FACTA UNIVERSITATIS (NIŠ)
SER.: ELEC. ENERG. vol. 21, no. 1, April 2008, 1-21

# On Strong Consistency of a Class of Recursive Stochastic Newton-Raphson Type Algorithms with Application to Robust Linear Dynamic System Identification

**Ivana Kovačević, Branko Kovačevic, and Željko Đurović**

**Abstract:** The recursive stochastic algorithms for estimating the parameters of linear discrete-time dynamic systems in the presence of disturbance uncertainty has been considered in the paper. Problems related to the construction of min-max optimal recursive algorithms are demonstrated. In addition, the robustness of the proposed algorithms has been addressed. Since the min-max optimal solution cannot be achieved in practice, an approximate optimal solution based on a recursive stochastic Newton-Raphson type procedure is suggested. The convergence of the proposed practically applicable robustified recursive algorithm is established theoretically using the martingale theory. Both theoretical and experimental analysis related to the practical robustness of the proposed algorithm are also included.

**Keywords:** Recursive algorithms, convergence, robustness, parameter estimation, nonlinear filtering, nongaussian noise.

## 1 Introduction

The presence of large unmodelled errors may severely degrade the performance of optimal statistical estimation methods [1–4]. Many convincing examples can be found in areas such as flight control, electric power systems, telecommunications, industrial process control, econometrics, biomedical systems, etc. [5–9]. Estimation algorithms based on the Gaussian random error model have been found to be especially inefficient when the real error distribution belongs to the heavy tailed variety, giving rise to occasionally very large errors named outliers [1–4].

Manuscript received on August 15, 2007.

I. Kovačević is with High School Of Electrical Engineering, Vojvode Stepe 183, 11000 Belgrade (e-mail: ivanak@vets.edu.yu). B. Kovačević and Ž. Đurović are with Faculty of Electrical Engineering, University of Belgrade, Bul. kralja Aleksandra 73, 110000 Belgrade Serbia (e-mails: [kovacevic_b, zdjurovic]@etf.bg.ac.yu).

Therefore, considerable efforts have been oriented towards the robust estimation algorithms possessing a low sensitivity to error distribution changes, usually valid locally within a prespecified class. The fundamental contribution to the field of robust estimation has been given by Huber, who introduced the concept of min-max robust estimation [3]. Further developments of this idea have led to applications on different type of problems, including system identification, state estimation, signal processing and adaptive control [10–16]. However, the strong theoretical results in robust identification are restricted mostly to static models [17, 18]. In the case of dynamic systems such an analysis is difficult, owing to both the dynamic nature of system model and nonlinear form of a robust algorithm itself [18–20]. As a consequence, the convergence study of robust dynamic system identification schemes, is rather complicated. In general, there are at least two approaches for a such analysis. The first one is the ordinary differential equation (ODE) approach [19]. The second one uses the martingale convergence theorem, and represents an extremely powerful method that relies on relatively weak assumptions [19, 20].

The purpose of this article is to extend the concept of min-max optimal estimation to the problem of robust recursive identification of linear, dynamic, discrete-time single input single output systems in the presence of disturbance uncertainty. Problems related to the construction of min-max optimal recursive algorithms of stochastic gradient type are demonstrated. The link between the min-max optimal recursive parameter estimation and robust recursive parameter estimation has been also established. Since the min-max optimal solution cannot be achieved in practice, a simple procedure for constructing a robustified recursive stochastic Newton-Raphson type algorithm, based on a realizable nonlinear transformation of the prediction error together with a suitable generation of the weighting matrix, is suggested. The convergence of the proposed algorithm is established theoretically using the martingale theory. This results in a set of rather weak conditions under which the proposed algorithm should perform satisfactorily in practice. Experimental analysis, based on Monte Carlo simulations, illustrate the discussion and show efficiency of the derived robustified recursive algorithm in a non Gaussian and impulsive noise environment.

## 2 Problem Formulation

Let an abstract linear, dynamic, discrete-time, time-invariant system under consideration be modeled by a linear difference equation with fixed parameters

$$y(i) = -\sum_{k=1}^{n} a_k y(i-k) + \sum_{k=1}^{m} u(i-k) + \xi(i) \tag{1}$$

where $y(i) \in R^1$, $u(i) \in R^1$, $\xi(i) \in R^1$ are system output, measurable input, and stochastic input or noise, respectively, while the constants $a_i$, $i = 1,\ldots,n$ and $b_j$, $j = 1,\ldots,m$ and represent the system parameters. It will be assumed that the sequence $\{\xi(i)\}$ is a stochastic process, generating an increasing sequence of sub-sigma algebras $\{F_i\}$, and that constants $m$ and $n$ are a priory known $(0 < m \leq n)$. Furthermore, the probability density function (pdf) of $\xi(i)$ is not completely known, but some knowledge of this pdf is available, which can be represented as a certain class $P$ of zero-mean symmetric pdf's to which the real disturbance pdf belongs.

Introducing the backward shift operator $q^{-k}y(i) = y(i-k)$, the equation (1) can be written in the polynomial form

$$A(q^{-1})y(i) = B(q^{-1})u(i) + \xi(i) \tag{2}$$

where

$$A\left(q^{-1}\right) = 1 + \sum_{k=1}^{n} a_k q^{-k},\ B\left(q^{-1}\right) = \sum_{k=1}^{m} b_k q^{-k} \tag{3}$$

are the characteristic and control polynomial, respectively. One can also rewrite (1) as a linear regression equation

$$y(i) = Z^T(i)\Theta + \xi(i) \tag{4}$$

where $Z^T(i) = [-y(i-1),...,-y(i-n),u(i-1),...,u(i-m)]$ is the vector of input-output data, and $\Theta^T = [a_1 \cdots a_n b_1 \cdots b_m]$ is the constant parameter vector. The system representation in (2) is known as the autoregressive model with exogenous input (ARX), or the infinite impulse response filter (IIR). The problem of recursive identification of a system described by (4) can be considered as the task of estimation the unknown parameter vector $\Theta$ in real-time, on the basis of current input-output measurements. Formulation of the identification problem reduces to the choice of a forecasting or prediction model $\hat{y}(i) = \hat{y}(i/\theta)$ and the choice of an identification criterion or average loss [17–20]

$$J(\theta) = E\{F[\nu(i,\theta)]\}. \tag{5}$$

Here $\nu(i,\theta) = y(i) - \hat{y}(i/\theta)$ is the output prediction error, while $F(\cdot)$ is the loss function. The solution reduces to determining an identification algorithm

$$\hat{\theta}(i) = f\left(\hat{\theta}(i-1), y(i), u(i)\right) \tag{6}$$

which determines the estimates $\hat{\theta}(i)$ of the system parameter vector $\theta$ from the preceding estimated $\hat{\theta}(i-1)$ and current input-output measurements $\{y(i), u(i)\}$.

The goal is to develop a forecasting model $\hat{y}(\cdot)$, an optimal loss function $F(\cdot)$ in (5) and, finally, an algorithm (6) which should be optimal in a certain sense.

Starting from (1) or (4), the mean-square optimal forecasting model, minimizing the criterion $E\left\{v^2(i,\theta)\right\}$, is given by [18–20]

$$\hat{y}(i/\theta) = \left[1 - A\left(q^{-1}\right)\right] y(i) + B\left(q^{-1}\right) u(i) = Z^T(i)\,\theta \tag{7}$$

where $Z(i)$ and $\theta$ are given by (4). The concept of optimality is closely related to the level of available prior information on the system and unobserved disturbances, as it will be discussed in the next section.

## 3 Review of Min-Max Optimal Robust Parameter Estimation

With incomplete prior information on disturbances, one can construct min-max optimal, robust identification algorithms, minimizing the performance index for the least favorable pdf within a given class [1–4]. Namely, let $\tau$ be a class of parameter vector estimates and $V(T,p)$ the asymptotic estimation error covariance matrix of $T \in \tau$ when the pdf is $p \in P$ . Consider the game in which we choose $T \in \tau$, while nature chooses $p \in P$, and $V(T,p)$ is the payoff. This game has a saddle point pair $(T_0, p_0)$ if $T_{-0}$ and $p_0$ satisfy

$$\min_{T\in\tau}\max_{p\in P} V(T,p) = V(T_0,p_0) = \max_{p\in P}\min_{T\in\tau} V(T,p). \tag{8}$$

This $T_0$ is referred to as the min-max optimal robust estimate and $p_0$ is the least favorable pdf.

Particularly, if $\tau$ is the class of recursive stochastic gradient type estimators of the vector $\Theta$ in (4), represented by [18–20]

$$\hat{\Theta}(i) = \hat{\Theta}(i-1) + \Gamma(i) Z(i)\,\psi\left(v\left(i,\hat{\Theta}(i-1)\right)\right) \tag{9}$$

where $v\left(i,\hat{\Theta}\right) = y(i) - \hat{\Theta}^T Z(i)$ is the prediction error, or measurement residual, the min-max optimal robust estimator $T_0$ is defined by [1–4]

$$F(.) = F_0(.) = -\log p_0(.) \tag{10}$$

$$\psi(\cdot) = \psi_0(\cdot) = F_0'(.) \tag{11}$$

$$\Gamma(i) = \Gamma_0(i) = \left[iE\left\{\psi_0'(\cdot)\right\} E\left\{Z(i) Z^T(i)\right\}\right]^{-1}. \tag{12}$$

Here $p_0(\cdot)$ is the least favorable pdf minimizing the Cramer-Rao bound within a given class $P$ [17,18]. In this way, $\psi_0(\cdot)$ in (11) is the maximum likelihood (ML) type function corresponding to a unique least favorable pdf $p_0$ within a prespecified

class $P$, and (9) reduces to the ML type identification procedure [17–20]. However, the problem of finding the pdf $p_0(\cdot)$ requires the solution of a non-classical variational problem, which is tractable only by numerical methods [17,18]. It is solvable analytically only in the case of finite memory systems, i.e. $A(\cdot)=1$ in (2), when it reduces to the task of minimizing the Fisher information $I(p)=E\left\{(p'/p)^2\right\}$ within the class $P$ [17, 18]. Numerous examples of classes $P$, and the corresponding solution for the pdf $p_0(\cdot)$ minimizing the Fisher information $I(p)$ within the given class $P$, can be found in [18]. Moreover, the optimal weighing matrix $\Gamma_0(\cdot)$ in (12) cannot be constructed in practice, since it requires a priory knowledge of the real disturbance pdf. Therefore, in robust estimation, one looks for estimators that are quite efficient if the underlying disturbance distribution is normal but are also very efficient even though the underlying distribution has long tails, generating the extreme values of a measurement signal named outliers. This properties is the so-called efficiency robustness [1–4].

In general, the construction of a practically applicable recursive robust identification algorithm requires further approximations of the min-max optimal solution. A possible approach based on a weighted least squares method, reducing the effects of extreme disturbances or outliers, has been proposed in [21]. An alternative approach, based on a stochastic gradient type algorithms (9), combined with the recursive generation of the weighting matrix in (12) by step-by-step optimization of the additional criterion and convenient approximations, is presented in [22]. Another possibility for generating an approximate optimal solution of (12), based on a recursive stochastic Newton-Raphson type procedure, is presented in the next section. It should be noted that a recursive estimator of the type (9), not necessarily using the particular $\psi$ function of the ML type in (11), is called an approximate maximum likelihood recursive estimator, or recursive M-estimator [3, 17–19].

## 4 Robustified Recursive Stochastic Newton-Raphson Type Algorithms

In order to apply a Newton-Raphson type procedure, one has to approximate the average loss (5) with the empirical average loss [17–19]

$$J_i(\theta)=i^{-1}\sum_{k=1}^{i}F(\nu(k,\theta)). \tag{13}$$

Under certain conditions, with $i$ growing $J_i(.)$ in (13) converges to $J(.)$ in (5) [18]. Thus, one can resort to the approximate Newton-Raphson type method for solving a set of nonlinear equations resulting from the optimality condition of (13). This leads to an iterative algorithm of the form [17–19]

$$\hat{\theta}(i) = \hat{\theta}(i-1) - \left[i\nabla_{\theta}^{2} J_i\left(\hat{\theta}(i-1)\right)\right]^{-1}\left[i\nabla_{\theta} J_i\left(\hat{\theta}(i-1)\right)\right] \tag{14}$$

with $\nabla_{\theta}(\cdot)$ being the partial derivative operator

$$\nabla_{\theta}(.) = \left[\frac{\partial(\cdot)}{\partial a_1}, \cdots, \frac{\partial(\cdot)}{\partial a_n}, \frac{\partial(\cdot)}{\partial b_1}, \cdots, \frac{\partial(\cdot)}{\partial b_m}\right]^{T}.$$

Moreover, with large $i$ and by virtue of approximate truth of the optimality condition, yielding $\nabla_{\theta} J_{i-1}\left(\hat{\theta}\right) \approx 0$ and $v\left(i, \tilde{\theta}(i-1)\right) \approx \xi(i)$, one obtains from (5) and (13) the following approximate expressions

$$\begin{aligned} i\nabla_{\theta} J_i\left(\hat{\theta}(i-1)\right) &= -Z(i)\,\psi\left(v\left(i, \hat{\theta}(i-1)\right)\right) \\ i\nabla_{\theta}^{2} J_i\left(\hat{\theta}(i-1)\right) &= \alpha \sum_{k=1}^{i} Z(k) Z^{T}(k) \end{aligned} \tag{15}$$

where $\psi(\cdot) = F'(\cdot)$ and $\alpha = E\{\psi'(\xi(i))\}$. In deriving he second relation in (15) is used the fact that $\nabla_{\theta}^{2} J_i\left(\hat{\theta}\right) \approx \nabla_{\theta}^{2} J\left(\hat{\theta}\right) = \alpha E\left\{Z(i) Z^{T}(i)\right\}$, together with the approximation of the mathematical expectation with the arithmetic mean, i.e. $E\left\{Z(i) Z^{T}(i)\right\} \approx i^{-1} \sum_{k=1}^{i} Z(k) Z^{T}(k)$ . By substituting the first relation from (15) in (14), one obtains the parameter update equation (9). Furthermore, by introducing

$$\Gamma(i) = \left[i\nabla_{\theta}^{2} J_i\left(\hat{\theta}(i-1)\right)\right]^{-1} \tag{16}$$

the second relation in (15) reduces to

$$\Gamma^{-1}(i) = \Gamma^{-1}(i-1) + \alpha Z(i) Z^{T}(i). \tag{17}$$

Additionally, the matrix inversion lemma states that if the matrices A, B, C and D satisfy the equation [18–20]

$$A^{-1} = B^{-1} + C^{T} D^{-1} C$$

then

$$A = B - BC^{T}\left(CBC^{T} + D\right)^{-1} CB.$$

By choosing

$$A = \Gamma(i), B = \Gamma(i-1), C^{T} = Z(i), D^{-1} = \alpha,$$

one obtains from the matrix inversion lemma and the relation (17)

$$\Gamma(i) = \Gamma(i-1) - \frac{\Gamma(i-1) Z(i) Z^{T}(i) \Gamma(i-1)}{\alpha^{-1} + Z^{T}(i) \Gamma(i-1) Z(i)}. \tag{18}$$

The relations (9) and (18) define a robustified recursive stochastic Newton-Raphson type algorithm, with some initial guesses $\hat{\theta}(0) = 0, \Gamma(0) = cI$, where $c$ is some positive constant and $I$ is the identity matrix of corresponding order.

Moreover, in practice one has to adopt a class of realizable procedures (9), with $\psi(\cdot)$ being a suitable chosen nonlinearity, which has to cut off the outliers. As mentioned before, regarding the practical importance of achieving robustness with respect to ouliers contaminating the Gaussian disturbances, this function has to provide high efficiency at the nominal Gaussian model, as well as for a strategically chosen set of outlier models(efficiency robustness [1–4]). Additionally, it is desirable that this function be bounded and continuous [1–3]. Namely, boundedness insures that no single observation can have an arbitrarily large influence on estimates, while continuity insures that patchy outliers will not have a major effects. This requirement is known as resistant robustness [1–4]. Thus, $\psi(z)$ should look like $z$ for small values of the argument $z$, in order to preserve the regular observations generated by normal distribution, but it has to grow slower than linear with $|z|$, in order to suppress the influence of outliers. This corresponds, for example, to the choice of the saturation type nonlinearity named Huber influence function [3]

$$\psi(z) = \min\left(\frac{z}{\sigma^2}, \frac{k}{\sigma^2}\right)\operatorname{sgn}(z) \tag{19}$$

where $\sigma$ is the disturbance standard deviation, and the tuning constant $k$ has to be chosen so as to give the desired efficiency at the nominal Gaussian model [3]. However, the noise variance $\sigma^2$ is usually unknown and it must be estimated. Although ad hock, a popular robust estimate of $\sigma^2$ is the median of the absolute median deviations [1]

$$d(i) = \operatorname{median}\left\{\left|y(k) - \operatorname{median}\left\{y(k)\right\}\right|\right\}/0.6745\ , k = i - L + 1, \cdots, i \tag{20}$$

The divisor 0.6745 is used because then $d \approx \sigma$ if the sample size L is large enough and if the sample actually arises from a normal distribution. This particular scheme of selecting d at each step i suggests appropriate values of the tuning constant k in (19). Namely, since $d \approx \sigma$, the parameter k is usually taken to approximately be the value close to 1.5. Moreover, a common choice for the sliding window length is from the interval $L \in (5, 30)$ [13, 15].

Other $\psi(\cdot)$ functions that are commonly used in robust estimation can be found in [1–3].

**Remark 1** *If one chooses*

$$\begin{aligned} \psi(\cdot) &= \psi_0(\cdot) = -\left[\log\left(p_0(\cdot)\right)\right]'; \\ p_0 &= \arg\min_{p \in P} I(p) \end{aligned} \tag{21}$$

*this function does not provide the optimality in the min-max sense (8), but it minimizes the conditional covariance of the parameter estimate increment* $E\left\{\left[\hat{\Theta}(i)-\hat{\Theta}(i-1)\right]\left[\hat{\Theta}(i)-\hat{\Theta}(i-1)\right]^{T}\middle|F_{i-1}\right\}$ *for the least favorable pdf* $p_0$ *in (21) [9]. Namely, if the aim is to desensitize the algorithm with respect to outliers occurring at instant i, the choice* $\psi(\cdot)=\psi_0(\cdot)$ *provides to make the incremental covariance as small as possible, having in mind the supposed accuracy achieved in the preceding iteration.*

**Remark 2** *The idea of introducing* $\psi(\cdot)$ *in (21) can be justified also by analyzing one-step optimal estimates. Supposing that* $\Theta$ *in (4) is a random vector, one can show that this function corresponds to the saddle point of the conditional error covariance* $E\left\{\left[\hat{\Theta}(i)-\Theta\right]\left[\hat{\Theta}(i)-\Theta\right]^{T}\middle|F_{i-1}\right\}$ [9].

**Remark 3** *The choice of* $\psi(\cdot)$ *in (19) corresponds to the* $\psi_0(\cdot)$ *function in (21) when P is the* $\varepsilon-$ *contaminated family, defined by* [1–4]

$$P=P_{\varepsilon}=\left\{p\middle|p=(1-\varepsilon)N\left(0,\sigma^{2}\right)+\varepsilon h\right\},\ \varepsilon\in[0,1) \tag{22}$$

*with* $h(\cdot)$ *being zero-mean symmetric pdf and* $N\left(0,\sigma^{2}\right)$ *is the zero-mean normal pdf with the variance* $\sigma^{2}$ *[3]. The least favorable pdf* $p_0$ *in (21) within the class (22) is normal with exponential tails, yielding* $I^{-1}(p_0)=2(1-\varepsilon)erf(k)\sigma^{2}$*, where erf is the error function* [3].

The weighting matrix $\Gamma(i)$ in (18) depends on $\alpha$ and, as a consequence, the rate of estimates convergence also depends on it. Moreover, the factor $\alpha$ allows to make practically very important connections between the nonlinear transformation $\psi(\cdot)$ and the weighting matrix sequence.Unfortunately, it is very difficult to express these dependences explicitly. However, it can be shown that appropriate choice of $\alpha$ results in some intuitively appealing robust identification procedures, derived in the literature within different contexts and in different ways. Namely, since the real disturbance pdf is not known, a convenient possibility is to adopt $\alpha=E_{p_0}\{\psi_0'(\cdot)\}=I(p_0)$, with $E_{p_0}\{\cdot\}$ being the expectation with respect to the least favorable pdf $p_0$ in (21). The resulting algorithm is formally similar to the robustified Kalman filter or least squares method [13, 15]. Furthermore, when $\alpha(i)=1$ the recursion (18) reduces to the Riccati equation corresponding to the recursive least-squares method [19]. This algorithm differs from the least-squares method only by the insertion of nonlinear transformation $\psi_0(\cdot)$. Finally, if one approximates the mathematical expectation $\alpha=E_{p_0}\{\psi_0'(\cdot)\}$ by a single realization $\psi_0'(\cdot)$, one obtains an algorithm with changing factor $\alpha=\alpha(i)=\psi_0'\left(\nu\left(i,\hat{\theta}(i-1)\right)\right)$ in each step. An algorithm of the same form, starting from off-line estimates of constant parameters in static plants, has been derived in [17, 18].

The proposed algorithm has been derived on the basis of approximations and somewhat heuristic reasoning. However, all the available practically applicable recursive robust estimators are obtained as a result of approximations and assumptions, requiring further practical and/or theoretical verifications. We shall give the figure of merit of the proposed algorithm on the basis of convergence analysis using martingale theory, combined with experimental analysis based on Monte Carlo simulations.

## 5 Convergence Analysis

The basic convergence result is the lemma of Neveu [19, 20]. To be precise, let $(\Omega, F, P)$ be a probability space and $F_1 \subset F_2 \subset \cdots$ a sequence of sub- $\sigma$ -algebras of $F$, and *x(t)* is a sequence of real random variables adapted to $F$. Then $\{x(t), F_t\}$ is a martingale provided that $E\{|x(t)|\} < \infty$ almost surely (a.s.), i.e. with probability 1 (w.p.1), and $E\{x(t)/F_{t-1}\} = x(t-1)$ w.p.1. Alternatively, if $E\{x(t)/F_{t-1}\} \leq x(t-1)$ w.p.1 we say that $\{x(t), F_t\}$ is a supermartingale. Then, the following lemma can be proven [19, 20].

**Lemma 1** *Let $\{z_n\}$ be a sequence of nonnegative random variables and $\{F_n\}$ a sequence of increasing adapted sigma algebras, i.e. $z_n \in F_n$. Suppose*

$$E\{z_n/F_{n-1}\} \leq z_{n-1} + \alpha_n$$

*and $\sum_{n=1}^{\infty} \alpha_n < \infty$ w.p.1. Then $\{z_n\}$ converges w.p.1 to a finite nonnegative random variable $z^*$ as $n \to \infty$, i.e. $\lim_{n\to\infty} z_n = z^*$ w.p. 1.*

This result is restated in a number of forms that suit better in specific theoretical analysis. A unified treatment of a number of almost sure convergence theorems, based on the facts that the processes involved possess a common almost supermartingale properties has been proposed by Robbins and Siegmund [23]. This result is stated below.

**Theorem 1** *For each n let $z_n$, $\beta_n$, $\xi_n$ and $\zeta_n$ be non-negative $F_n$ -measurable random variables such that*

$$E\{z_{n+1}|F_n\} \leq z_n(1+\beta_n) + \xi_n - \zeta_n$$

*Then $\lim_{n\to\infty} z_n$ exists and is finite, i.e. $\lim_{n\to\infty} z_n = z^*$ w.p. 1, and $\sum_{n=1}^{\infty} \zeta_n < \infty$ w.p.1, on $\{\sum_{n=1}^{\infty} \beta_n < \infty\ ,\ \sum_{n=1}^{\infty} \xi_n < \infty\}$.*

The results of the Theorem 1 and/or Lemma 1 can be used to prove the convergence of the proposed robustified recursive stochastic gradient type algorithm (9), (18). The results are summarized in the theorem stated below.

**Theorem 2** *Consider the model (2), (3) and the algorithm (9), (18) subject to the conditions:*

- *A1 All zeros of the polynomial $A\left(q^{-1}\right)$ are inside the unit circle, and the sequence $\{u(i)\}$ is bounded.*
- *A2 $\{\xi(i)\}$ is a sequence of independent and identically distributed (i.i.d.) random variables, such that the probability distribution function $P(\cdot)$ is symmetric, and $E\{\xi(i)/F_{i-1}\}=0$, $E\left\{\xi^2(i)|F_{i-1}\right\}=\sigma^2<\infty$, while $v(i)-\xi(i)\in F_{i-1}$, with $F_{i-1}$ being the $\sigma$ -algebra generated by $\xi(0),\cdots,\xi(i-1),Z(0),\cdots,Z(i-1)$.*
- *A3 The function $\psi(\cdot)$ is odd and continuous almost everywhere.*
- *A4 The function $\psi(\cdot)$ grows slower than linear, i.e. $|\psi(z)|\leq k_1'\left(1+k_2'|z|\right)$; $k_1'\in(0,\infty)$, $k_2'\in[0,\infty)$.*
- *A5 The coefficient $\alpha$ in (18) is positive and bounded, i.e. $\alpha\in(0,k'')$, $k''<\infty$.*
- *A6 If $\phi_1(z)=E\{\psi(-\xi(i)+z)|F_{i-1}\}$ then $z\phi_1(z)\geq\frac{1}{2}\alpha z^2$ for each $z\neq 0$ and $\alpha$ given by (18).*
- *A7 The observation vector grows as $||Z(i)||^2\leq M\log^d r(i)$, $M>0,\delta>0$, where $r(i)=Tr\left\{\Gamma^{-1}(i)\right\}$ with $Tr\{.\}$ being the trace of a matrix and $||.||$ represents the Euclidean norm.*
- *A8 The persistent excitation conditions*

$$\lim_{i\to\infty}\lambda_{\min}\left\{\Gamma^{-1}(i)\right\}=\infty \ w.p.1$$

$$\lim_{i\to\infty}\frac{\log^k r(i)}{\lambda_{\min}\left\{\Gamma^{-1}(i)\right\}}=0 \ w.p.1$$

*for some $k>1+\delta$, with $\delta$ being given by A7, where $\lambda_{\min}\{.\}$ denotes the minimal eigenvalue of a matrix.*

*Then $\hat{\Theta}(i)$ converges to the true parameter value $\Theta$ with probability one (w.p.1), i.e. $P\{\lim_{i\to\infty}\hat{\Theta}(i)=\Theta\}=1$.*

The proof of the theorem is given in the appendix. A similar result is derived in the literature, but it relies on a rather strong assumption of bounded condition number of the inverse of algorithm weighting matrix [22]. As a consequence, a practical application of a such algorithm requires a condition number monitoring scheme. Moreover, a numerical difficulty may also arise when the condition number gets too large. This problem is overcome by introducing the assumption A8.

The assumptions A1 and A2 are commonly used in the convergence study of recursive stochastic gradient type identification algorithms, based on martingale

theory [19,20]. The assumption A1 means that system under consideration in (1) or (4) is bounded input-bounded output (BIBO) stable, while A2 denotes that additive measurements noise in (2) is a zero-mean white sequence. It should be noted that the assumption A1 is not formally needed to prove the convergence Theorem 2 (see, appendix), but it is introduced since the BIBO stability is one of the most desired properties of a system. Thus, the convergence result of Theorem 2 is in charge for both stable and unstable systems. On the other hand, the convergence result exposed in [22] can be applied only for a stable system.

The assumptions A3 and A4 define the class of nonlinear functions which provide for the consistent parameter estimates. Many $\psi(\cdot)$ functions that are commonly used in robust estimation, such as Huber's, Hampel's, Tukey's, and Andrew's nonlinearity satisfy the above assumptions [1–4]. Although the assumption A4 means that $\psi(\cdot)$ function may be unbounded, all the mentioned $\psi(\cdot)$ functions are bounded and continuous from practical reasons related to the resistant robustness. Moreover, the noise variance $\sigma^2$ in A2 has not to be finite provided $\psi(\cdot)$ function is bounded. Finally, it is hoped, and some numerical simulations seem to substantiate this hope, that the robust estimators approach their asymptotic behavior provided $\psi(\cdot)$ function is bounded [9–17].

The assumption A6 is a new one. A condition for A6 to be satisfied is that

$$\phi_1(a) = \int_0^\infty [\psi(u+a) - \psi(u-a)]\, dP(u)$$

is monotonically nondecreasing. Bearing in mind A2 and A3, this will be fulfilled if both the functions $\psi(\cdot)$ and $P(\cdot)$ have a common raising point, i.e. $\psi(z+\varepsilon) > \psi(z-\varepsilon)$; $P(z+\varepsilon) > P(z-\varepsilon)$ for some $z$ and every $\varepsilon > 0$, yielding $a\phi_1(a) > 0$ for $a \neq 0$ and $\phi_1(0) = 0$. Thus, the assumption A6 is fulfilled if $\psi(\cdot)$ function is odd, continuous almost everywhere, monotone increasing and piecewise continuously differentiable. As mentioned before, a desirable practical property is that $\psi(\cdot)$ function is also bounded ( $k_2 = 0$ in A4).

Particularly, if $P$ is the class of pdf's with bounded variance, i.e.

$$P = \left\{ p \left| \int_{-\infty}^{\infty} z^2 p(z)\, dz \leq \sigma^2 \right. \right\}$$

the least favourable pdf in (21) is zero-mean Gaussian with the variance $\sigma^2$, yielding $\psi_0(z) = z/\sigma^2$ [17, 18]. Thus, the resulting algorithm is the recursive least squares method [19,20]. Moreover, $\alpha = I(p_0) = \sigma^{-2}$ and $\phi_1(z) = z/\sigma^2$, so that the condition A6 is fulfilled. On the other hand, for the $\varepsilon$ -contaminated class of pdf's (22) the optimal nonlinearity $\psi_0(\cdot)$ in (21) is defined by (19), yielding $\phi_1(z) = \beta z$ where $\beta = \sigma^{-2} \int_{-k}^{k} p(u)\, du$ [17, 18]. Therefore, the hypothesis A6 is also satisfied,

and the resulting algorithm is a robust version of the conventional linear recursive least squares method [19, 20].

The assumption A7 determines the rate of the observation vector growth, but it is not restrictive since $M$ is an arbitrary positive constant.

Furthermore, it is fairly obvious that some condition on the input sequence must be introduced in order to secure a reasonable identification results. Clearly, an input that is identically zero will not be able to yield full information about the system input-output properties. Required input should excite all natural modes of the system. Such an input is called persistently exciting, and A8 represents one of the weakest versions of the persistent excitation assumption [19, 20].

**Remark 4** *The results of Theorem 2 are also valid if the noise sequence $\xi(i)$ in (1) is no more a zero mean white, but represents a colored or correlated zero mean sequence generated by a moving average process $C\left(q^{-1}\right)e(i)=\xi(i)$ with e(i) being a zero mean white sequence, while $C\left(q^{-1}\right)=1+\sum_{i=1}^{l}c_i q^{-i}$ represents a polynomial whose roots lies inside the unit circle. The system representation (2) with a such noise model is known as autoregressive moving average model with exogenous input (ARMAX). Then, all assumptions of the Theorem 2 remains the same with the exceptions that A6 changes to $z\left[\Phi_1\left(z/C\left(q^{-1}\right)\right)-1/2\alpha z\right]>0$ for every $z\neq 0$. Moreover, the forecasting model (7) becomes [19, 20]*

$$\hat{y}(i/\theta)=\frac{B\left(q^{-1}\right)}{C\left(q^{-1}\right)}u(i)+\left[1-\frac{A\left(q^{-1}\right)}{C\left(q^{-1}\right)}\right]y(i)=Z^T(i/\theta)\,\theta \tag{23}$$

*where*

$$\begin{aligned}\theta^T=&\{a_1,\cdots,a_n,b_1,\cdots,b_m,c_1,\cdots,c_l\}\\ Z^T(i/\theta)=&\{-y(i-1),\cdots,-y(i-n),u(i-1),\\ &\cdots,u(i-m),\nu(i-1,\theta),\nu(i-l,\theta)\}\end{aligned}$$

*with the output prediction error $\nu(i,\theta)=y(i)-\hat{y}(i/\theta)$. In the implementation of the algorithm, one has to replace the unknown $\theta$ with the corresponding most recent estimate, i.e. to replace $\nu(j,\theta)$ with $\nu\left(j,\hat{\theta}(j-1)\right)$, $j=1,\cdots,i$.*

## 6 Numerical Examples

In order to investigate more precisely the practical robustness of the proposed algorithm, Monte Carlo simulations have been undertaken. The results presented are related to the fourth-order model, given in the form (4): $Z^T(i)=\ [-y(i-1)\ ...\ -y(i-4)\ u(i-1)]$, and $\Theta^T=[1\ -0.18\ 0.78\ -0.656\ 2]$.

The sequence $\{u(i)\}$ is adopted to be a white noise, with $u(i)$ being the standard normal random variable $N(0,1)$, while the disturbance $\xi(i)$ is confined to the class of $\varepsilon$ -contaminated pdf's (22) with $\varepsilon = 0.1$ and $\sigma^2 = 1$. The following algorithms have been tested: 1) recursive least-squares algorithm, denoted as RLS; 2) recursive robust algorithm (9), (18) with the nonlinearity (19) with *k=1.5*, and the scale factor (20), calculated on the sliding frame length of $L$=10 samples, denoted as RRA.

The effect of desensitizing the estimates to the influence of outliers is illustrated in Tab. 1, depicting the average square error norm for different outlier characteristics, calculated on the basis of 100 Monte Carlo trials and 500 iterations. Obviously, the least-squares algorithm is slightly superior to robust algorithm in the case of Gaussian disturbances (Tab. 1; $h(.) = N(0,1)$ in (22)). The robust algorithm is superior than the least-squares method even for smaller outlier variances (Tab. 1, $h(.) = N(0,10)$ *or* $h(.) = L(0,1)$. It should be noted that the total noise variance is equal to $0.9 + 0.1 * var\{h(.)\}$, where $var\{L(0,1)\} = 2$, $var\{C(0,1)\} = \infty$, $var\{N(0,10)\} = 10$ ), and gives significantly better performances for larger outlier variances (Tab. 1, $h(.) = C(0,1)$ ). Moreover, the robust method performs quite well uniformly for different outlier statistics $h(\cdot)$, under the circumstances characterized by a small or moderate value of the contamination degree $\varepsilon \leq 0.1$. This is encouraging from the point of view of its application, since the real outlier pdf $h(\cdot)$, as well as the contamination degree $\varepsilon$, are not known in practice.

Table 1. Average mean-square error norm for ARX model (2) and different outlier statistics $h$ ( $L$ - Laplace, $C$- Cauchy, $N$ - normal pdf; $\sigma^2 = 1$, $\varepsilon = 0.1$, $k = 1.5$, $\hat{\Theta}(0) = 0$, $\Gamma(0) = 0.1I$ )

| Algor. | $L(0,1)$ | $C(0,1)$ | $N(0,1)$ | $N(0,10)$ |
|---|---|---|---|---|
| **RLS** | 0.1249 | 0.9721 | 0.0122 | 0.1235 |
| **RRA** | 0.0167 | 0.0194 | 0.0139 | 0.0160 |

Moreover, the proposed robust procedure is nonlinear and, consequently, the estimates may be highly influenced by initial conditions $\hat{\Theta}(0)$ and $\Gamma(0)$. However, a low sensitivity to initial conditions is important for achieving practical robustness. The problem of initial conditions can be circumvented by using a good starting value as a result of off-line robust M-estimation [3, 9, 13–15].

Additionally, the application of algorithm requires the exact knowledge of $\alpha = I(p_0)$, which depends on the contamination degree $\varepsilon$ (see, remark 3). However, $\varepsilon$ is not known in practice, and one has to adopt it a priory. The usual values of $\varepsilon$ are from the range $(0.02, 0.1)$ [1–4]. The experiments have shown that the algorithm is rather insensitive to the choice of $\varepsilon$ belonging to the above region.. As mentioned before, another possibility is to approximate the expectation by a single realization,

i.e. $\alpha = \alpha(i) \approx \psi'\left(v\left(i, \hat{\Theta}(i-1)\right)\right)$, resulting in an algorithm with adaptive factor $\alpha$ in each step. The results obtained are similar to those presented above, but this algorithm is more sensitive to the initial conditions.

To illustrate the characteristic of the proposed algorithm in a case of correlated measurement noise, we have applied the RLS and RRA algorithms to the ARMAX model (2) (see, remark 4)

$$A\left(q^{-1}\right) = 1 + 0.83q^{-1} - 0.167q^{-2}, B\left(q^{-1}\right) = 0.167q^{-1}, C\left(q^{-1}\right) = 1 + 0.2q^{-1}$$

As before, $\{u(i)\}$ is taken to be white normal sequence *N(0,1)*, while $\{x(i)\}$ is contaminated normal, distributed as (22) with $\varepsilon = 0.1$ and $\sigma^2 = 1$. Table 2 depicts the average mean-square error norm, obtained on the basis of 100 Monte Carlo trials and 500 iterations, for different outlier statistics $h(\cdot)$.

Table 2. Average mean-square error norm for ARMAX model and different outlier statistics $h$ ( $L$ - Laplace, $C$- Cauchy, $N$ - normal pdf; $\sigma^2 = 1$, $\varepsilon = 0.1$, $k = 1.5$, $\hat{\Theta}(0) = 0$, $\Gamma(0) = 0.1I$)

| Algor. | $L(0,1)$ | $C(0,1)$ | $N(0,1)$ | $N(0,10)$ |
|---|---|---|---|---|
| **RLS** | 0.1544 | 1.9151 | 0.0220 | 0.1635 |
| **RRA** | 0.0360 | 0.0484 | 0.0309 | 0.0278 |

Similarly as in the previous example, linear RLS is slightly superior than the nonlinear robust RRA for the Gaussian noise (Tab. 2; $h(.) = N(0,1)$ in (22)). However, in the presence of outliers (Tab. 2, $h(.) = L(0,1)$ *or* $C(0,1)$ *or* $N(0,10)$ ) it leads to the biased RLS estimates, while RRA performs quite well in all situations. As before, the reason lies not only in the nonlinear transformation of the prediction errors, but also in an adequate way of generating the weighting matrix $\Gamma(i)$, where the factor $\alpha$ keeps the eigenvalues of $\Gamma(i)$ at values high enough to provide for noise immunity.

## 7 Conclusion

The problem of recursive robust identification of linear dynamic discrete-time single input-single output systems has been considered in the paper. A theoretical analysis has shown that the asymptotically min-max optimal, robust algorithms cannot be constructed in practice. Arguments are given indicating possibilities of applying realizable, but nonoptimal, nonlinear transformation of the prediction error. As a result, a general form of robustified recursive stochastic Newton-Raphson type identification schemes is adopted. The convergence of the proposed recursive algorithm is established theoretically using the martingale theory. In order to investigate practical robustness of the algorithm, Monte-Carlo simulations have

been undertaken. The results obtained have shown that the efficiency of the robust algorithm is generally better than for the conventional recursive least-squares method. Moreover, the implementation of the robust recursive algorithm is inexpensive, since it requires effectively no additional cost in either computer time or program complexity.

## Appendix

## A Proof of the Convergence Theorem 2

If we denote $\tilde{\Theta}(i) = \hat{\Theta}(i) - \Theta$ and introduce the Lyapunov's stochastic function $V(i) = \tilde{\Theta}^T(i)\Gamma^{-1}(i)\tilde{\Theta}(i)$, we obtain from (9)

$$\begin{aligned} V(i) =& \tilde{\Theta}^T(i-1)\Gamma^{-1}(i)\tilde{\Theta}(i-1) \\ &+ 2\tilde{\Theta}^T(i-1)Z(i)\psi\left(\nu\left(i,\hat{\Theta}(i-1)\right)\right) \\ &+ Z^T(i)\Gamma(i)Z(i)\psi^2\left(\nu\left(i,\hat{\Theta}(i-1)\right)\right). \end{aligned} \tag{24}$$

By adding and subtracting $V(i-1)$ to the right hand side of (24) and taking into account (17), one obtains

$$\begin{aligned} V(i) =& V(i-1) + 2\tilde{\Theta}^T(i-1)Z(i)\psi\left(\nu\left(i,\hat{\Theta}(i-1)\right)\right) \\ &+ \alpha\left(\tilde{\Theta}^T(i-1)Z(i)\right)^2 \\ &+ Z^T(i)\Gamma(i)Z(i)\psi^2\left(\nu\left(i,\hat{\Theta}(i-1)\right)\right). \end{aligned} \tag{25}$$

Let us define the function

$$\phi_2\left(\tilde{\Theta}^T(i-1)Z(i)\right) = E\left\{\psi^2\left(-\nu\left(i,\hat{\Theta}(i-1)\right)\right)\middle| F_{i-1}\right\} \tag{26}$$

where

$$\nu\left(i,\hat{\Theta}(i-1)\right) = -\tilde{\Theta}^T(i-1)Z(i) + \xi(i) \tag{27}$$

is the prediction error, or residual. The function $\Phi_2(.)$ exists under the same conditions as the function $\Phi_1(.)$. Taking into account the hypothesis A4, one concludes

$$\phi_2\left(\tilde{\Theta}^T(i-1)Z(i)\right) \leq k_1\left[1 + k_2\left(\tilde{\Theta}^T(i-1)Z(i)\right)^2\right] \tag{28}$$

where $k_1$ and $k_2$ are a finite positive and nonnegative constants, respectively. Since $Tr\left\{bb^T\right\} = b^T b$ for a column vector $b$, by applying the trace operation on (17) and choosing $b = Z(i)$, one obtains

$$r(i) = r(i-1) + \alpha Z^T(i)Z(i). \tag{29}$$

By dividing (25) with $\log^k r(i)$ and taking the expectation, as well as by applying A2, A3 and (26)-(28), one obtains further

$$\begin{aligned} E\left\{\frac{V(i)}{\log^k r(i)}|F_{i-1}\right\} \leq \frac{V(i-1)}{\log^k r(i)} - \frac{2\tilde{\Theta}^T(i-1)Z(i)}{\log^k r(i)} \\ \times\left[\phi_1\left(\tilde{\Theta}^T(i-1)Z(i)\right) - \frac{\alpha}{2}\tilde{\Theta}^T(i-1)Z(i)\right] \\ +k_1k_2\left(\tilde{\Theta}^T(i-1)Z(i)\right)^2\frac{Z^T(i)\Gamma(i)Z(i)}{\log^k r(i)} \\ +k_1\frac{Z^T(i)\Gamma(i)Z(i)}{\log^k r(i)}. \end{aligned} \tag{30}$$

By analyzing the third term in (30), we have

$$\begin{aligned} k_1k_2\left(\tilde{\Theta}(i-1)Z(i)\right)^2\frac{Z^T(i)\Gamma(i)Z(i)}{\log^k r(i)} = k_1k_2\tilde{\Theta}^T(i-1)\Gamma^{-1/2}(i-1) \\ \times\Gamma^{1/2}(i-1)Z(i)Z^T(i)\Gamma^{1/2}(i-1)\Gamma^{-1/2}(i-1) \\ \times\tilde{\Theta}(i-1)\frac{Z^T(i)\Gamma(i)Z(i)}{\log^k r(i)}. \end{aligned} \tag{31}$$

Moreover, the well known result from matrix analysis stated for a square matrix $A$ and vector $b$ [18–20]

$$\begin{aligned} |\lambda_{\min}|\,\|b\| \leq \|Ab\| \leq |\lambda_{\max}|\,\|b\|, \\ |\lambda_{\min}|\,\|b\|^2 \leq |b^TAb| \leq |\lambda_{\max}|\,\|b\|^2 \end{aligned} \tag{32}$$

with $\lambda_{\min}$ and $\lambda_{\max}$ being the eigenvalues of $A$ with smallest and largest absolute values, and $\|b\|^2 = b^Tb$. Thus, by choosing in (32)

$$b = \Gamma^{-1/2}(i-1)\tilde{\Theta}(i-1)\ \ , A = \Gamma^{1/2}(i-1)Z(i)Z^T(i)\Gamma^{1/2}(i-1)$$

further follows

$$\begin{aligned} &\tilde{\Theta}^T(i-1)\Gamma^{-1/2}(i-1)\Gamma^{1/2}(i-1)Z(i)Z^T(i)\Gamma^{1/2}(i-1)\Gamma^{-1/2}(i-1)\tilde{\Theta}(i-1) \\ &\leq \lambda_{\max}\left\{\Gamma^{1/2}(i-1)Z(i)Z^T(i)\Gamma^{1/2}(i-1)\right\}\left\|\Gamma^{-1/2}(i-1)\tilde{\Theta}(i-1)\right\|^2 \\ &= \lambda_{\max}\left\{\Gamma^{1/2}(i-1)Z(i)Z^T(i)\Gamma^{1/2}(i-1)\right\}\tilde{\Theta}^T(i-1)\Gamma^{-1/2}(i-1) \\ &\quad\times\Gamma^{-1/2}(i-1)\tilde{\Theta}(i-1) \\ &\leq Tr\left\{\Gamma^{1/2}(i-1)Z(i)Z^T(i)\Gamma^{1/2}(i-1)\right\}\tilde{\Theta}^T(i-1)\Gamma^{-1}(i-1)\tilde{\Theta}(i-1) \\ &= Tr\left\{\Gamma^{1/2}(i-1)Z(i)Z^T(i)\Gamma^{1/2}(i-1)\right\}V(i-1). \end{aligned} \tag{33}$$

Bearing in mind that $Tr\{AB\} = Tr\{BA\}$ and adopting

$$B = \Gamma^{1/2}(i-1) \ , \ A = \Gamma^{1/2}(i-1) Z(i) Z^T(i)$$

one obtains

$$Tr\left\{\Gamma^{1/2}(i-1) Z(i) Z^T(i) \Gamma^{1/2}(i-1)\right\} = Tr\left\{\Gamma(i-1) Z(i) Z^T(i)\right\}. \tag{34}$$

Furthermore, since for a square matrix $A$ and some column vector $b$ is fulfilled $Tr\left\{Abb^T\right\} = b^T Ab$, one concludes from (32) and (34), after choosing $A = \Gamma(i-1)$ and $b = Z(i)$

$$Tr\left\{\Gamma(i-1) Z(i) Z^T(i)\right\} \leq \lambda_{\max}\left\{\Gamma(i-1)\right\} \|Z(i)\|^2. \tag{35}$$

By subtracting (35) into (33), one obtains

$$\begin{aligned} &\tilde{\Theta}^T(i-1)\Gamma^{-1/2}(i-1)\Gamma^{1/2}(i-1) Z(i) Z^T(i) \Gamma^{1/2}(i-1)\Gamma^{-1/2}(i-1)\tilde{\Theta}(i-1) \\ &\leq \lambda_{\max}\left\{\Gamma(i-1)\right\} \|Z(i)\|^2 V(i-1). \end{aligned} \tag{36}$$

Moreover, by subtracting (36) into (31), further follows from A7

$$\begin{aligned} &k_1 k_2 \frac{\left(\tilde{\theta}^T(i-1) Z(i)\right)^2}{\log^k r(i)} Z^T(i)\Gamma(i) Z(i) \\ &\leq k_1 k_2 \frac{V(i-1)}{\log^k r(i)} \lambda_{\max}\left\{\Gamma(i-1)\right\} \| Z(i) \|^2 Z^T(i)\Gamma(i) Z(i) \\ &\leq k_1 k_2 \frac{M \log^\delta r(i)}{\lambda_{\min}\left\{\Gamma^{-1}(i-1)\right\}} \frac{V(i-1)}{\log^k r(i)} Z^T(i)\Gamma(i) Z(i) \\ &= k_3 \frac{V(i-1)}{\log^k r(i-1)} \frac{\log^k r(i-1)}{\lambda_{\min}\left\{\Gamma^{-1}(i-1)\right\}} \frac{Z^T(i)\Gamma(i) Z(i)}{\log^c r(i)} \end{aligned} \tag{37}$$

where $k_3 = k_1 k_2 M$ and $c = k - \delta > 1$. Bearing in mind (37), the relation (30) ca be rewritten as

$$\begin{aligned} E\left\{\frac{V(i)}{\log^k r(i)} | F_{i-1}\right\} &\leq \frac{V(i-1)}{\log^k r(i-1)} \left[1 + k_3 \frac{\log^k r(i-1)}{\lambda_{\min}\left\{\Gamma^{-1}(i-1)\right\}} \frac{Z^T(i)\Gamma(i) Z(i)}{\log^c r(i)}\right] \\ &- 2\frac{\tilde{\theta}^T(i-1) Z(i)}{\log^k r(i)} \left[\phi_1\left(\tilde{\theta}^T(i-1) Z(i)\right) - \frac{\alpha}{2}\tilde{\theta}(i-1) Z(i)\right] \\ &+ k_1 \frac{Z^T(i)\Gamma(i) Z(i)}{\log^k r(i)}. \end{aligned} \tag{38}$$

Let us introduce the notations

$$\begin{aligned}
z_{i-1} &= V(i-1)/\log^k r(i-1);\\
\beta_{i-1} &= k_3 \frac{\log^k r(i-1)}{\lambda_{\min}\{\Gamma^{-1}(i-1)\}} \frac{Z^T(i)\Gamma(i)Z(i)}{\log^c r(i)}\\
\xi_{i-1} &= k_1 \frac{Z^T(i)\Gamma(i)Z(i)}{\log^k r(i)};\\
\zeta_{i-1} &= 2\frac{\tilde{\Theta}^T(i-1)Z(i)}{\log^k r(i)}\left[\phi_1\left(\tilde{\Theta}^T(i-1)Z(i)\right) - \frac{\alpha}{2}\tilde{\Theta}^T(i-1)Z(i)\right].
\end{aligned} \tag{39}$$

In order to apply the convergence Theorem 2, one has to prove that $z_{i-1}$, $\beta_{i-1}$, $\xi_{i-1}$ and $\zeta_{i-1}$ are nonnegative $F_{i-1}$ measurable random variables, satisfying $\sum_{i=1}^{\infty}\beta_i < \infty$, $\sum_{i=1}^{\infty}\xi_i < \infty$. Starting from (29), one concludes that $r(i)$ is a sequence of non decreasing values, i.e. $r(i) \geq r(i-1)$. Furthermore, starting from A8 and since

$$r(i) = Tr\left\{\Gamma^{-1}(i)\right\} = \sum_{i=1}^{n+m}\lambda_i\left\{\Gamma^{-1}(i)\right\} \geq \lambda_{\min}\left\{\Gamma^{-1}(i)\right\} \tag{40}$$

where $\lambda_i$ are the nonnegative eigenvalues of the positive-semidefinite matrix $\Gamma^{-1}(.)$, with $\lambda_{\min}$ being the minimal eigenvalue, one concludes that $\lim_{i\to\infty} r(i) = \infty$, or equivalently for $i$ large enough $r(i) > 1$, from which it follows $\log r(i) > 0$. Furthermore, due to A6 and since the quadratic forms $V(i) = \tilde{\theta}^T(i)\Gamma^{-1}(i)\tilde{\theta}(i)$ and $Z^T(i)\Gamma(i)Z(i)$ are nonnegative, one also concludes that $z_{i-1}$, $\beta_{i-1}$, $\xi_{i-1}$ and $\zeta_{i-1}$ are nonnegative $F_{i-1}$ measurable random variables. Thus, it still remains to show that $\sum_{i=1}^{\infty}\beta_i < \infty$, $\sum_{i=1}^{\infty}\xi_i < \infty$, and this is equivalent to the condition

$$\sum_{i=1}^{\infty}\frac{Z^T(i)\Gamma(i)Z(i)}{\log^c r(i)} < \infty \text{ w.p.1} \tag{41}$$

for some $c > 1$. To prove (41), let us define the following matrix in the portioned form

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} \alpha^{-1} & Z^T(i) \\ Z(i) & \Gamma^{-1}(i) \end{bmatrix}. \tag{42}$$

Then, Schurs formula allows the determinant of a portioned matrix to be written as a product of component determinants [24, 25], i.e.

$$\det A = \det A_{11}\det\left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right) = \det A_{22}\det\left(A_{11} - A_{12}A_{11}^{-1}A_{21}\right). \tag{43}$$

By substituting (42) into (43), further follows

$$\alpha^{-1}\det\left(\Gamma^{-1}(i) - \alpha Z(i)Z^T(i)\right) = \det\Gamma^{-1}(i)\left(\alpha^{-1} - Z^T(i)\Gamma(i)Z(i)\right). \tag{44}$$

Starting from (17), one obtains

$$\Gamma^{-1}(i-1) = \Gamma^{-1}(i) - \alpha Z(i) Z^T(i) \tag{45}$$

and after substituting (45) in (44), we have

$$\det\Gamma^{-1}(i-1) = \det\Gamma^{-1}(i)\left(1 - \alpha Z^T(i)\Gamma(i)Z(i)\right)$$

from which it follows

$$\alpha Z^T(i)\Gamma(i)Z(i) = \frac{\det\Gamma^{-1}(i) - \det\Gamma^{-1}(i-1)}{\det\Gamma^{-1}(i)}. \tag{46}$$

Furthermore, since

$$\det\Gamma^{-1}(i) = \prod_{i=1}^{n+m} \lambda_i\left\{\Gamma^{-1}(i)\right\} \le \lambda_{\max}^{n+m}\left\{\Gamma^{-1}(i)\right\} \tag{47}$$

and

$$r(i) = Tr\left\{\Gamma^{-1}(i)\right\} = \sum_{i=1}^{n+m} \lambda_i\left\{\Gamma^{-1}(i)\right\} \ge \lambda_{\max}\left\{\Gamma^{-1}(i)\right\} \tag{48}$$

one concludes from (47) and (48)

$$\begin{aligned} r(i) &\ge \left[\det\Gamma^{-1}(i)\right]^{\frac{1}{n+m}}, \\ \log^c r(i) &\ge \frac{\log^c\left(\det\Gamma^{-1}(i)\right)}{(n+m)^c}. \end{aligned} \tag{49}$$

Thus, starting from the assumption A5, (46) and (49), one obtains

$$\frac{1}{\alpha}\sum_{i=i_0}^{\infty} \frac{\alpha Z^T(i)\Gamma(i)Z(i)}{\log^c r(i)} \le \frac{(n+m)^c}{\alpha}\sum_{i=i_0}^{\infty} \frac{\det\Gamma^{-1}(i) - \det\Gamma^{-1}(i-1)}{\det\Gamma^{-1}(i)\log^c\left(\det\Gamma^{-1}(i)\right)} \tag{50}$$

The relation (50) is easier to interpret if one views it as a discrete-time approximation to a continuously relation defined as a function of a continuously-time $t$. That is, the sum in (50) should be a reasonable approximation to the corresponding integral, and vice verse, yielding

$$\begin{aligned} \sum_{i=i_0}^{\infty} \frac{\det\Gamma^{-1}(i) - \det\Gamma^{-1}(i-1)}{\det\Gamma^{-1}(i)\log^c\left(\det\Gamma^{-1}(i)\right)} &\le \int_{\det\Gamma^{-1}(i_0)}^{\det\Gamma^{-1}(\infty)} \frac{dt}{t\log^c t} \\ &= \frac{1}{(c-1)\log^{c-1}\left(\det\Gamma^{-1}(i_0)\right)} < \infty \quad \text{w.p.1.} \end{aligned} \tag{51}$$

Here is used the fact that $r(i) = Tr\left\{\Gamma^{-1}(i)\right\}$ is the sequence of non decreasing reals satisfying, due to (40) and A8, $\lim_{i\to\infty} r(i) = \infty$ w.p.1, while the sequence $\det\Gamma^{-1}(i)$ has the same properties, owing to (49). Thus, starting from some $i_0$, one can write $\det\Gamma^{-1}(i) > 0$ for $i \geq i_0$ and $\det\Gamma^{-1}(\infty) = \infty$ w.p.1. Moreover, starting from (41), (50), (51) and A5, one concludes that the assumptions of the convergence Theorem 2 are satisfied.

By applying this theorem on (39), one obtains $\lim_{i\to\infty} z_i = z^*$ w.p.1, where $z^*$ is a finite, nonnegative random variable. Additionally, starting from the definition of $z_i$ in (39) and using the second relation in (32) with $b = Z(i)$, $A = \Gamma^{-1}(i)$ one obtains

$$z_i \geq \frac{\lambda_{\min}\left\{\Gamma^{-1}(i)\right\} \parallel \tilde{\theta}(i) \parallel^2}{\log^k r(i)}. \tag{52}$$

Taking into account the fact that $z_i$ converges towards a finite variable and using the assumption A8, one concludes from (52) that $\lim_{i\to\infty} \parallel \tilde{\theta}(i) \parallel = 0$ w.p.1, which completes the proof.

Moreover, bearing in mind (52), one also concludes that the ratio of estimates convergence is defined by

$$\parallel \tilde{\theta}(i) \parallel = O\left\{\frac{\log^k r(i)}{\lambda_{\min}\left(\Gamma^{-1}(i)\right)}\right\}^{\frac{1}{2}} \tag{53}$$

where $\lim_{|x|\to\infty} O(|x|)/|x| = 0$. Obviously, the expression (53) depends on the parameter $\alpha$ , since $r(i)$ in (29) depends on $\alpha$. Thus, $\alpha$ influences implicitly the rate of estimates convergence, although it is very difficult to find the explicit expression for this dependence.

## References


[1] W. J. Rey, *Robust Statistical Methods*. Springer Verlag, 1977.

[2] V. Barnet and T. V. Lewis, *Outliers in Statistical Data*. London, New York: John Wiley, 1978.

[3] P. J. Huber, *Robust Statistics*. London, New York: John Wiley, 1981.

[4] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*. London, New York: John Wiley, 1987.

[5] E. Handsdin, F. C. Schweppe, J. Kohlas, and A. Fiechter, "Bad data analysis for power system state estimation," *IEEE Trans. on Power Systems and Apparatures*, vol. 94, no. 3, pp. 329–337, 1975.

[6] H. Biernes, *Robust Methods and Asymptotic Theory in Nonlinear Econometrics*. Springer Verlag, 1981.

[7] S. A. Kassam and V. H. Poor, "Robust techniques in signal processing," *Proc. IEEE*, vol. 73, no. 3, pp. 433–481, 1985.

[8] R. D. Martin and D. J. Thomson, "Robust resistant spectrum estimation," *Proc. IEEE*, vol. 70, no. 9, pp. 1097–1014, 1982.

[9] S. S. Stanković and B. D. Kovačević, "Analysis of robust stochastic approximation algorithms for process identification," *Automatica*, vol. 22, no. 4, pp. 483–488, 1986.

[10] C. H. Lee, "On robust linear prediction," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 36, no. 2, pp. 642–650, 1988.

[11] Z. D. Banjac, B. D. Kovačević, M. D. Veinović, and M. M. Milosavljević, "Robust adaptive filtering with variable forgetting factor," *WSEAS Trans. on Circuits and Systems*, vol. 3, no. 2, pp. 223–229, 2004.

[12] S. Chan and Y. Zou, "A recursive least m-estimate algorithm for robust adaptive filtering in impulsive noise: fast algorithm and convergence performance analysis," *IEEE Transaction on Signal Processing*, vol. 52, no. 4, pp. 975–991, 2004.

[13] Ž. M. Đurović and B. D. Kovačević, "Robust estimation with unknown noise statistics," *IEEE Transaction on Automatic Control*, vol. 44, no. 6, pp. 1292–1296, 1999.

[14] V. Ž. Filipović and B. D. Kovačević, "On robustified minimum variance controller," *Internation Journal of Control*, vol. 65, no. 1, pp. 117–129, 1996.

[15] B. D. Kovačević, Ž. M. Đurović, and S. T. Glavaški, "On robust Kalman filtering," *International Journal of Control*, vol. 56, no. 3, pp. 547–562, 1992.

[16] G. A. Williamson, P. M. Clarkson, and W. A. Sethares, "Performance characteristics of median LMS adaptive filter," *IEEE Transaction on Signal Processing*, vol. 41, no. 2, pp. 667–680, 1993.

[17] B. T. Poljak and Ya. Z. Tsypkin, "Robust identification," *Automatica*, vol. 16, no. 2, pp. 53–63, 1980.

[18] Ya. Z. Tsypkin, *Foundations of informational identification theory*. Moscow: Nauka, 1984.

[19] L. Ljung and T. Soderstrom, *Theory and practice of recursive identification*. MIT Press, 1983.

[20] G. C. Goodwin and K. S. Sin, *Adaptive Filtering, Prediction and Control*. Prentice Hall, 1984.

[21] H. Dai and N. K. Sinha, "Robust recursive least-squares method with modified weights for bilinear system identification," *Proc. IEE*, vol. 136, no. 6, pp. 122–126, 1988.

[22] I. P. Kovačević, B. D. Kovačević, and Ž. M. Đurović, "On strong consistency of a class of robust stochastic gradient system identification algorithms," *WSEAS Trans. on Circuits and Systems*, vol. 5, no. 8, pp. 1244–1253, 2006.

[23] J. S. Rusbagi, *Optimizing Methods in Statistics*. Academic Press, 1971.

[24] R. F. Stengel, *Stochastic Optimal Control*. London, New York: John Wiley, 1986.

[25] F. L. Lewis, *Optimal Estimation with an Introduction to Stochastic Control Theory*. London, New York: John Wiley, 1986.