8.45	8.45	10.31	11.23	11.69
$\overline{a}_k$	60	30.82	16.08	16.08	12.39	12.39	12.39	12.39
e_k	59	29.82	15.08	7.63	3.94	2.08	1.16	0.70
L_k	1	0.51	0.51	0.51	0.52	0.53	0.56	0.61

Consequently, consider the center of each unknown parameter interval $\widehat{a}_1, \widehat{a}_2, \widehat{a}_3$ as the estimate of a_1, a_2, a_3. For y^*, let

$$u_k = \frac{1}{\widehat{a}_1}(y^* - \widehat{a}_2 u_k - \widehat{a}_3 u_{k-2}).$$

At some k, change the parameters to $a_1 = 14, a_2 = 5, a_3 = 8$. Then y_k no longer tracks y_k. However, the quantized sensor can detect it and the parameters can be estimated again (Figures 10.2 and 10.3).

Example 10.28 (Comparison of estimations with quantized and binary sensors). Consider

$$y_k = au_k + d_k, \tag{10.23}$$

where the real parameter $a = 12$ but is unknown. The prior information is that $a \in [1, 60]$, $\delta = \|dl\|_\infty \leq 1$, and $u_{\max} = 30$. There are two-threshold sensors with $C_1 = 95$, $C_2 = 100$ and a binary sensor with $C = 95$. The estimation aim is to reduce the unknown parameter bound to be less than 1.

Then,

(i) increasing thresholds makes the estimation faster (Tables 10.1, 10.2, and Figure 10.4);

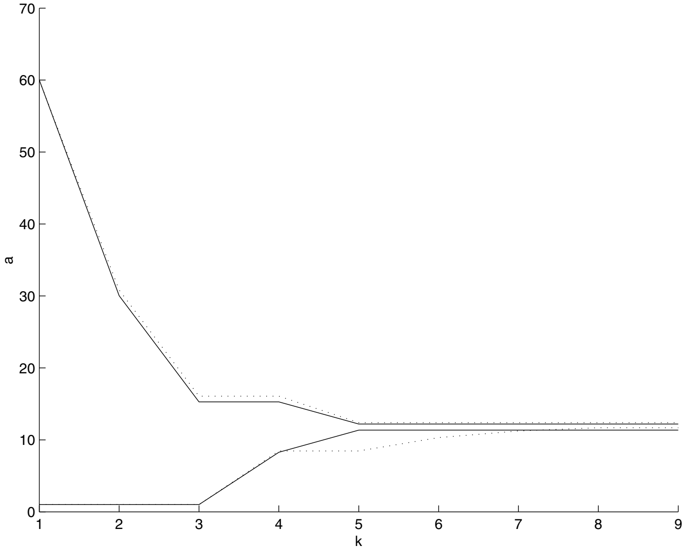

FIGURE 10.4. Comparison of estimations with two thresholds and binary sensors

(ii) from Table 10.1, we can find that $L_k < 1/m_0$, because the m_0 thresholds divided the parameter space into $m_0 + 1$ intervals, and L_k is decided by the longest interval in the worst case. However, for a given question, L_k may be decided by a smaller interval.

10.6 Notes

Based on the shortest decoupling input and analysis of gain system, the identification of an FIR system with quantized observations is studied.

In this chapter, the shortest decoupling input starting from k_0 is defined and designed without utilizing the information of inputs before k_0 and the length is $\nu(n_0)$. In some special cases assuming $u_k = 0$ for $k = k_0 - n_0 + 2, \ldots, k_0$, the shortest length of inputs to decouple each parameter once can be designed to be $n_0(n_0 + 1)/2$, which can be generalized to decouple each parameter p times with only $pn_0(n_0 + 1)/2$ (see [14]). For example, if $n_0 = 3$ and $u_k = 0$ for $k = -1, 0$, by input $u = \{u_1, u_2, 0, 0, u_5, 0\}$, one can get

$$y_2 = a_1 u_1, \quad y_5 = a_3 u_2, \quad y_7 = a_2 u_5.$$

That is, u is a three-parameter-decoupling input segment and the length is $n_0(n_0 + 1)/2 = 6$.

However, the condition that $u_k = 0$ for $k = k_0 - n_0 + 2, \ldots, k_0$ may not be reasonable. For example, after the input segments, in the above example, we have $k_0' = 6$, but the condition that $u_k = 0$ for $k = k_0' - 1, k_0'$ is not satisfied. The decoupling method developed in this chapter has a benefit of being independent of such conditions, and it is the optimal one in this case.