On the Worst-Case Divergence of the Least-Squares Algorithm

Hüseyin Akçay [a,1] and Brett Ninness [a,2]

[a] *Centre for Integrated Dynamics and Control (CIDAC) and Department of Electrical and Computer Engineering, University of Newcastle, Callaghan, NSW 2308, Australia*

Abstract

In this paper, we provide a $\mathcal{H}_\infty$–norm lower bound on the worst–case identification error of least–squares estimation when using FIR model structures. This bound increases as a logarithmic function of model complexity and is valid for a wide class of inputs characterized as being quasi–stationary with covariance function falling off sufficiently quickly.

Key words: Least–squares, identification in $\mathcal{H}_\infty$, time–domain data, divergence.

Technical Report EE9765, Centre for Integrated Dynamics and Control (CIDAC), Department of Electrical and Computer Engineering, University of Newcastle,AUSTRALIA

1 Introduction

The least–squares algorithm, due to Gauss, has been extensively studied and used in system identification and under certain stochastic assumptions on the disturbances it has been shown to enjoy various optimality properties [9,22].

Recently however, many researchers have considered using deterministic 'worst–case' noise descriptions; see for example [21,14,16] for surveys of this area and

[1] On leave from TÜBİTAK, Marmara Research Center, Division of Mathematics, P.0. Box 21, Gebze–KOCAELİ, Turkey. This author gratefully acknowledges support for this work from TÜBİTAK and CIDAC.

[2] This author gratefully acknowledges the support of CIDAC and the Australian Research Council.

lists of references. Since the least–squares method is in very widespread and successful use, it is of interest to analyze its behavior under worst–case assumptions and indeed some literature already exists on this topic [1–4,17–19]. The purpose of the current paper is to pursue this analysis further and to analyze the dependence of the estimation error on the model complexity. To simplify the exposition, we shall restrict our attention to finite–impulse response models.

In a stochastic setting, and under suitable conditions on the unknown system, noise, and input, an infinite-dimensional system can be recovered asymptotically as the size of data and model order grows [10,9]. A key condition required for this property to hold is that the input signal must be persistently exciting of infinite order.

In a deterministic worst-case setting, when the unknown system is finite-dimensional, it is known that its least-squares estimate converges to the true system as noise amplitude decreases to zero and the number of data tends to infinity [1]. This property has been called *robust convergence* in the terminology introduced in [7]. On the other hand, in [2] it is shown that if the input is chosen to be a pseudo random binary sequence (PRBS), then the worst–case identification error of the least–squares algorithm diverges as $O(\sqrt{n})\,\epsilon$, where n is the model order and ϵ is an upper bound on measurement corruptions.

Therefore, for a particular highly specialised input, the popular least-squares algorithm is divergent in a worst-case setting. The main result of this paper is to show that this same divergence occurs for a much broader class of inputs than the specialised PRBS one, and hence this paper extends previous results on this topic [2].

The worst case result presented here is with respect to the frequency domain $\mathcal{H}_\infty$ supremum norm. However, in the context of worst-case identification it is pertinent to observe that the time–domain ℓ_1 and ℓ_2–norms have drawn more attention than the $\mathcal{H}_\infty$–norm although the latter is no less important owing to its use in robust control. This concentration on time domain formulation is mostly due to the complexity of the ensuing analysis. More specifically, it should be appreciated that particular difficulties arise when working with the $\mathcal{H}_\infty$–norm since this choice involves measuring the performance of the identification algorithm in terms of a non–quadratic function (the $\mathcal{H}_\infty$ norm) while the free variables (noise and input sequence) are ℓ_∞ constrained.

As a result, and in contrast to the worst–case identification in $\mathcal{H}_\infty$ using time-domain data setting of this paper, the performance, sample complexity, and input design issues of ℓ_1 and ℓ_2 worst–case identification are by now well understood and a variety of results are available in the literature [4,5,8,11,13,15,17–20].

2 Divergence Result

Consider the identification of a single–input/single–output, linear–time invariant, discrete–time system represented by a finite impulse response model

$$y(t) = \sum_{k=1}^{n} g(k)u(t-k) + v(t) \tag{1}$$

where $u(t)$ is the input sequence and $v(t)$ is a bounded disturbance

$$|v(t)| \leq \epsilon < \infty, \quad \forall t. \tag{2}$$

The order of the system, n is assumed to be known and finite and the inputs are assumed to be bounded as

$$|u(t)| \leq C_u < \infty. \tag{3}$$

Let

$$U_N \triangleq \frac{1}{\sqrt{N}} \begin{bmatrix} u(n) & \cdots & u(1) \\ \vdots & \ddots & \vdots \\ u(N+n-1) & \cdots & u(N) \end{bmatrix},$$

$$Y_N \triangleq \frac{1}{\sqrt{N}} \begin{bmatrix} y(n+1) & \cdots & y(N+n) \end{bmatrix}^T$$

where Y_N^T denotes transpose of Y_N. Then assuming U_N has full column rank, the least–squares estimates of impulse–response coefficients in (1) are calculated as

$$\widehat{g}_N = (U_N^T U_N)^{-1} U_N^T Y_N. \tag{4}$$

Let $\widetilde{g}_N \triangleq g - \widehat{g}_N$ denote the error, which can be written as

$$\widetilde{g}_N = (U_N^T U_N)^{-1} U_N^T V_N$$

where

$$V_N \triangleq \frac{1}{\sqrt{N}} \begin{bmatrix} v(n+1) & \cdots & v(N+n) \end{bmatrix}^T$$

and let

$$\widetilde{G}_N(e^{j\omega}) \triangleq \sum_{k=1}^{n} \widetilde{g}_N(k) e^{-j\omega k}, \ -\pi \leq \omega \leq \pi$$

denote the error transfer function. The $\mathcal{H}_\infty$–norm of $\widetilde{G}_N$ is defined by

$$\|\widetilde{G}_N\|_\infty \triangleq \max_{|\omega|\leq\pi} |\widetilde{G}_N(e^{j\omega})|$$

which allows the further definition of the worst-case identification error as

$$e_N \triangleq \sup_{|v(t)|\leq\epsilon} \|\widetilde{G}_N\|_\infty. \tag{5}$$

The main result of this paper is the derivation of a lower bound on e_N and which applies for a much broader class of input signals than originally considered in [2] where the only other similar result appears. The nature of the lower bound derived here is such as to tend to infinity with increasing n, so that worst case divergence of the least-squares algorithm is implied.

To proceed, a key tool is Lemma 2 which depends on the following classical result [6] used in Fourier analysis literature for estimating so-called Lebesgue constants.

Fact 1 *(Hardy's inequality) Let c_k be arbitrary complex numbers. Then*

$$\int_0^{2\pi} \left| \sum_{k=1}^{n} c_k e^{-jk\omega} \right| \mathrm{d}\omega \geq 2 \sum_{k=1}^{n} \frac{|c_k|}{k+1}. \tag{6}$$

Lemma 2 *Let e_N be as in (5). Then*

$$e_N \geq \epsilon \sum_{k=1}^{n} \frac{1}{k+1} \frac{1}{N} \sum_{t=n+1}^{N+n} \left| \sum_{l=1}^{n} S_N(k,l) u(t-l) \right|. \tag{7}$$

where $S_N = (U_N^T U_N)^{-1}$.

PROOF. Let

$$P_{N,t}(z) = \sum_{k=1}^{n} \sum_{\ell=1}^{n} S_N(k,\ell) u(t-\ell) z^k \tag{8}$$

Interchanging the order of summation, $\widetilde{G}_N(e^{j\omega})$ can be written as

$$\widetilde{G}_N(e^{j\omega}) = \frac{1}{N} \sum_{t=n+1}^{N+n} v(t) P_{N,t}(e^{-j\omega}). \tag{9}$$

Let $\mathcal{R}$ and $\mathcal{I}$ denote respectively real and imaginary parts. Fix θ and let

$$\eta(t) = \epsilon \ \mathrm{sign} \left(\mathcal{R} \left[P_{N,t}(e^{-j\theta}) \right] \right), \quad \forall t \tag{10}$$

where sign(x) is the real valued function defined as sign$(x) = x/|x|$ for $x \neq 0$ and sign$(0) = 0$. Then $\eta(t)$ is admissible noise and we obtain from (9) and (10)

$$\begin{aligned}\sup_{\|v\|_\infty \leq \epsilon} |\widetilde{G}_N(e^{j\theta})| &= \sup_{\|v\|_\infty \leq \epsilon} \frac{1}{N} \left| \sum_{t=n+1}^{N+n} v(t) P_{N,t}(e^{-j\theta}) \right| \\ &\geq \frac{1}{N} \left| \sum_{t=n+1}^{N+n} \eta(t) P_{N,t}(e^{-j\theta}) \right| \\ &\geq \frac{\epsilon}{N} \sum_{t=n+1}^{N+n} \left| \mathcal{R}\left[P_{N,t}(e^{-j\theta}) \right] \right|. \end{aligned} \tag{11}$$

Letting now $\eta(t) = \epsilon$ sign $\left(\mathcal{I}\left[P_{N,t}(e^{-j\theta}) \right] \right)$ for all t, use of (9) provides

$$\sup_{\|v\|_\infty \leq \epsilon} |\widetilde{G}_N(e^{j\theta})| \geq \frac{\epsilon}{N} \sum_{t=n+1}^{N+n} \left| \mathcal{I}\left[P_{N,t}(e^{-j\theta}) \right] \right|. \tag{12}$$

Thus it follows from (11) and (12) that for each fixed θ

$$\begin{aligned} e_N &= \sup_{\|v\| \leq \epsilon} \max_{|\theta| \leq \pi} |\widetilde{G}(e^{j\theta})| \\ &\geq \sup_{\|v\|_\infty \leq \epsilon} |\widetilde{G}_N(e^{j\theta})| \\ &\geq \frac{\epsilon}{2N} \sum_{t=n+1}^{N+n} \left(\left| \mathcal{R}\left[P_{N,t}(e^{-j\theta}) \right] \right| + \left| \mathcal{I}\left[P_{N,t}(e^{-j\theta}) \right] \right| \right). \\ &\geq \frac{\epsilon}{2N} \sum_{t=n+1}^{N+n} \left| P_{N,t}(e^{-j\theta}) \right|. \end{aligned} \tag{13}$$

Integrating (13) with respect to θ and using Fact 1 for (8) allows (7) to be obtained as follows

$$\begin{aligned} e_N &\geq \frac{\epsilon}{2N} \sum_{t=n+1}^{N+n} \frac{1}{2\pi} \int_0^{2\pi} \left| P_{N,t}(e^{-j\theta}) \right| \, d\theta \\ &\geq \frac{\epsilon}{N} \sum_{t=n+1}^{N+n} \sum_{k=1}^{n} \frac{1}{k+1} \left| \sum_{\ell=1}^{n} S_N(k,\ell) u(t-\ell) \right| \\ &= \epsilon \sum_{k=1}^{n} \frac{1}{k+1} \frac{1}{N} \sum_{t=n+1}^{N+n} \left| \sum_{\ell=1}^{n} S_N(k,\ell) u(t-\ell) \right|. \end{aligned}$$

□

From now on, it will be assumed that $u(t)$ is quasi–stationary in the sense of [9] or, equivalently, amenable to the 'generalised harmonic analysis' of Wiener [23]. This implies that $R_N = U_N^T U_N$ converges to a symmetric Toeplitz autocorrelation matrix

$$R = \begin{pmatrix} R(0) & \cdots & R(n-1) \\ \vdots & \ddots & \vdots \\ R(n-1) & \cdots & R(0) \end{pmatrix}. \tag{14}$$

The associated (power) spectrum of $u(t)$ is defined as

$$\Phi(\omega) = \sum_{\tau=-\infty}^{\infty} R(\tau) e^{-j\tau\omega} \tag{15}$$

assuming that $R(\tau)$ decays sufficiently quickly for the infinite sum to exist.

It is well known that for finite dimensional systems [12] and under mild stochastic assumptions on the disturbance that involve it being second order stationary with covariance function decaying sufficiently quickly then $\|\tilde{G}_N\|_\infty \to 0$ w.p.1 as $N \to \infty$ if the input is quasi–stationary and sufficiently rich. This conclusion can be extended to strictly stable infinite–dimensional linear systems if the model order is allowed to monotonically increase to infinity at a suitable rate relative to the growth in available data [10]. A common theme of these results is the assumption that R in (14) is positive definite for all n, which is equivalent to $\Phi(\omega) > 0$.

The autocorrelation matrix R in (14) is positive definite for all n provided $R(0) > \sum_{\tau=1}^{\infty} |R(\tau)|$. To see this, let $\sigma_1(R_N) \geq \cdots \geq \sigma_n(R_N)$ denote singular values of R_N. Then the largest singular value of R_N can be bounded above as

$$\sigma_1(R_N) \leq \max_{1 \leq k \leq n} \sum_{\ell=1}^{n} |R_N(k,\ell)|. \tag{16}$$

Now if R_N is split as $R_N = \overline{R}_N + \tilde{R}_N$, where $\overline{R}_N$ is the diagonal matrix defined by $\overline{R}_N(k,k) = R_N(k,k)$ and $\tilde{R}_N$ is perturbation matrix containing only off-diagonal entries of R_N, then the smallest singular value of R_N may be bounded below as

$$\sigma_n(R_N) \geq \sigma_n(\overline{R}_N) - \sigma_1(\tilde{R}_N). \tag{17}$$

Thus from (17) and (16), the following inequality is obtained

$$\sigma_n(R_N) \geq \min_{1\leq k\leq n} R_N(k,k) - \max_{1\leq k\leq n} \sum_{\ell\neq k}^{n} |R_N(k,\ell)| \tag{18}$$

which as $N \to \infty$ converges by the quasi-stationarity assumption to

$$\sigma_n(R) \geq R(0) - \sum_{\tau=1}^{n} |R(\tau)|.$$

Therefore the assumption $R(0) > \sum_{\tau=1}^{\infty} |R(\tau)|$ ensures $\sigma_n(R) > 0$ for all n.

In the derivation of the main result, the following stronger condition will in fact be required

$$R(0) > \sum_{\tau=1}^{\infty} |R(\tau)| + C_u \left(\sum_{\tau=1}^{\infty} |R(\tau)| \right)^{1/2}. \tag{19}$$

This is satisfied by a large class of inputs including white-noise inputs and the pseudo random binary sequences used in [2]. As well, a coloured input generated according to

$$u(t) = s(t) + H(q)p(t) \tag{20}$$

where $s(t)$ and $p(t)$ are zero mean, uncorrelated, white-noise sequences with unit variance and $H(q)$ is an exponentially stable transfer function such that $\|H(e^{j\omega})\|_2 \ll 1$ can be shown to satisfy (19).

However, it is acknowledged that a limitation of the main result in Theorem 3 is that the requirement (19) cannot be weakened. Nevertheless, Theorem 3 constitutes a significant advance over previous work where the conditions [2] were (by assuming PRBS input) much stronger. As well, the fact that the class of inputs may be broadened at all from the PRBS case considered in [2] indicates a more fundamental nature of the worst case divergence than merely being a pathology confined to a particular input.

Given this preliminary comments, the main result of the paper is as follows.

Theorem 3 *Let $u(t)$ be quasi–stationary deterministic signal satisfying (19) and e_N be as in (5). Then for some absolute constant C*

$$\liminf_{N\to\infty} e_N \geq C \log n \ \epsilon. \tag{21}$$

PROOF. First by the triangle inequality,

$$\frac{1}{N}\sum_{t=n+1}^{N+n}\left|\sum_{l=1}^{n} S_N(k,l)u(t-l)\right| \geq \quad S_N(k,k)\ \frac{1}{N}\sum_{t=n+1}^{N+n}|u(t-k)| - \frac{1}{N}\sum_{t=n+1}^{N+n}\left|\sum_{\ell\neq k} S_N(k,l)u(t-\ell)\right|, \tag{22}$$

A lower bound on the first term of the right hand side of (22) can be derived from the following inequality

$$R_N(k,k) = \frac{1}{N}\sum_{t=n+1}^{N+n} u^2(t-k) \leq C_u \frac{1}{N}\sum_{t=n+1}^{N+n}|u(t-k)| \tag{23}$$

and an upper bound on the second term of the right hand side of (22) is obtained as

$$\begin{aligned}\frac{1}{N}\sum_{t=n+1}^{N+n}\left|\sum_{\ell\neq k} S_N(k,\ell)u(t-\ell)\right| &\leq \left[\frac{1}{N}\sum_{t=n+1}^{N+n}\left(\sum_{\ell\neq k} S_N(k,\ell)u(t-\ell)\right)^2\right]^{1/2} \\ &= \left(\sum_{\ell\neq k}\sum_{s\neq k} S_N(k,\ell)S_N(k,s)R_N(\ell,s)\right)^{1/2} \\ &= S_N^{1/2}(k,k)\ \left(S_N(k,k)R_N(k,k)-1\right)^{1/2} \end{aligned} \tag{24}$$

where the inequality in (24) holds by the Cauchy–Schwartz inequality and the last equality by the fact that R_N is symmetric and $S_N = R_N^{-1}$. Thus from (22), (23), and (24)

$$\frac{1}{N}\sum_{t=n+1}^{N+n}\left|\sum_{\ell=1}^{n} S_N(k,\ell)u(t-\ell)\right| \geq \frac{1}{C_u}S_N(k,k)R_N(k,k) - S_N^{1/2}(k,k)\ \left(S_N(k,k)R_N(k,k)-1\right)^{1/2}. \tag{25}$$

Now, split R_N as $R_N = \overline{R}_N + \tilde{R}_N$ as in the discussion preceding the theorem. Notice that $S_N(k,k)R_N(k,k) - 1$ equals the k'th diagonal entry of $-R_N^{-1}\tilde{R}_N$. Hence it is bounded in magnitude by $\sigma_1(R_N^{-1}\tilde{R}_N)$ which is in turn over–bounded by $\sigma_n^{-1}(R_N)\sigma_1(\tilde{R}_N)$. Since R_N converges to R it can be asserted that for all k

$$\limsup_{N\to\infty}\left[R_N(k,k)S_N(k,k)-1\right] \leq \left(R(0) - \sum_{\tau=1}^{\infty}|R(\tau)|\right)^{-1}\sum_{\tau=1}^{\infty}|R(\tau)|, \tag{26}$$

which implies

$$\limsup_{N\to\infty} S_N(k,k) \le \left(R(0) - \sum_{\tau=1}^{\infty} |R(\tau)| \right)^{-1}. \tag{27}$$

Thus from (25)–(27), for all k

$$\begin{aligned} \liminf_{N\to\infty} \quad & \frac{1}{N} \sum_{t=n+1}^{N+n} \left| \sum_{\ell=1}^{n} S_N(k,\ell) u(t-\ell) \right| \\ & \ge \frac{R(0) - \sum_{\tau=1}^{\infty} |R(\tau)| - C_u \left(\sum_{\tau=1}^{\infty} |R(\tau)| \right)^{1/2}}{C_u \left(R(0) - \sum_{\tau=1}^{\infty} |R(\tau)| \right)}. \end{aligned} \tag{28}$$

Let γ denote the right hand side of (28). Then from Lemma 2

$$e_N \ge \gamma \sum_{k=1}^{n} \frac{1}{k+1} \epsilon = O(\log n)\, \epsilon.$$

This completes the proof. □

Theorem 3 shows that when the inputs are bounded quasi-stationary subject to the requirement (19), then the worst-case supremum norm frequency domain error e_N diverges as model order is increased. It remains an open problem whether the lower bound in (21) is attained for some input signal in the class (19) or even in the largest class $\Phi(\omega) > 0$. Under the same input assumptions, however it is easy to show that e_N is bounded above by a term $O(\sqrt{n})\ \epsilon$.

3 Conclusion

The purpose of this paper was to illustrate that for a class of inputs that are persistently exciting, the least squares algorithm is worst-case divergent at a rate logarithmic in model order n and with respect to the $\mathcal{H}_\infty$ norm error of transfer function estimation. This represents an extension of previous results related to time domain norms to the $\mathcal{H}_\infty$ frequency domain norm, and represents an extension of previous results derived for the $\mathcal{H}_\infty$ norm to a much wider class of input signals. Nevertheless, the main limitation of the work is that a certain technical condition on the input needs to be imposed that precludes the results applying to all inputs that are persistently exciting

of all orders. Work is in progress to use the techniques of this paper to remedy the deficiency.

References

[1] H. Akçay and P. P. Khargonekar, The least–squares algorithm, parametric identification, and bounded noise, *Automatica* **29** (1993) 1535–1540.

[2] H. Akçay and H. Hjalmarsson, The least-squares identification of FIR systems subject to worst-case noise, *Syst. Contr. Lett.* **23** (1994) 329–338.

[3] E. W. Bai and K. M. Nagpal, Least-squares type algorithms for identification in the presence of modelling uncertainty, *IEEE. Trans. Automat. Contr.* **40** (1995) 756–761.

[4] E. W. Bai, R. Tempo, and H. Cho, Membership set estimation: size, optimal input, complexity and relations with least-squares, *IEEE Trans. CAS: Part I* **4** (1995) 266–277.

[5] M. A. Dahleh, T. Theodosopoulos, and J. N. Tsitsiklis, The sample complexity of worst-case identification of F.I.R. systems, *Syst. Contr. Lett.* **2** (1993) 157–166.

[6] G. H. Hardy and J. E. Littlewood, Some new properties of Fourier constants *Math. Ann.* **97** (1927) 159–209.

[7] A. J. Helmicki, C. A. Jacobson, and C. N. Nett, Control oriented system identification: A worst-case/deterministic approach in $\mathcal{H}_\infty$, *IEEE Trans. Automat. Contr.* **36** (1991) 1163–1176.

[8] J. A. Jacobson, C. N. Nett, and J. R. Partington, Worst-case identification in ℓ_1: optimal algorithms and error bounds, *Syst. Contr. Lett.* **19** (1992) 419–424.

[9] L. Ljung, *System Identification, Theory for the User*, (Prentice–Hall, Englewood Cliffs, NJ, 1987).

[10] L. Ljung and Z. D. Yuan, Asymptotic properties of black-box identification of transfer functions, *IEEE Trans. Automat. Contr.* **30** (1985) 514–530.

[11] B. Kacewicz and M. Milanese, Optimality properties in finite-sample ℓ_1 identification with bounded noise, *Int. J. Adaptive Contr. Signal Processing* **9** (1995) 87–96.

[12] L.Ljung. Asymptotic variance expressions for identified black-box transfer function models. *IEEE Transactions on Automatic Control*, **30** (1985) 834–844,

[13] P. M. Mäkilä, Robust identification and Galois sequences, *Int. J. Contr.* **54** (1991) 1189–1200.

[14] P.M. Mäkilä, J.R. Partington, and T.K. Gustafsson. Worst-case control-relevant identification. *Automatica*, **31** (1995) 1799–1820.

[15] M. Milanese, Properties of least-squares estimates in set membership identification, *Automatica* **31** (1995) 327–332.

[16] B. Ninness and G.C. Goodwin. Estimation of model quality. *Automatica*, **31** (1995) 32–74.

[17] J. R. Partington, Worst-case identification in ℓ_2: linear and nonlinear algorithms, *Syst. Contr. Lett.* **22** (1994) 93–98.

[18] J. R. Partington, Worst-case analysis of the least-squares method and related identification methods, *Syst. Contr. Lett.* **24** (1995) 193–200.

[19] J. R. Partington and P. M. Mäkilä, Analysis of Linear Methods for Robust Identification in ℓ_1, *Automatica* **31** (1995) 755–758.

[20] K. Poolla and A. Tikku, On the time complexity of worst-case system identification, *IEEE Trans. Automat. Contr.*, **39** (1994) 944–950.

[21] R. S. Smith and M. Dahleh, *The Modeling of Uncertainty in Control Systems*, Lecture Notes in Control and Information Sciences, (Springer-Verlag, 1994).

[22] T. Söderström and P. Stoica, *System Identification*, (Prentice-Hall Int., Hemel Hempstead, Hertfordshire, 1989).

[23] Norbert Wiener. *The Fourier Integral and Certain of its Applications.* Cambridge University Press, 1933.