The Least Squares Algorithm, Parametric System Identification and Bounded Noise*

HÜSEYIN AKÇAY† and PRAMOD P. KHARGONEKAR‡

Key Words—Error analysis; identification; least squares estimation.

Abstract—The least squares parametric system identification algorithm is analyzed assuming that the noise is a bounded signal. A bound on the worst-case parameter estimation error is derived. This bound shows that the worst-case parameter estimation error decreases to zero as the bound on the noise is decreased to zero.

1. Introduction

The LEAST SQUARES ALGORITHM, due to Gauss, is one of the most widely used algorithms in science. It has been extensively studied and used for parametric system identification—see, for example, the book by Ljung (1987). It is very well known that the least squares algorithm enjoys certain optimality properties under suitable stochastic assumptions about exogeneous noise. In contrast, recent papers have taken a worst-case deterministic approach to identification. See the recent paper by Khargonekar (1993) for a discussion of this general area and an extensive list of references. In particular, our work is most closely related to the work on time-domain worst-case identification problems (Chen et al., 1992; Dahleh et al., 1992; Kacewicz and Milanese, 1992; Mäkilä, 1991; Poolla and Tikku, 1992; Tse et al., 1991; and the references cited in these papers).

Our work has grown out of a need to make connections between the classical identification theory and the more recent work in the area of robust identification. Towards this goal, in this paper we have investigated the performance of the least squares algorithm in the presence of worst-case bounded noise. In the result of this paper, we derive a bound on the worst-case parameter estimation error using the least squares algorithm in the presence of arbitrary bounded noise. This error bound shows that if the input is chosen to be a pseudorandom binary sequence, the worst-case parameter estimation error decreases to zero as the bound on the noise decreases to zero. In the terminology introduced by Helmicki et al. (1991), the least squares algorithm is robustly convergent. [We note that there is an important technical difference between the notion of robust convergence used in this paper and that in Helmicki et al. (1991).] While the problem formulation is motivated from the deterministic worst-case identification theory, the techniques employed in this paper draw upon the results in classical identification theory (Ljung, 1987). These results are similar in spirit to those in Wahlberg and Ljung (1992).

1.1. Notation. We end this introduction with some remarks on notation. Let $\mathcal{K}$ denote the Hardy space of bounded analytic functions in the open unit disk $D$ of the complex plane $C$. Let $G(z)$ denote the transfer function of a single-input/single-output (SISO), finite-dimensional linear time-invariant (FDLTI), discrete-time system. A transfer function $G$ is called stable if all its poles have magnitude strictly greater than one. Thus the transfer function of a stable LTI system belongs to $\mathcal{K}$. For any stable FDLTI discrete-time transfer function $G$, its $\mathcal{K}$ norm is defined as

$$\|G\|_{\mathcal{K}} := \max_{z \in D} |G(z)|.$$  

Let $D_p := \{z : |z| < p\}$, and let $\mathcal{K}(D_p, M)$ denote a subset of $\mathcal{K}$, defined as

$$\mathcal{K}(D_p, M) := \{G : G \text{ analytic in } D_p, |G(z)| = M, \forall z \in D_p\}.$$  

Thus if $p > 1$, the systems in $\mathcal{K}(D_p, M)$ are exponentially stable with a decay rate greater than $p$, and a gain bounded by $M$. Let $\mathcal{K}(n)$ denote the set of nth order proper, stable, real-rational transfer functions. Define the class of the systems

$$\mathcal{K}(n, m, p, M) := \{G : G \text{ analytic in } D_p, |G(z)| = M, \forall z \in D_p\},$$  

where $m \leq n$. Let $g$ be the impulse response of $G$. Then $l_1, l_2, l_3$ norms for $g$ are defined as

$$\|g\|_{L_1} := \sum_{k=1}^{\infty} |g(k)|, \quad \|g\|_{L_2} := \left( \sum_{k=1}^{\infty} |g(k)|^2 \right)^{1/2}, \quad \|g\|_{L_3} := \sup_{0 \leq k < \infty} |g(k)|.$$  

Given an $r \times s$ matrix $R$, its $\|\cdot\|_F$ norm is defined as

$$\|R\|_F := \max_{1 \leq k \leq r} \sum_{i=1}^{s} |R_{ik}|.$$  

For vectors, $\|\cdot\|_2$ will denote the Euclidean norm. Finally, $q^{-i}$ will denote the delay operator, i.e.,

$$(q^{-i}u)(t) := u(t-i).$$

2. Parametric system identification

We will consider the identification of an SISO, LTI, discrete-time system represented by the ARX model

$$A(q^{-1})y(t) = B(q^{-1})u(t) + v(t),$$

where

$$A(q^{-1}) = 1 + a_1 q^{-1} + \cdots + a_m q^{-m}, \quad B(q^{-1}) = b_1 q^{-1} + \cdots + b_n q^{-n}.$$
Let
\[ \theta^T = (a_1, a_2, \ldots, a_p, b_1, \ldots, b_m) \]
denote the vector of unknown parameters. Let
\[ \phi^T(t) = (-y(t-1), -y(t-2), \ldots, -y(t-n), u(t-1), \ldots, u(t-m)) \]
denote the regression vector of previous output and input signals. The model (2.1) can also be written
\[ y(t) = \theta^T \phi(t) + v(t). \]
Define the transfer function
\[ G(z) = \frac{b_1 z + \cdots + b_m z^m}{1 + a_1 z + \cdots + a_p z^p} = \frac{B(z)}{A(z)}. \]
We will assume that \( G \in \mathbb{R}(n, m, p, M) \).

The signal \( v(t) \) will be assumed to be bounded
\[ |v(t)| \leq \epsilon, \quad \forall t. \] (2.7)
The past inputs \( u(t), \ t \leq 0 \) will be assumed to be bounded, but otherwise unknown, i.e.
\[ \sup_{i<0} |u(i)| \leq C_1. \] (2.8)
This bound corresponds to an upper bound on the size of the initial state of the unknown system. For example, if it is known that the system is at rest at \( t = 0 \), we may take \( C_1 = 0 \).

The input sequence \( \{u(t), t > 0\} \) will be chosen by the experimenter and will only be required to satisfy an upper bound
\[ \sup_{i \geq 1} |u(i)| \leq C_2. \] (2.9)
The choice of the input sequence will be discussed in greater detail in the following sections.

Suppose that \( N + n \) data have been collected
\[ \mathcal{Z}^N := \{y(i), u(i) : i = 1, \ldots, N + n\}. \] (2.10)
An identification algorithm produces an estimate \( \hat{\theta}_N \) of \( \theta \) by using the prior information and the set of measured data (2.10). The performance of the identification algorithm is measured by the worst-case identification error defined as
\[ e_N(G, \epsilon) := \sup_{\|\theta - \hat{\theta}_N\|_2} \|y - \theta^T \phi\|. \] (2.11)
Note that in the definition of the worst-case identification error, we have not taken a supremum over the set of unknown systems as is done in Helmicki et al. (1991). The reason for this will become clear after the statement of the result. Following the terminology introduced in Helmicki et al. (1991), we call an identification algorithm convergent if
\[ \lim_{\epsilon \to 0} e_N(G, \epsilon) = 0. \] (2.12)
Moreover, the identification algorithm is said to be 'robustly convergent (and untuned)', if in addition the algorithm does not depend on the a priori information \( p \) and \( M \).

3. Pseudorandom binary sequences and the least squares identification algorithm

Given the above problem formulation, the well known least squares algorithm is a potential choice for the identification algorithm. It is well known that the least squares algorithm enjoys many optimality and convergence properties in stochastic formulations of the identification problem. However, relatively little is known about the performance of the least squares algorithm in the presence of bounded but otherwise arbitrary noise. In other words, relatively little seems to be known about the performance of the least squares algorithm in the worst-case deterministic problem setting described in the previous section. In this section, we will demonstrate that the least squares method is robustly convergent for the identification problem posed above for a special input design using pseudorandom binary sequences. This result follows from a bound on the worst-case identification error which is the main technical contribution of this paper.

First, we will briefly review some important properties of the Galois sequences. See Schroeder (1986) and references therein for more details.

3.1. Galois sequences. Let \( \sigma_k(x) = \sum_{i=0}^{p} c_i x^i \) be an irreducible polynomial over the binary field \( \{0, 1\} \) of degree \( p \) such that \( x \) is a primitive element. Define
\[ e_k \epsilon p = \sum_{i=0}^{p} c_k \epsilon i, \mod (2^p - 1), \quad k = 1, 2, \ldots, \] (3.1)
and
\[ e_1 = 1, \quad e_i = 0, \quad i = 2, 3, \ldots, p. \] (It is not necessary to choose \( e_i \), \( i = 2, 3, \ldots, p \) to be all zero; these can be chosen to be either zero or one.) The resulting sequence \( e_k \), \( k = 1, 2, \ldots \) is called a Galois sequence. A Galois sequence generated by an irreducible polynomial of degree \( p \) as above is periodic with period \( 2^p - 1 \). Using this periodicity property, the Galois sequence \( e_k \) can be extended for \( k = -1, -2, \ldots \). Irreducible polynomials \( x_k \) over the binary field \( \{0, 1\} \) up to degree \( p = 168 \), have been published (Stahnke, 1973). Using \( p = 168 \), we can generate a Galois sequence with period \( 2^{168} - 1 \) which is of the order of \( 10^{51} \). Each polynomial has the minimum number of terms possible (at most five) so that the linear recursion formula (3.1) is easy to implement.

Let us convert the Galois sequence \( e_k \) into another sequence \( s_k \) of integers according to the rule
\[ s_k := (-1)^{e_k}, \quad k = \pm 1, \pm 2, \ldots. \] (3.2)
A key property of this sequence is as follows. Define the correlation function of \( s_k \) as
\[ R_n := \sum_{k=0}^{2^n-2} s_k s_{k+n}. \] (3.3)
It turns out that (Schroeder, 1986)
\[ R_n = 2^n - 1, \quad n = 0 \mod (2^p - 1); \quad R_n = -1 \quad \text{otherwise.} \] (3.4)
Thus, for \( p \) sufficiently large, the ratio \( R_n/R_p \) is approximately zero for \( n \neq 0 \mod (2^p - 1) \). Hence the sequence \( s_k \) is approximately uncorrelated. (Except for the \( p \)-tuple of all 1s, this sequence contains all possible \( p \)-tuples of \( \pm 1 \).) The sequence \( s_k \) will be called a pseudorandom sequence. It should be noted that the sequence \( s_k \) depends on \( p \) as well as \( x_k \). For simplicity of notation, we will not show this dependence explicitly.

Galois sequences have found wide applications in communications and signal processing (Schroeder, 1986). Mäkilä (1991) has constructed a robustly convergent identification algorithm in \( l_1 \). Tse et al. (1991) have used pseudorandom binary signals in their identification studies. Pseudorandom binary signals have also been used in the study of time complexity of identification (Dahlke et al., 1992; Kacwicz and Milanese, 1992; Poolla and Tikku, 1992). In this paper, we will exploit the approximate pseudorandomness property for the analysis of worst-case performance of the least algorithm.

3.2. Main result—the worst-case performance of the least squares identification. In this subsection, we derive an upper bound on the worst-case identification error of the least squares algorithm for a special input. The length of the identification experiment will be taken to be \( N + n \), where \( N = 2^n - 1 \) for some positive integer \( p \) which will be fixed later. We choose the inputs \( \{u(t), t = 1, 2, \ldots, N + n\} \), to be a scaled version of the pseudorandom sequence
\[ u(t) = C s_t, \quad t = 1, 2, \ldots, N + n. \] (3.5)
From (2.1), we can write the system of equations
\[ Y_n = \Theta \Phi_n + V_n, \]
where
\[ Y_n := [y(n + 1)y(n + 2) \cdots y(n + N)]^T, \]
\[ \Phi_n := [\varphi(n + 1)\varphi(n + 2) \cdots \varphi(n + N)]^T, \]
\[ V_n := [v(n + 1)v(n + 2) \cdots v(n + N)]^T. \]

The usual least squares estimate of \( \theta \) given \( Y_n, \Phi_n \) is
\[ \hat{\theta}_n := (\Phi_n^T \Phi_n)^{-1} \Phi_n^T Y_n, \]
assuming that \( \Phi_n^T \Phi_n \) is invertible.

For the model (2.1) and the pseudorandom input \( u(t) \) described above, we obtain an upper bound on the worst-case error (2.11). This bound holds when the noise bound \( \varepsilon \) is small and the duration of the experiment \( N + n \) is large. The derivation is long but straightforward and is included in the Appendix.

First, we need some definitions. Let \( G \in \mathbb{R}_* \). Let \( n, m \) be fixed. For a vector \( x \in \mathbb{R}^{N + m} \), let
\[ X := \sum_{i=1}^{n} x_i e^{i(n-1)}, \quad Y := \sum_{i=1}^{m} x_i e^{i(n-1)}. \]

Now define the quadratic form
\[ Q(x) := C_2 x^T (G - Y)^2 \]
and set \( \lambda(G, C_2, n, m) \) by
\[ \lambda(G, C_2, n, m) := \min \{ Q(x) : x \in \mathbb{R}^{N + m}, \|x\|_2 = 1 \}. \]

In other words, \( \lambda(G, C_2, n, m) \) is the smallest eigenvalue of the quadratic form \( Q \). Define constants \( \varepsilon_0, T, K_1, K_2, K_3 \) as follows
\[ \varepsilon_0 := (M + 1)C_2(1 - \rho^{-1})^n, \]
\[ T := \max \left\{ \frac{8M^2C_2^2}{C_2^2(1 - \rho^{-1})^2} \right\}, \]
\[ K_1 := \frac{nC_2(2M + 1)}{(1 - \rho^{-1})^2}, \quad K_2 := \frac{3C_2(nM + 1)}{(1 - \rho^{-1})^n}, \]
\[ K_3 := \frac{(M + 1)(1 + \rho^{-1})^n + n^{-1/2}}{C_3}, \]
where \( C := C_1 + C_2 + C_3 \).

**Theorem 3.1.** Let \( G \in \mathbb{R}_*(n, m, \rho, M) \). Let \( \lambda(G, C_2, n, m) \), \( \varepsilon_0, T, K_1, K_2, K_3 \) be as defined in (3.13)-(3.16), respectively. Let \( N \rightarrow \infty \). Suppose the input \( u(t) \), \( t = 1, 2, \ldots, N + n \) is a pseudorandom binary sequence as given by (3.5). For all \( \varepsilon \leq \varepsilon_0 \), if \( \lambda > K_2(N^{-1} + K_2) \), the worst-case identification error (2.11) of the least squares algorithm (3.10) is bounded as
\[ e_0(G, \varepsilon) \leq \frac{K_1 K_2}{\lambda - (K_2(N^{-1} + K_2) \varepsilon)^n} \frac{1}{\lambda - (K_2(N^{-1} + K_2) \varepsilon)^n} K_3. \]

Consequently, the least squares is robustly convergent.

In contrast, using an ARX representation, the input length can possibly be reduced quite significantly.

* The upper bound on \( e_0(G, \varepsilon) \) depends on the quantity \( \lambda \). This dependence prevents us taking the supremum in 2.11 with respect to \( \mathbb{R}_*(n, m, \rho, M) \). If an \textit{a priori} lower bound for \( \lambda \) is available then this can be done and appropriate robust convergence results can be obtained.

A proof of this result is given in the Appendix.

**4. Conclusions**

In this paper, we have analyzed the worst-case performance of the least squares algorithm. We have obtained an upper bound on the worst-case parameter identification error which is used to establish appropriate convergence properties.

There are many different directions in which further research is necessary. We feel that suitable modifications of the least squares algorithm for the problem of parametric system identification in the presence of arbitrary but bounded noise may be very useful. Also, it is necessary to develop good methods that allow for undermodeling in this problem context.

**References**


**A. Appendix—Proofs**

We will prove the theorem in several steps. First, we need some definitions. Given two signals \( \{u(t) : -\infty < t \leq N + n\} \) and \( \{y(t) : -\infty < t \leq N + n\} \), we define the autocorrelation
and cross correlation functions as follows

\[
R_{uu}(k, l; n, N) := \frac{1}{N^{n+1}} \sum_{t=0}^{N} u(t-k)u(t-l), 0 \leq k, l \leq \infty
\] (A.1)

\[
R_{uy}(k, l; n, N) := \frac{1}{N^{n+1}} \sum_{t=0}^{N} u(t-k)y(t-l), 0 \leq k, l \leq \infty
\] (A.2)

\[
R_{uy}(k, l; n, N) := \frac{1}{N^{n+1}} \sum_{t=0}^{N} u(t-k)y(t-l), 0 \leq k, l \leq \infty
\] (A.3)

Let us define three matrices \(R_{uy}, R_{uu}, R_{uy}\) and two vectors \(R_{uy}, R_{uu}\) as follows

\[
R_{uy}(k, l) := R_{uy}(k, l; n, N), l \leq k, l \leq n;
\]

\[
R_{uu}(k, l) := R_{uu}(k, l; n, N), l \leq k, l \leq m;
\]

\[
R_{uy}(k, l) := R_{uy}(k, l; n, N), l \leq k \leq n, l \leq l \leq m;
\]

\[
R_{uy}(l) := R_{uy}(0, l; n, N), l \leq l \leq n;
\]

Notice that

\[
R_{u} := \left( \begin{array}{c}
R_{uy} \\
-\Psi_{u} \end{array} \right)_{n \times m} = \frac{1}{N} \Psi_{u} \Phi_{u}^{T};
\] (A.5)

**Lemma A.1.** Let the future inputs \(u(t); 1 \leq t < N + 1\) to a system \(G \in \mathcal{H}(D_{p}, M)\) be as in (3.5). Let the past inputs \(\{u(t); -\infty < t \leq 1\}\) to \(G\) be bounded by \(C_{1}\). Let \(y\) denote the output of \(G\) in response to the input \(u\). Then for \(1 \leq k, l \leq n\)

\[
R_{uy}(k, l; n, N) = C_{2} g(l - k) + AR_{uy}(k, l; n, N),
\] (A.6)

where

\[
|A|_{R_{uy}}(k + s, l; n, N) := \sum_{t=1}^{N} g(s)g(k - l + s),
\]

\[
|A|_{R_{uy}}(k, l; n, N) := \sum_{t=1}^{N} g(s)g(k - l + s).
\]

**Proof.** Let \(\bar{u}\) be periodic extension of the inputs in (3.5). Let \(\bar{u} = u - \bar{u}\). Notice that \(\bar{u}(t) = 0, 1 \leq t \leq 1\) and \(\|\bar{u}(t)\|_{\infty} \leq C_{1} + C_{2}\). Then, we have for \(1 \leq k, l \leq n\)

\[
R_{uy}(k + s, l; n, N) \leq C_{2}^{2} + \Delta R_{uy}(k + s, l; n, N)
\]

\[
R_{uy}(k + s, l + s; n, n) = C_{2}^{2} + \Delta R_{uy}(k + s, l + s; n, n),
\] (A.6)

where

\[
|\Delta R_{uy}(k + s, l; n, N)\|_{N} \leq C_{2}^{2},
\]

\[
|\Delta R_{uy}(k + s, l + s; n, n)\|_{N} \leq C_{2}^{2} + \sum_{t=1}^{N} g(s)g(k - l + s).
\]

Moreover

\[
|\Delta R_{uy}(k, l; n, N)\|_{N} \leq C_{2}^{2} + \sum_{t=1}^{N} g(s)g(k - l + s),
\] (A.7)

Since \(G \in \mathcal{H}(D_{p}, M)\), by an application of Cauchy's formula \(G' \in \mathcal{H}(D_{p}, M'(p - 1))\). Hence

\[
|\bar{u}(t)|_{\infty} \leq \frac{M}{p - 1},
\] (A.19)

\[
\sum_{t=1}^{N} s g(t) = \|G(t)\|_{1} \leq \frac{M_{p}}{(p - 1)^{2}},
\] (A.20)

Therefore for \(1 \leq k, l \leq n\)

\[
|\Delta R_{uy}(k, l; n, N)\|_{N} \leq \frac{C_{2}^{2}}{N(p - 1)} + \frac{M_{p} C_{2} C}{(p - 1)^{2}},
\] (A.21)

\[
|\Delta R_{uy}(k, l; n, N)\|_{N} \leq \frac{M_{p} C_{2}^{2}}{(p - 1)^{2}} + \frac{3 C_{2}^{2} M_{p}^{2} p}{(p - 1)^{3}},
\] (A.22)
Corollary A.2. Consider the system given by (2.1) and let $G \in \mathcal{H}(l_{\infty}, M)$. Let the future inputs $(u(t): 1 \leq t < N + n)$ to the system be as in (3.5). Let the past inputs $(u(t): -\infty < t < 0)$ to $G$ be bounded by $C_1$. Let the noise $v$ be bounded by $\phi$. Let $y$ denote the output of the system, in response to the input $u$. Then for $1 \leq k, l < n$; $R_{u u}, R_{u y}, R_{y y}$ are given by (A.6), where

$$
\|AR_{u u}(k, l; n, N)\| \leq \frac{C_2}{N},
$$

(A.23)

$$
\|AR_{u y}(k, l; n, N)\| \leq \frac{MC_2(2p - 1)}{N(p - 1)^2} + \frac{C_2}{(1 - \rho^{-1})^2},
$$

(A.24)

$$
\|AR_{y y}(k, l; n, N)\| \leq \frac{M^2C_2^2(4p - 1)}{(N(p - 1))^2} + \frac{C_2}{(1 - \rho^{-1})^2} + \frac{M^2C_2^2}{2N} + \frac{C_2}{(1 - \rho^{-1})^2}.
$$

(A.25)

Proof. (2.1) can be rewritten as

$$
y = \hat{y} + \eta,
$$

(A.26)

where

$$
\eta := A^{-1}v, \eta(t) = 0, t \geq 0; \quad \hat{y} = Gu.
$$

(A.27)

Let $u = u_1 + u_2$ be a decomposition of $u$ into past inputs and future inputs, i.e.

$$
u_1(t) = u(t), t \leq 0
$$

and

$$
u_2(t) = 0, t > 0;
$$

(A.28)

$$
u_2(t) = 0, t < 0; \quad u_2(t) = u(t), t > 0.
$$

(A.29)

Let $y_1 := Gu_1$ and $y_2 := Gu_2$. Then easy calculations show that

$$
\sum_{t = n + 1}^{N+n} \eta(t - k)2^{1/2} \leq \sqrt{N} \|A^{-1}\|_{\infty, \epsilon}, 1 \leq k \leq n.
$$

(A.30)

We can write

$$
R_{u u}(k, l; n, N) = R_{u u}(k, l; n, N) + R_{u y}(k, l; n, N),
$$

(A.31)

$$
R_{u y}(k, l; n, N) = R_{y y}(k, l; n, N) + R_{u y}(k, l; n, N) + R_{y y}(k, l; n, N) + R_{y u}(k, l; n, N),
$$

(A.32)

where $R_{u u}$ and $R_{y y}$ are given by (A.6). Now an application of Schwarz inequality results in

$$
\|AR_{u u}(k, l; n, N)\| \leq \sqrt{N} \|A^{-1}\|_{\infty, \epsilon}, 1 \leq k \leq n.
$$

(A.33)

Finally, since $G$ is analytic in $D_\rho$, $A(z)$ does not have zeros in $D_\rho$. Therefore $A(z)$ can be factored as

$$
A(z) = \prod_{i=1}^{n} (1 + p_i z^{-1}).
$$

(A.35)

where $p_i, i = 1, \ldots, n$ are the poles of $G$. These observations lead to

$$
(1 - p^{-1})^n \leq \|A\|_{\infty, \epsilon} \leq (1 + p^{-1})^n.
$$

(A.36)

As a consequence of Corollary A.2, we may define a matrix $R$ and a vector $\Psi$

$$
R := \lim_{N \to \infty} R_{N}; \Psi := \lim_{N \to \infty} \Psi_{N}.
$$

(A.37)

Let $\Delta R_{N} := R_{N} - R$ and $\Delta \Psi_{N} := \Psi_{N} - \Psi_{N}$. The frequency-domain expression of $R$ allows us to show that the smallest eigenvalue of $R$ is $\lambda_l(G, C_2, n, m)$ and the following fact.

Fact A.3. Let $R$ be as in A.37 for a system $G \in \mathcal{H}(n)$. Then $R$ is invertible. If $G \in \mathcal{H}(n, m)$, then $\lim_{\epsilon \to 0, N \to \infty} \hat{\theta}_N = \theta$.

Proof of Theorem 3.1. We only need to derive the error bounds. According to Theorem 1.18 in Grenander and Szego (1958), the largest eigenvalue $\lambda_{\max}(\Delta R_{N})$ of $\Delta R_{N}$ is bounded above by $\|\Delta R_{N}\|_{\infty, \epsilon}$. Hence if $\lambda > \lambda_{\max}(\Delta R_{N})$ for some $\lambda$, the following decomposition

$$
\|\Delta R_{N}\|_{\infty, \epsilon} \leq \lambda \|\Delta R_{N}\|_{\infty, \epsilon}
$$

(A.38)

we get

$$
\|\hat{\theta}_N - \theta\|_{\infty, \epsilon} \leq \|\Delta \Psi_{N}\|_{\infty, \epsilon} + \|\theta\|_{\infty, \epsilon} \|\Delta R_{N}\|_{\infty, \epsilon}, \lambda \|\Delta R_{N}\|_{\infty, \epsilon}.
$$

(A.39)

From Corollary A.2, since $m \leq n$ and $C_2 \leq C$, we have

$$
\|\Delta R_{y y}\|_{\infty, \epsilon} + \|\Delta R_{y u}\|_{\infty, \epsilon} \leq \left\{\begin{array}{ll}
M^2 \rho - 1 + 2\rho - 1 + \rho^{-1} + 1 & \frac{M^2C_2^2}{N(p - 1)^2} \\
(2M + 1 + M)\rho & \frac{C_2}{N(1 - \rho^{-1})^3}
\end{array}\right\} + \left(\frac{2M + 1 + M}{N}\right)^{1/2} + \left(\frac{2M + 1 + M}{N}\right)^{1/2} \frac{C_2}{N(1 - \rho^{-1})^3}.
$$

(A.40)

Thus

$$
\|\Delta R_{u u}\|_{\infty, \epsilon} + \|\Delta R_{u y}\|_{\infty, \epsilon} \leq \left\{\begin{array}{ll}
M^2 \rho - 1 + 2\rho - 1 + \rho^{-1} + 1 & \frac{M^2C_2^2}{N(p - 1)^2} \\
(2M + 1 + M)\rho & \frac{C_2}{N(1 - \rho^{-1})^3}
\end{array}\right\} + \left(\frac{2M + 1 + M}{N}\right)^{1/2} + \left(\frac{2M + 1 + M}{N}\right)^{1/2} \frac{C_2}{N(1 - \rho^{-1})^3}.
$$

(A.41)

Let $T := \max \{NMC_1/C_2^2(1 - \rho^{-1})^{3}, 2M\}$. If $N \geq T$, then $\sqrt{2N} \|\Delta R_{N}\|_{\infty, \epsilon} \leq C_2$. Let $\epsilon_0 := (M + 1)C_2(1 - \rho^{-1})^3$. Then, for $N \geq T$ and $\epsilon \leq \epsilon_0$, we have

$$
\left\{\begin{array}{ll}
M^2 \rho - 1 + 2\rho - 1 + \rho^{-1} + 1 & \frac{M^2C_2^2}{N(p - 1)^2} \\
(2M + 1 + M)\rho & \frac{C_2}{N(1 - \rho^{-1})^3}
\end{array}\right\} + \left(\frac{2M + 1 + M}{N}\right)^{1/2} + \left(\frac{2M + 1 + M}{N}\right)^{1/2} \frac{C_2}{N(1 - \rho^{-1})^3}.
$$

(A.42)

Since $\rho > 1$, the following inequalities hold

$$
\left\{\begin{array}{ll}
M^2 \rho - 1 + 2\rho - 1 + \rho^{-1} + 1 & \frac{M^2C_2^2}{N(p - 1)^2} \\
(2M + 1 + M)\rho & \frac{C_2}{N(1 - \rho^{-1})^3}
\end{array}\right\} + \left(\frac{2M + 1 + M}{N}\right)^{1/2} + \left(\frac{2M + 1 + M}{N}\right)^{1/2} \frac{C_2}{N(1 - \rho^{-1})^3}.
$$

(A.43)
Then, from the inequalities (A.40)-(A.44), an upper bound on \( \| \Delta R_n \|_1 \) is obtained as

\[
\| \Delta R_n \|_1 \leq \lambda_0 := \frac{nC^2(2M + 1)^2}{N(1 - \rho^{-1})^2} + \frac{3(M + 1)C_2n}{(1 - \rho^{-1})^p} \epsilon. \tag{A.45}
\]

Since the upper bounds on \( \Delta R_{mn}(k, l; n, N) \), \( \Delta R_{mp}(k, l; n, N) \), \( \Delta R_{pp}(k, l; n, N) \) given in Corollary A.2 are uniform in \( k \) and \( l \) in the range of interest, an upper bound on \( \| \Delta \Psi_n \|_2 \) is obtained as follows

\[
\| \Delta \Psi_n \|_2 \leq n^{-1/2} \left( \frac{nC^2(2M + 1)^2}{N(1 - \rho^{-1})^2} + \frac{3(M + 1)C_2n}{(1 - \rho^{-1})^p} \epsilon \right) \leq n^{-1/2} \lambda_0. \tag{A.46}
\]

Finally, we have the following estimate on \( \| \theta \|_2 \)

\[
\| \theta \|_2 \leq (\| A \|_\infty - 1 + \| B \|_\infty^{1/2}) = (M + 1) \| A \|_\infty \leq (M + 1)(1 + \rho^{-1})^p. \tag{A.47}
\]

Using (A.45), (A.46), (A.47), the desired error bound follows.