On the Numerical Stability and Accuracy of the Conventional Recursive Least Squares Algorithm

Athanasios P. Liavas and Phillip A. Regalia, *Senior Member, IEEE*

***Abstract*—We study the nonlinear round-off error accumulation system of the conventional recursive least squares algorithm, and we derive bounds for the relative precision of the computations in terms of the conditioning of the problem and the exponential forgetting factor, which guarantee the numerical stability of the finite-precision implementation of the algorithm; the positive definiteness of the finite-precision inverse data covariance matrix is also guaranteed. Bounds for the accumulated round-off errors in the inverse data covariance matrix are also derived. In our simulations, the measured accumulated roundoffs satisfied, in steady state, the analytically predicted bounds. We consider the phenomenon of explosive divergence using a simplified approach; we identify the situations that are likely to lead to this phenomenon; simulations confirm our findings.**

I. Introduction

ADAPTIVE signal processing algorithms are widely used in many application areas because of their ability to adapt to changing and/or unknown environments. The theory of adaptive FIR recursive least squares (RLS) filters is well developed and provides the user, at each time instant, a set of parameters optimal in the least-squares sense [1], [2].

A very important "real-life" problem that is inherent in the continuous use of adaptive algorithms is their behavior in finite-precision environments. This problem contains the following subproblems:

- round-off error generation, which is strongly implementation dependent;
- round-off error propagation, in which a single perturbation is introduced at an arbitrary iteration, and its influence on subsequent computations is studied;
- round-off error accumulation, in which we study the composite influence of the previous two subproblems.

If, for a given algorithm, the round-off errors accumulate without bound, then the algorithm is unsuitable for continuous use without further intervention.

For the conventional RLS algorithms, the round-off error propagation is the best studied of the three aforementioned subproblems [3]–[5]. Such studies typically examine the *linearized* round-off error propagation system and focus on the derivation of its *exponential stability*; this, in turn, implies *local* exponential stability of the corresponding *nonlinear* round-off error propagation system [6, pp. 17–19]. No clear indication has surfaced, however, as to "how small" the accumulated error should be so that the influence of the nonlinear terms does not destroy the stability properties of the overall system.

An examination of the nonlinear round-off error accumulation system of the conventional RLS algorithm appeared in [7], where a scenario for *explosive divergence* was developed. Explosive divergence is the occurence of "sudden big" errors due to finite-precision effects. This phenomenon has been linked to the loss of the positive definiteness of the finite-precision inverse data covariance matrix and the negative value of a theoretically positive quantity [7]. However, the approach of [7] is mostly qualitative, and finally, numerical stability is not guaranteed.

This is our main subject. We perform a detailed study of the stability properties of the *nonlinear* round-off error accumulation system of the conventional RLS algorithm and we derive the following:

- an upper bound for the relative precision of the computations in terms of
  - —the conditioning of the problem;
  - —the exponential forgetting factor;

  which guarantees that the nonlinear round-off error accumulation system remains BIBO stable; it is also guaranteed that the finite-precision inverse data covariance matrix remains positive definite;
- a corresponding upper bound for the accumulated round-off error.

In Section II, we review the linear least squares problem and two classical RLS algorithms. In Section III, we derive the nonlinear round-off error accumulation system of the conventional RLS algorithm, which is studied in Section IV. In Section V, we present simulation results, and conclusions are drawn in Section VI.

II. Recursive Least Squares Algorithms

For the standard least squares problem, we are given a sequence of M-dimensional input vectors ϕ_t, plus a reference sequence u_t, $t = 1,\ldots,k$ and are asked to compute an M-dimensional parameter vector θ_k such that

$$\theta_k \triangleq \arg\min_{\theta\in R^M} \sum_{t=1}^{k} \lambda^{k-t}(u_t - \theta^t\phi_t)^2 \tag{1}$$

where λ is the so-called forgetting factor $(0 \ll \lambda < 1)$.

Manuscript received May 5, 1997; revised June 8, 1998. This work was supported by the Training and Mobility of Researchers (TMR) Program of the European Commission under Contract ERBFMBICT960659. The associate editor coordinating the review of this paper and approving it for publication was Prof. M. H. Er.

The authors are with the Département Signal et Image, Institut National des Télécommunications, Evry, France (e-mail: liavas@sim.int-evry.fr; regalia@sim.int-evry.fr.).

Publisher Item Identifier S 1053-587X(99)00138-5.

The recursive solution of (1) gives rise to an RLS algorithm, for which a standard form appears as

$$\theta_k = \theta_{k-1} + R_k^{-1}\phi_k\big(u_k - \theta_{k-1}^t\phi_k\big) \tag{2}$$

$$R_k = \lambda R_{k-1} + \phi_k\phi_k^t. \tag{3}$$

The conventional recursive least squares (CLS) algorithm results if we introduce $P_k = R_k^{-1}$ and apply the matrix inversion lemma to (3); this gives the familiar recursions

$$r_k^e = \lambda + \phi_k^t P_{k-1}\phi_k, \tag{4}$$

$$\theta_k = \theta_{k-1} + \frac{P_{k-1}\phi_k}{r_k^e}\big(u_k - \theta_{k-1}^t\phi_k\big) \tag{5}$$

$$P_k = \frac{1}{\lambda}\left(P_{k-1} - \frac{P_{k-1}\phi_k\phi_k^t P_{k-1}}{r_k^e}\right). \tag{6}$$

In order to study the finite-precision implementation of the CLS algorithm, we denote by $\widetilde{P}_k$ the finite-precision version of P_k, we define the intermediate quantity

$$\widetilde{r}_k^e \triangleq \lambda + \phi_k^t\widetilde{P}_{k-1}\phi_k \tag{7}$$

and we express the finite-precision time update of P_k as

$$\widetilde{P}_k = \frac{1}{\lambda}\left(\tilde{P}_{k-1} - \frac{\tilde{P}_{k-1}\phi_k\phi_k^t\tilde{P}_{k-1}}{\tilde{r}_k^e}\right) + \epsilon\tilde{P}_k \tag{8}$$

where the term $\epsilon\tilde{P}_k$ denotes the local round-off error in the computation of P_k; in this way, we separate the propagated from the local round-off errors. In Appendix A, we perform a detailed analysis of the finite-precision time update of P_k, and we compute the local round-off error $\epsilon\tilde{P}_k$.

III. The Nonlinear Round-Off Error Accumulation System

Let us denote by Δx the accumulated round-off error in the quantity x. Then

$$\Delta P_k \triangleq \tilde{P}_k - P_k, \tag{9}$$

$$\Delta r_k^e \triangleq \tilde{r}_k^e - r_k^e = \phi_k^t\Delta P_{k-1}\phi_k. \tag{10}$$

Assuming that $\left|\frac{\Delta r_k^e}{r_k^e}\right| < 1$, we can expand the second term of (8) as (11), shown at the bottom of the page.[1] Thus, the nonlinear round-off error accumulation system is described by (12), shown at the bottom of the page. The study of (12) is of primordial importance for the "real-life" finite-precision implementation of the CLS algorithm. However, it seems that the existence of the higher order terms has been a major obstacle toward this purpose. In the sequel, we perform a detailed study of the nonlinear difference equation (12), and we derive sufficient conditions for its BIBO stability, implying numerical stability of the CLS algorithm.

A. Assumptions

In order to study (12), the following assumptions are invoked.

1) The regression vector ϕ_t is persistently exciting, that is, there exist a, b and k_0 such that $0 < a < b < \infty$ and

$$aI \le \sum_{t=1}^{k}\lambda^{k-t}\phi_t\phi_t^t \le bI, \quad \text{for all } k > k_0. \tag{13}$$

[1] Formulas (8) and (9) of [8] contain typos.

$$\frac{\tilde{P}_{k-1}\phi_k\phi_k^t\tilde{P}_{k-1}}{\tilde{r}_k^e} = \frac{P_{k-1}\phi_k\phi_k^tP_{k-1} + P_{k-1}\phi_k\phi_k^t\Delta P_{k-1} + \Delta P_{k-1}\phi_k\phi_k^tP_{k-1} + \Delta P_{k-1}\phi_k\phi_k^t\Delta P_{k-1}}{r_k^e} \times \underbrace{\left\{\underbrace{\underbrace{1 - \frac{\Delta r_k^e}{r_k^e}}_{t_1(k,\Delta P_{k-1})} + \underbrace{\left(\frac{\Delta r_k^e}{r_k^e}\right)^2 - \cdots}_{t_2(k,\Delta P_{k-1})}}_{t_0(k,\Delta P_{k-1})}\right\}}. \tag{11}$$

$$\Delta P_k = \underbrace{\frac{1}{\lambda}\left\{\Delta P_{k-1} + \frac{P_{k-1}\phi_k\phi_k^tP_{k-1}}{r_k^e}\frac{\Delta r_k^e}{r_k^e} - \frac{P_{k-1}\phi_k\phi_k^t\Delta P_{k-1} + \Delta P_{k-1}\phi_k\phi_k^tP_{k-1}}{r_k^e}\right\}}_{\text{first order terms}} - \underbrace{\frac{1}{\lambda}\left\{\frac{P_{k-1}\phi_k\phi_k^tP_{k-1}}{r_k^e}t_2(k,\Delta P_{k-1}) + \frac{P_{k-1}\phi_k\phi_k^t\Delta P_{k-1}}{r_k^e}t_1(k,\Delta P_{k-1})\right.}_{\text{higher order}} \underbrace{\left. + \frac{\Delta P_{k-1}\phi_k\phi_k^tP_{k-1}}{r_k^e}t_1(k,\Delta P_{k-1}) + \frac{\Delta P_{k-1}\phi_k\phi_k^t\Delta P_{k-1}}{r_k^e}t_0(k,\Delta P_{k-1})\right\}}_{\text{terms}} + \epsilon\tilde{P}_k. \tag{12}$$

This is a well-known condition, which implies that the data covariance matrix and its inverse exist and are bounded. Thus, there exist bounded constants $\mathcal{R}$ and $\mathcal{P}$ such that

$$\|R_k\| \leq \mathcal{R}, \quad \text{and} \quad \|P_k\| \leq \mathcal{P}, \quad \text{for all } k \tag{14}$$

where, here and throughout, $\|\cdot\|$ denotes the 1-norm of its argument. The validity of the bounds in (14) for $k < k_0$ can be guaranteed by a "soft" start.

2) The regression vector ϕ_t is bounded as

$$\|\phi_t\| \leq \Phi, \quad \text{for all } t. \tag{15}$$

In Appendix A, we show that if the *relative precision* of the computations, which is denoted by ϵ, belongs to a certain interval, then the local round-off $\epsilon\tilde{P}_k$ is bounded as

$$\|\epsilon\tilde{P}_k\| \leq \mathcal{E}\epsilon, \quad \text{for all } k \tag{16}$$

where $\mathcal{E}$ is a bounded constant.

Using (3) and (15), we see that an upper bound for $\mathcal{R}$ is given by

$$\mathcal{R} \leq \bar{\mathcal{R}} \triangleq \frac{\Phi^2}{1-\lambda} \tag{17}$$

and an upper bound for the condition number of the data covariance matrix $\mathcal{K}_k \triangleq \|R_k\|\|P_k\|$ is provided by

$$\mathcal{K}_k \leq \mathcal{K} \triangleq \mathcal{R}\mathcal{P} \leq \bar{\mathcal{K}} \triangleq \frac{\mathcal{P}\Phi^2}{1-\lambda}. \tag{18}$$

IV. Stability Analysis of the Nonlinear Round-Off Error Accumulation System

Let us denote by $f(k, \Delta P_{k-1})$ the higher order terms appearing in (12). Using (10), we can rearrange (12) as

$$\Delta P_k = \frac{1}{\lambda}\left(I - \frac{P_{k-1}\phi_k\phi_k^t}{r_k^e}\right)\Delta P_{k-1}\left(I - \frac{\phi_k\phi_k^t P_{k-1}}{r_k^e}\right) + f(k, \Delta P_{k-1}) + \epsilon\tilde{P}_k. \tag{19}$$

From (6), we obtain

$$I - \frac{P_{k-1}\phi_k\phi_k^t}{r_k^e} = \lambda P_k R_{k-1} \tag{20}$$

so that (19) becomes

$$\Delta P_k = \lambda P_k R_{k-1}\Delta P_{k-1} R_{k-1} P_k + f(k, \Delta P_{k-1}) + \epsilon\tilde{P}_k. \tag{21}$$

Looking for a moment at the linearized homogeneous system, we see that

$$\Delta P_k = \lambda^{k-i} P_k R_i \Delta P_i R_i P_k \tag{22}$$

which gives

$$\begin{aligned}\|\Delta P_k\| &\leq \lambda^{k-i}\|P_k R_i\|\|\Delta P_i\|\|R_i P_k\| \\ &\leq \lambda^{k-i}\|P_k\|^2\|R_i\|^2\|\Delta P_i\| \leq \lambda^{k-i}\mathcal{K}^2\|\Delta P_i\|.\end{aligned} \tag{23}$$

Thus, if the data covariance matrix is bounded from above and below, then the linearized round-off error propagation system is exponentially stable, with base of decay λ [3].[2] This implies that the nonlinear round-off error propagation system is *locally* (i.e., for small ΔP_k) exponentially stable. However, no study exists, to our knowledge, that provides an indication as to "how small" ΔP_k should be so that the influence of the nonlinear and the additive terms in (21) does not destroy these stability properties. This is our main task in what follows.

At first, we derive the solution of (21) as

$$\Delta P_k = \sum_{i=1}^{k} \lambda^{k-i} P_k R_i (f(i, \Delta P_{i-1}) + \epsilon\tilde{P}_i) R_i P_k \tag{24}$$

where $f(1, \Delta P_0) = 0$; this can be trivially forced by initializing the algorithm with P_0, which can be represented with no round-off error. Then, if there exists a constant r such that $\|\Delta P_i\| < r$ for $i = 1, \ldots, k-1$, the following theorem provides an upper bound for $\|\Delta P_k\|$.

Theorem 1: If $\|\Delta P_i\| < r$ for $i = 1, \ldots, k-1$ and $r < \frac{\lambda}{\Phi^2}$, then

$$\|\Delta P_k\| < \frac{\overbrace{\Phi^2(\mathcal{K} + \mathcal{P}\Phi^2)^2}^{\mathcal{A}_1} r^2}{\lambda(1-\lambda)(\lambda - \Phi^2 r)} + \frac{\mathcal{K}^2\mathcal{E}\epsilon}{1-\lambda}. \tag{25}$$

The proof is given in Appendix B.

Now, if we can find an r independent of k in the range $0 < r < \frac{\lambda}{\Phi^2}$ such that the right-hand side of (25) is less than or equal to r, i.e.,

$$\|\Delta P_k\| < \frac{\mathcal{A}_1 r^2}{\lambda(1-\lambda)(\lambda - \Phi^2 r)} + \frac{\mathcal{K}^2\mathcal{E}\epsilon}{1-\lambda} \leq r \tag{26}$$

then we obtain by induction that $\|\Delta P_k\| < r$ for all k, thereby implying that the accumulated round-off error remains bounded.

If we momentarily set $\epsilon = 0$ in (26), we deduce that r can be no larger than

$$r \leq r_0 \triangleq \frac{\lambda^2(1-\lambda)}{\mathcal{A}_1 + \underbrace{\lambda(1-\lambda)\Phi^2}_{\mathcal{A}_2}}. \tag{27}$$

Next, for each $r \in (0, r_0]$, the upper bound on ϵ, for which (26) remains true and, hence, $\|\Delta P_k\| < r$ for all k, is found as

$$\epsilon \leq \frac{1-\lambda}{\mathcal{K}^2\mathcal{E}} \underbrace{\left(r - \frac{\mathcal{A}_1 r^2}{\lambda(1-\lambda)(\lambda - \Phi^2 r)}\right)}_{F_0(r)}. \tag{28}$$

In order to maximize the relative precision ϵ, that is, minimize the wordlength, which guarantees that the accumulated round-off error remains bounded, we have to maximize the function $F_0(r)$ in the interval $(0, r_0]$. The extremal points of $F_0(r)$ are the solutions to $dF_0(r)/dr = 0$, leading to the second-order equation

$$\Phi^2(\mathcal{A}_1 + \mathcal{A}_2)r^2 - 2\lambda(\mathcal{A}_1 + \mathcal{A}_2)r + \frac{\lambda^2}{\Phi^2}\mathcal{A}_2 = 0. \tag{29}$$

[2] For the $\lambda = 1$ case, see [1, p. 756] and [9].

This has as solutions

$$r = \rho_1^0 = \frac{\lambda}{\Phi^2}\left(1 - \frac{\sqrt{\mathcal{A}_1^2 + \mathcal{A}_1\mathcal{A}_2}}{\mathcal{A}_1 + \mathcal{A}_2}\right)$$
$$r = \rho_2^0 = \frac{\lambda}{\Phi^2}\left(1 + \frac{\sqrt{\mathcal{A}_1^2 + \mathcal{A}_1\mathcal{A}_2}}{\mathcal{A}_1 + \mathcal{A}_2}\right) \tag{30}$$

and the maximum of $F_0(r)$ is attained at $r = \rho_1^0 < r_0$. Thus, an upper bound for the relative precision that quarantees BIBO stability of the round-off error accumulation system given $\mathcal{P}$, $\mathcal{R}$, Φ, λ, and $\mathcal{E}$ is provided by

$$\epsilon \le \epsilon_0 \triangleq \frac{1-\lambda}{\mathcal{K}^2\mathcal{E}} F_0(\rho_1^0). \tag{31}$$

The corresponding bound for the accumulated round-off error is

$$||\Delta P_k|| < \rho_1^0, \quad \text{for all } k. \tag{32}$$

Remark: Bound (31) seems to be conservative, mainly because during the calculations in Appendix B, we have used the condition number $\mathcal{K}$ as an upper bound for $||P_k R_i||$. We emphasize that the bound so obtained applies in the general nonstationary case. A sharper bound, however, can be obtained if the input data are stationary, as we now pursue.

A. The Stationary Case

When the input sequence ϕ_t is stationary, then the approximation

$$P_k R_{k-1} \approx I \tag{33}$$

is often used in steady-state and for λ very close to 1 [1, p. 713], [10]. This approximation affords the derivation of bounds that are much less pessimistic than (31) and (32).

In this case, the round-off error accumulation system is given by

$$\Delta P_k = \lambda \Delta P_{k-1} + f(k, \Delta P_{k-1}) + \epsilon \tilde{P}_k \tag{34}$$

and thus

$$||\Delta P_k|| \le \frac{1}{1-\lambda}\Big(\max_i ||f(i, \Delta P_{i-1})|| + \mathcal{E}\epsilon\Big). \tag{35}$$

Assuming that $||\Delta P_i|| < r$ for $i = 1, \ldots, k-1$, the next theorem provides a bound for $||\Delta P_k||$.

Theorem 2: Let ϕ_t be a stationary sequence, λ be very close to 1, $||\Delta P_i|| < r$ for $i = 1, \ldots, k-1$, and $r < \frac{\lambda}{\Phi^2}$. Then

$$||\Delta P_k|| < \frac{\overbrace{\Phi^2((1-\lambda)\mathcal{P}\Phi^2 + 3 - 2\lambda)}^{\alpha_1} r^2}{\lambda(1-\lambda)(\lambda - \Phi^2 r)} + \frac{\mathcal{E}\epsilon}{1-\lambda} (\le r). \tag{36}$$

The proof is given in Appendix C.

BIBO stability is guaranteed if the right-hand side of (36) is less than or equal to r. Putting in (36) $\epsilon = 0$, gives an upper bound for r as

$$r \le r_1 \triangleq \frac{\lambda^2(1-\lambda)}{\alpha_1 + \mathcal{A}_2}. \tag{37}$$

For each $r \in (0, r_1]$, if

$$\epsilon \le \frac{1-\lambda}{\mathcal{E}} \underbrace{\left(r - \frac{\alpha_1 r^2}{\lambda(1-\lambda)(\lambda - \Phi^2 r)}\right)}_{F_1(r)} \tag{38}$$

then $||\Delta P_k|| < r$ for all k. We can derive an upper bound for ϵ by using the maximization point of $F_1(r)$ for $r \in (0, r_1]$ as

$$\epsilon \le \epsilon_1 \triangleq \frac{1-\lambda}{\mathcal{E}} F_1(\rho_1^1) \tag{39}$$

where

$$\rho_1^1 = \frac{\lambda}{\Phi^2}\left(1 - \frac{\sqrt{\alpha_1^2 + \alpha_1\mathcal{A}_2}}{\alpha_1 + \mathcal{A}_2}\right). \tag{40}$$

The corresponding bound for the round-off accumulated error is

$$||\Delta P_k|| < \rho_1^1, \quad \text{for all } k. \tag{41}$$

Remark: Bound (39) is much less conservative than (31), mainly because during the calculations in Appendix C, we have approximated $||P_k R_{k-1}||$ by 1 and not by $\mathcal{K}$, which was the case in Appendix B.

By simple manipulations of (37) and (38) and using (71) from Appendix A, we can show that

$$\epsilon \le \epsilon_1 < \frac{1-\lambda}{\mathcal{K}} \tag{42}$$

which establishes a relation between the conditioning of the problem and the numerical stability of the CLS algorithm; this is a well-known claim in the general context of adaptive algorithms [1, p. 738], [11]. We also observe that the relative precision is proportional to $1-\lambda$, which means that for λ very close to 1, the round-off error accumulation is more significant, which is as to be expected.

It has been widely observed that the sudden big errors due to finite-precision effects are related to the loss of the positive definiteness of the finite-precision inverse data covariance matrix $\tilde{P}_k$ [7]. In the sequel, we show that if we use (39) for the calculation of the relative precision of the computations, then $\tilde{P}_k$ remains positive definite.

Theorem 3: If a persistently exciting data sequence ϕ_t is fed as input to the CLS algorithm, which is implemented with relative precision ϵ calculated by (39), then the finite-precision inverse data covariance matrix $\tilde{P}_k$ remains positive definite for all k.

Proof: It is a well-known result in matrix perturbation theory [12, p. 118] that if

$$\tilde{A} = A + \Delta A \quad \text{and} \quad ||\Delta A|| < \frac{1}{||A^{-1}||} \tag{43}$$

then $\tilde{A}$ is perforce nonsingular.

Thus, if $||\Delta P_k||\ \mathcal{R} < 1$, then $||\Delta P_k||\ ||R_k|| < 1$, meaning that $\tilde{P}_k$ is perforce nonsingular. From (36) and (37), we obtain

$$||\Delta P_k|| < r_1 = \frac{\lambda^2(1-\lambda)}{\alpha_1 + \mathcal{A}_2} < \frac{1-\lambda}{\alpha_1} < \frac{1-\lambda}{\Phi^2} = \frac{1}{\bar{\mathcal{R}}} \le \frac{1}{\mathcal{R}} \tag{44}$$

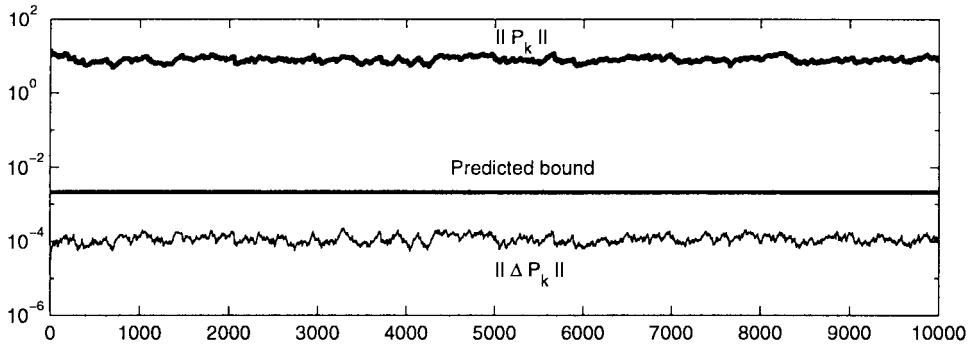

Fig. 1. Twenty-bit precision; 1-norm of P_k (thick line), 1-norm of ΔP_k (solid line), and bound predicted by formula (37).

which means that the sufficient condition (43) is always satisfied, implying nonsingularity of the finite-precision inverse data covariance matrix $\tilde{P}_k$ for all k. This fact implies that

$$||\Delta P_k||_2 < \frac{1}{||R_k||_2} = \lambda_{\min}(P_k) \tag{45}$$

where $||\cdot||_2$ denotes the matrix 2-norm and $\lambda_{\min}(P_k)$ the minimum eigenvalue of P_k because we can always find ΔP_k such that

$$||\Delta P_k||_2 = \frac{1}{||R_k||_2} = \lambda_{\min}(P_k) \tag{46}$$

which renders $\tilde{P}_k = P_k + \Delta P_k$ nonsingular [12, p. 120]. Then, since

$$|\lambda_i(\Delta P_k)| \leq ||\Delta P_k||_2 < \lambda_{\min}(P_k)$$

we obtain [13, p. 411]

$$\lambda_{\min}(\tilde{P}_k) \geq \lambda_{\min}(P_k) + \lambda_{\min}(\Delta P_k) > 0 \tag{47}$$

implying that the finite-precision inverse data covariance matrix remains positive definite. □

V. Simulations

In the previous section, we studied the nonlinear round-off error accumulation system of the CLS algorithm, and we derived an upper bound on the relative precision ϵ_0 (resp., ϵ_1), which guarantees that the accumulated round-off error is bounded by ρ_1^0 (resp. ρ_1^1). This gives sufficient conditions for the BIBO stability of the round-off error accumulation system and provides an upper bound for the accuracy of the computations. In this section, we perform simulations to check our theoretical results.

In the first simulation study, we generate input data by passing white Gaussian noise with standard deviation 0.1, through the AR model with poles 0.85, $0.7 \pm .4j$, and $-0.4 \pm .6j$. We run, in double-precision floating-point arithmetic, the CLS algorithm with order $M = 5$ and $\lambda = 0.99$, and we derive the following estimates: $\mathcal{P} = 8.0467$, $\Phi = 1.3913$, yielding $\bar{\mathcal{K}} = \frac{\mathcal{P}\Phi^2}{1-\lambda} = 1.5577 \times 10^3$. Then, we use (39), (41), and (71) to derive upper bounds for the relative precision and the accumulated round-off error as $\epsilon_1 = 1.3322 \times 10^{-6}$ and $\rho_1^1 = 0.0021$, respectively. We implement the CLS algorithm with 20-bit precision floating-point arithmetic, as indicated by the value of ϵ_1 (after each floating-point operation, we truncate the mantissa to 20 bits without affecting the exponent). In Fig. 1, we plot the 1-norm of the inverse data covariance matrix $||P_k||$, the 1-norm of the accumulated round-off error $||\Delta P_k||$, and the bound predicted by our theory; we observe that the bound is always satisfied (we used the double precision variables as the reference variables). We have repeated this experiment with many different types of data and for millions of iterations; the accumulated error satisfied, in steady-state, the bound predicted by our theory.

An interesting question, which is not answered directly by our theory, is if these sufficient conditions for BIBO stability are necessary as well. We attempt to give an answer to this question using a simplified approach. An approximate round-off error accumulation system for the CLS algorithm can be given as

$$\Delta P_k = \lambda \Delta P_{k-1} + \epsilon \tilde{P}_k. \tag{48}$$

This expression is more likely to be valid in cases in which the round-off terms dominate the nonlinear terms; a careful inspection of (36) reveals that this may happen when Φ is neither very large nor very small (this can be achieved by scaling the input data); Φ should not be very large because it magnifies the nonlinear terms, and it should not be very small because, in that case, for a given condition number, $\mathcal{P}$ becomes very large; thus, $||\Delta P_k||$ may become large, resulting in a domination of the nonlinear terms with respect to the local round-off terms. Then, using (71) from Appendix A, we obtain

$$||\Delta P_k|| \leq \frac{\mathcal{P}\epsilon}{1-\lambda}. \tag{49}$$

A usual rule of thumb is that if a numerical analysis round-off error bound is $A\epsilon$, then it is more realistic to expect that the roundoff is typically of order $\sqrt{A}\epsilon$ [14, p. 52]. This results from the independence of the various roundoffs and the central limit theorem. Thus, a more realistic estimation of $||\Delta P_k||$ is

$$||\Delta P_k|| \approx \sqrt{\frac{\mathcal{P}}{1-\lambda}}\epsilon. \tag{50}$$

Since the matrix 1-norm is subordinate to the vector 1-norm, there is a matrix ΔA with $||\Delta A|| = \frac{1}{||A^{-1}||}$ such that $A + \Delta A$ is nonsingular [12, p. 120]. If $||\Delta P_k|| \approx \frac{1}{\bar{\mathcal{K}}}$, then we are close

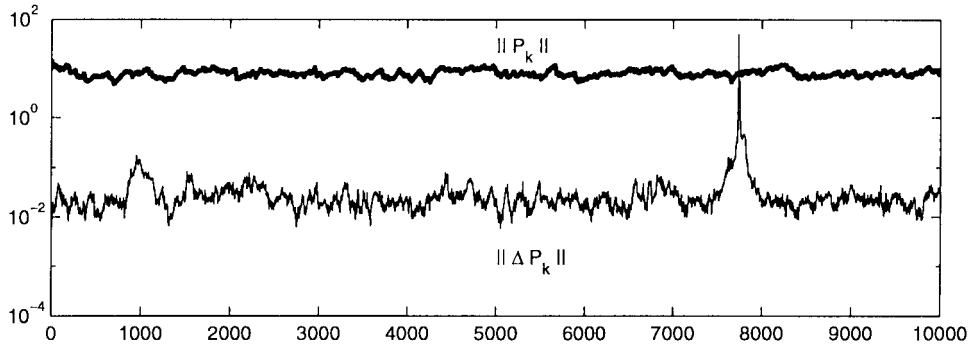

Fig. 2. Occurence of explosive divergence with 11-bit precision; 1-norm of P_k (thick line), 1-norm of ΔP_k (solid line).

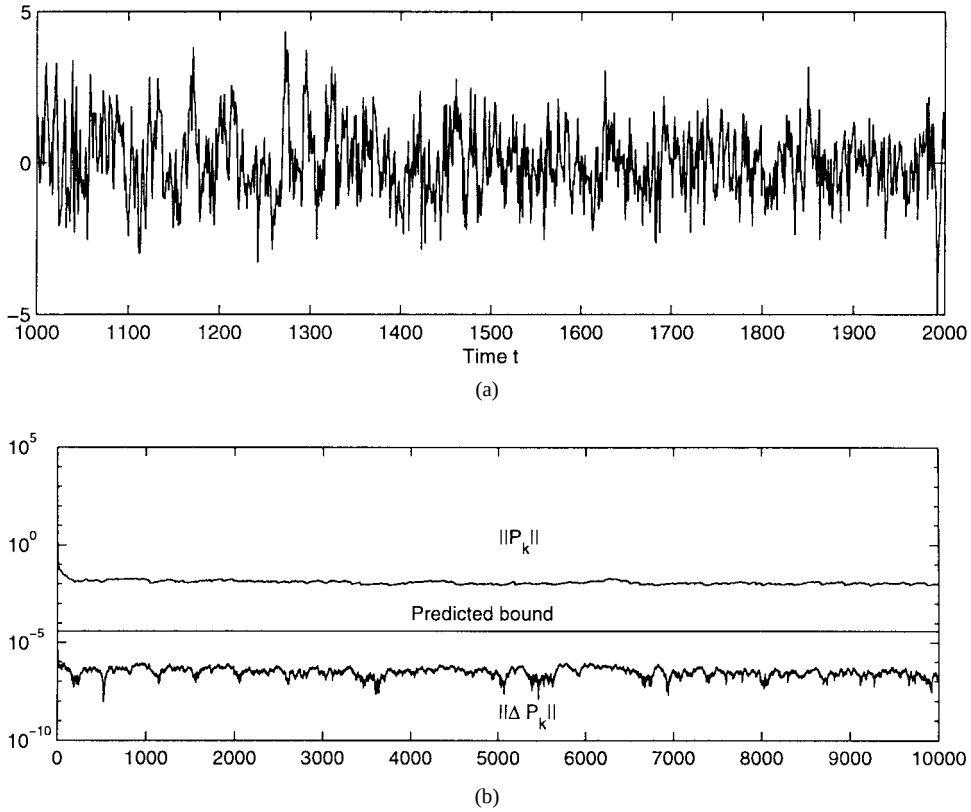

Fig. 3. (a) Nonstationary AR data. (b) Eighteen-bit precision: 1-norm of P_k (thick line), 1-norm of ΔP_k (solid line) and bound predicted by the stationary case.

to loss of the positive definiteness of $\tilde{P}_k$. Thus, if

$$\epsilon \approx \frac{\sqrt{1-\lambda}}{\sqrt{\mathcal{P}\mathcal{R}}} \tag{51}$$

then it is likely that explosive divergence may occur.

In order to check this approach, we use the same data set. Now, we compute the relative precision using (51), yielding $\epsilon = 6.4608 \times 10^{-4}$, implying 11-bit precision; we expect that explosive divergence will occur. We plot the results in Fig. 2. We observe that the accumulated roundoff exhibits a "stationary" behavior, and suddenly, a "big error" occurs. Then, the algorithm reconverges and the same scenario may re-commence. At the moments of "big" errors, the finite-precision inverse data covariance matrix loses its positive definiteness, as has been observed in the literature, e.g., [7]. Similar behavior has been observed in most of the cases studied in our simulations.

For data obtained by slowly time-varying systems, the approximation

$$P_k R_{k-1} \approx I$$

holds for λ close to 1, and the results derived for the stationary case can be applied to these cases as well. We present such a case in Fig. 3. In Fig. 3(a), we plot a segment of the data derived by passing 10^4 samples of a zero-mean unit-variance white Gaussian noise sequence e_t through the second order AR model

$$y_t = \sum_{i=1}^{2} a_i(t) y_{t-i} + e_t \tag{52}$$

whose coefficients vary according to the model

$$\begin{aligned} a_i(t) &= 0.999 * a_i(t-1) + w_i(t), \quad i = 1, 2 \\ a_1(1) &= 1.4, \quad a_2(1) = 0.65 \end{aligned} \tag{53}$$

where $w_i(t)$ is a white noise sequence with variance $\sigma_w^2 = 10^{-4}$. We use (39), (41), and (71) to derive $\epsilon_1 = 5.0783 \times 10^{-6}$, implying 18-bit precision, and $\rho_1^1 = 3.9985 \times 10^{-5}$. In Fig. 3(b), we plot $\|P_k\|$, $\|\Delta P_k\|$ and the bound predicted by the theory concerning the stationary case. We observe that the bound is satisfied. However, we note that in the general nonstationary case, these results may not apply, and bound (31) should be used.

VI. Conclusion

We considered the problem of the finite-precision implementation of the CLS algorithm. Most previous studies have focused on the study of the linearized round-off error propagation system and, thus, fall short of establishing conditions guaranteeing the numerical stability of the finite-precision implementation of the algorithm.

We derived upper bounds for the relative precision of the computations, which guarantee the BIBO stability of the nonlinear round-off error accumulation system, implying numerical stability of the implementation. These bounds depend on the conditioning of the problem through the quantities $\mathcal{P}$, $\mathcal{R}$, and Φ and the forgetting factor λ. Preservation of the positive definiteness of the finite-precision inverse data covariance matrix is also guaranteed.

Our approach resembles a numerical analysis one, where the derivation of the bounds is based on the application of the triangle and submultiplicative norm inequalities. This fact makes the derived bounds somewhat conservative, especially in the general nonstationary case. In all the simulations we performed, the accumulated round-off error satisfied, in steady-state, the analytically predicted bounds.

We considered the phenomenon of explosive divergence using a simplified approach; we identified the conditions that are likely to lead to this kind of numerical instability; simulations are in agreement with our findings.

Appendix A

In order to compute the round-off error generated during one iteration of the CLS algorithm, we must model the respective floating-point matrix operations. Assuming that A and B are $(M \times M)$ matrices, x and y are M-dimensional vectors, and a is a scalar, we obtain [14, pp. 69–76]

$$\begin{aligned} fl[x^t y] &= (x + \delta x)^t y = x^t (y + \delta y) \\ \|\delta x\| &\le \gamma_M \|x\|, \quad \|\delta y\| \le \gamma_M \|y\| \end{aligned} \tag{54}$$

$$fl[Ab] = (A + \delta A)b, \quad \|\delta A\| \le \gamma_M \|A\| \tag{55}$$

$$fl[xy^t] = xy^t + \delta, \quad \|\delta\| \le \epsilon \|xy^t\|. \tag{56}$$

It can be shown that

$$\begin{aligned} fl[a(A-B)] &= a(A-B) + \delta \\ \|\delta\| &\le 2\epsilon \|a(A-B)\| + O(\epsilon^2). \end{aligned} \tag{57}$$

(For the definition of γ_M and a discussion of the importance of the constant and the $O(\epsilon^2)$ terms in the error bounds, see [14, pp. 70–74]). Then

$$fl\big(\phi_k^t P_{k-1} \phi_k\big) = (\phi_k + \delta\phi_k)^t (P_{k-1} + \delta P_{k-1}) \phi_k \tag{58}$$

where $\|\delta\phi_k\| \le \gamma_M \|\phi_k\|$, $\|\delta P_{k-1}\| \le \gamma_M \|P_{k-1}\|$, and

$$\begin{aligned} &fl\big(P_{k-1}\phi_k\phi_k^t P_{k-1}\big) \\ &\quad = (P_{k-1} + \delta P_{k-1})\phi_k\phi_k^t (P_{k-1} + \delta P_{k-1}) + \delta_1 \end{aligned} \tag{59}$$

where $\|\delta_1\| \le \epsilon \|P_{k-1}\phi_k\phi_k^t P_{k-1}\| + O(\epsilon^2)$. The computed $\hat{P}_k$ can be expressed as

$$\begin{aligned} \hat{P}_k = \frac{1}{\lambda}\Bigg(& P_{k-1} \\ & - \frac{(P_{k-1} + \delta P_{k-1})\phi_k\phi_k^t (P_{k-1} + \delta P_{k-1}) + \delta_1}{\lambda + (\phi_k + \delta\phi_k)^t (P_{k-1} + \delta P_{k-1})\phi_k + \delta_2} \Bigg) + \delta_3 \end{aligned} \tag{60}$$

with $|\delta_2| \le \epsilon r_k^e + O(\epsilon^2)$, and $\|\delta_3\| \le 2\epsilon \|P_k\| + O(\epsilon^2)$ so that $\hat{P}_k$ becomes

$$\hat{P}_k = \frac{1}{\lambda}\left(P_{k-1} - \frac{P_{k-1}\phi_k\phi_k^t P_{k-1} + \Delta_1 + \delta_1}{\lambda + \phi_k^t P_{k-1}\phi_k + \Delta_2} \right) + \delta_3 \tag{61}$$

where $\|\Delta_1\| \le 2\gamma_M \|P_{k-1}\| \|\phi_k\phi_k^t P_{k-1}\| + O(\epsilon^2)$, and $\|\Delta_2\| \le 2\gamma_M \|\phi_k^t\| \|P_{k-1}\| \|\phi_k\| + O(\epsilon^2)$. (Term δ_2 has been considered small with respect to Δ_2 and has been neglected). Thus

$$\begin{aligned} \hat{P}_k = \frac{1}{\lambda}\Bigg\{ & P_{k-1} - \frac{1}{r_k^e}\left(1 - \frac{\Delta_2}{r_k^e} + O(\epsilon^2)\right) \\ & \times \left(P_{k-1}\phi_k\phi_k^t P_{k-1} + \Delta_1 + \delta_1\right) \Bigg\} + \delta_3. \end{aligned} \tag{62}$$

The error due to roundoff is

$$\begin{aligned} \hat{P}_k - P_k &= \frac{P_{k-1}\phi_k\phi_k^t P_{k-1}}{\lambda r_k^e}\frac{\Delta_2}{r_k^e} - \frac{\Delta_1}{\lambda r_k^e} - \frac{\delta_1}{\lambda r_k^e} + \delta_3 + O(\epsilon^2) \\ &= \left(P_k - \frac{1}{\lambda}P_{k-1}\right)\frac{\Delta_2}{r_k^e} - \frac{\Delta_1}{\lambda r_k^e} - \frac{\delta_1}{\lambda r_k^e} + \delta_3 + O(\epsilon^2). \end{aligned} \tag{63}$$

In the nonstationary case

$$\begin{aligned} \left\| P_k - \frac{1}{\lambda}P_{k-1} \right\| &\le \frac{1+\lambda}{\lambda}\mathcal{P}, \quad \|\Delta_1\| \le 2\gamma_M \mathcal{P}^2\Phi^2 + O(\epsilon^2) \\ \|\Delta_2\| &\le 2\gamma_M \mathcal{P}\Phi^2 + O(\epsilon^2) \end{aligned} \tag{64}$$

and

$$\left\| \frac{\delta_1}{\lambda r_k^e} \right\| = \epsilon \left\| P_k - \frac{1}{\lambda}P_{k-1} \right\| + O(\epsilon^2) \le \frac{1+\lambda}{\lambda}\mathcal{P} + O(\epsilon^2). \tag{65}$$

In order to derive the simplest possible bound, we ignore the small multiplicative terms, i.e., $\frac{1+\lambda}{\lambda}$, $\frac{1}{\lambda r_k^e}$, 2, and the $O(\epsilon^2)$ terms, and we replace the γ_M term, which is strongly dependent on the summation strategy [14, p. 70], by ϵ. Then

$$\|\hat{P}_k - P_k\| \le \epsilon\mathcal{E} \equiv \epsilon(\mathcal{P}^2\Phi^2 + \mathcal{P}). \tag{66}$$

Note the difference between a first-order approximation to the relative precision ϵ, which is assumed to be "small," and a

first-order approximation to the accumulated round-off error performed in most of the existing studies, which, however, may be much larger than ϵ.

In the stationary case and for λ very close to 1, using (33) and (83), we derive

$$\begin{aligned} \left\| P_k - \frac{1}{\lambda} P_{k-1} \right\| &\approx (1-\lambda)\mathcal{P} \\ \|\Delta_1\| &\le 2\gamma_M (1-\lambda)\mathcal{P} + O(\epsilon^2) \\ \left\| \frac{\delta_1}{\lambda r_k^e} \right\| &\approx \epsilon(1-\lambda)\mathcal{P} + O(\epsilon^2). \end{aligned} \tag{67}$$

Thus

$$\|\hat{P}_k - P_k\| \le \epsilon((1-\lambda)\mathcal{P}^2\Phi^2 + (1-\lambda)\mathcal{P} + \mathcal{P}) \tag{68}$$

and a very simple bound is

$$\|\hat{P}_k - P_k\| \le \epsilon\mathcal{E} \equiv \epsilon\mathcal{P}. \tag{69}$$

Note that in a finite-precision implementation of the algorithm, P_k should be replaced by its finite-precision version $\tilde{P}_k = P_k + \Delta P_k$. In the sequel, we show that $\|\Delta P_k\| < \|P_k\|$; recall that in (44), we obtained

$$\|\Delta P_k\| < r_1 < \frac{1-\lambda}{\Phi^2} = \frac{1}{\overline{\mathcal{R}}} \le \frac{1}{\|R_k\|} \le \|P_k\|.$$

Thus, by replacing P_k with $\tilde{P}_k$ in the previous calculations, we obtain

$$\|\hat{P}_k - \tilde{P}_k\| \le \epsilon\mathcal{E} \equiv 2\epsilon\mathcal{P} \tag{70}$$

or by ignoring the term 2

$$\|\hat{P}_k - \tilde{P}_k\| \le \epsilon\mathcal{E} \equiv \epsilon\mathcal{P}. \tag{71}$$

APPENDIX B

From (12), we obtain that the higher order terms $f(k, \Delta P_{k-1})$ are given by

$$\begin{aligned} f(k,\Delta P_{k-1}) = &\underbrace{-\frac{1}{\lambda}\frac{P_{k-1}\phi_k\phi_k^t P_{k-1}}{r_k^e} t_2(k,\Delta P_{k-1})}_{T_1(k)} \\ &\underbrace{-\frac{1}{\lambda}\frac{P_{k-1}\phi_k\phi_k^t \Delta P_{k-1}}{r_k^e} t_1(k,\Delta P_{k-1})}_{T_2(k)} \\ &\underbrace{-\frac{1}{\lambda}\frac{\Delta P_{k-1}\phi_k\phi_k^t P_{k-1}}{r_k^e} t_1(k,\Delta P_{k-1})}_{T_3(k)} \\ &\underbrace{-\frac{1}{\lambda}\frac{\Delta P_{k-1}\phi_k\phi_k^t \Delta P_{k-1}}{r_k^e} t_0(k,\Delta P_{k-1})}_{T_4(k)}. \end{aligned} \tag{72}$$

The following relation, which can be easily proved using (4) and (6), will be used in the sequel:

$$\frac{P_{k-1}\phi_k}{r_k^e} = P_k\phi_k. \tag{73}$$

Proof of Theorem 1: From (24), we obtain

$$\|\Delta P_k\| \le \frac{1}{1-\lambda} \max_i \|P_k R_i (f(i,\Delta P_{i-1}) + \epsilon\tilde{P}_i) R_i P_k\|. \tag{74}$$

We will bound $\|P_k R_i (f(i,\Delta P_{i-1}) + \epsilon\tilde{P}_i) R_i P_k\|$ by the sum of the bounds of the norms of each subterm.

In order to compute a bound for $\|P_k R_i T_1(i) R_i P_k\|$, we express the argument of the norm as

$$\frac{1}{\lambda} P_k R_i \frac{P_{i-1}\phi_i\phi_i^t P_{i-1}}{r_i^e} R_i P_k \sum_{n=2}^{\infty} (-1)^n \frac{\left(\phi_i^t \Delta P_{i-1}\phi_i\right)^n}{\left(r_i^e\right)^n} \tag{75}$$

which, using (73), becomes

$$\frac{1}{\lambda} P_k\phi_i\phi_i^t P_k \sum_{n=2}^{\infty} (-1)^n \frac{\left(\phi_i^t \Delta P_{i-1}\phi_i\right)^n}{\left(r_i^e\right)^{n-1}}. \tag{76}$$

Since P_{k-1} is positive definite, (4) gives that $r_k^e > \lambda$. Thus, $\frac{1}{r_k^e} < \frac{1}{\lambda}$ and using (14), (15), $r < \frac{\lambda}{\Phi^2}$, and $\|\Delta P_i\| < r$ for $i = 1, \ldots, k-1$, (76) gives

$$\|P_k R_i T_1(i) R_i P_k\| < \mathcal{P}^2\Phi^2 \sum_{n=2}^{\infty} \left(\frac{\Phi^2 r}{\lambda}\right)^n = \frac{\mathcal{P}^2\Phi^6 r^2}{\lambda(\lambda - \Phi^2 r)}. \tag{77}$$

From (72) and (73), we get that the second term is

$$-\frac{1}{\lambda} P_k R_i P_i \phi_i \phi_i^t \Delta P_{i-1} R_i P_k t_1(i, \Delta P_{i-1}) \tag{78}$$

and thus

$$\begin{aligned} \|P_k R_i T_2(i) R_i P_k\| &< \frac{1}{\lambda}\mathcal{P}\Phi^2 r\mathcal{K}\|t_1(i,\Delta P_{i-1})\| \\ &< \frac{\mathcal{P}\Phi^4\mathcal{K}r^2}{\lambda(\lambda - \Phi^2 r)}. \end{aligned} \tag{79}$$

The same bound holds for the third term. In the same manner, we can show that

$$\|P_k R_i T_4(i) R_i P_k\| < \frac{\Phi^2\mathcal{K}^2 r^2}{\lambda(\lambda - \Phi^2 r)}. \tag{80}$$

For the term concerning the roundoff, we get

$$\|P_k R_i \epsilon\tilde{P}_i R_i P_k\| \le \mathcal{K}^2\mathcal{E}\epsilon. \tag{81}$$

Combining (77), (79), (80), (81), and (74), we verify (25) to prove Theorem 1. □

APPENDIX C

Proof of Theorem 2: In view of (35), in order to prove Theorem 2, we must compute a bound for $\|f(i,\Delta P_{i-1})\|$. We will bound $\|f(i,\Delta P_{i-1})\|$ by the sum of the bounds of the subterms. From (72) and (73), we obtain

$$T_1(i) = -\frac{1}{\lambda} P_i\phi_i\phi_i^t P_i \sum_{n=2}^{\infty} \frac{\left(\phi_i^t \Delta P_{i-1}\phi_i\right)^n}{\left(r_i^e\right)^{n-1}}. \tag{82}$$

From (3) and (33), we obtain

$$P_i\phi_i\phi_i^t = I - \lambda P_i R_{i-1} \approx (1-\lambda) I. \tag{83}$$

Thus

$$||T_1(i)|| < \frac{(1-\lambda)\mathcal{P}\Phi^4 r^2}{\lambda(\lambda-\Phi^2 r)}. \quad (84)$$

From (72) and (73), we get

$$T_2(i) = -\frac{1}{\lambda} P_i \phi_i \phi_i^t \Delta P_{i-1} t_1(i, \Delta P_{i-1}). \quad (85)$$

Thus

$$||T_2(i)|| < \frac{1}{\lambda}(1-\lambda) r ||t_1(i, \Delta P_{i-1})|| < \frac{(1-\lambda)\Phi^2 r^2}{\lambda(\lambda-\Phi^2 r)}. \quad (86)$$

The same bound holds for the third term. In the same manner, we can show

$$||T_4(i)|| < \frac{\Phi^2 r^2}{\lambda(\lambda-\Phi^2 r)}. \quad (87)$$

Combining (35), (84), (86), and (87), we obtain (36), which proves Theorem 2. □

ACKNOWLEDGMENT

The first author would like to thank Prof. G. V. Moustakides for useful comments on an early draft of this paper.

REFERENCES

[1] S. Haykin, *Adaptive Filter Theory*, 3rd ed. Englewood Cliffs, NJ: Prentice-Hall, 1996.
[2] N. Kalouptsidis and S. Theodoridis, *Adaptive System Identification and Signal Processing Algorithms*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[3] S. Ljung and L. Ljung, "Error propagation properties of recursive least-squares adaptation algorithms," *Automatica*, vol. 21, no. 2, pp. 157–167, 1985.
[4] M. H. Verhaegen, "Round-off error propagation in four generally-applicable, recursive, least-squares estimation schemes," *Automatica*, vol. 25, no. 3, pp. 437–444, 1989.
[5] D. T. M. Slock, "Backward consistency concept and round-off error propagation dynamics in resursive least-squares algorithms," *Opt. Eng.*, vol. 31, no. 6, pp. 1153–1169, June 1992.
[6] B. D. O. Anderson *et al.*, *Stability of Adaptive Systems: Passivity and Averaging Analysis*. Cambridge, MA: MIT Press, 1986.
[7] G. Bottomley and S. T. Alexander, "A novel approach for stabilizing recursive least squares filters," *IEEE Trans. Signal Processing*, vol. 39, pp. 1770–1779, Aug. 1991.
[8] A. P. Liavas and P. A. Regalia, "Numerical stability issues of the conventional recursive least squares algorithm," in *Proc. ICASSP*, Seattle, WA, May 1998.
[9] D. T. M. Slock and T. Kailath, "Numerically stable fast transversal filters for recursive least squares adaptive filtering," *IEEE Trans. Signal Processing*, vol. 39, pp. 92–114, Jan. 1991.
[10] E. Eleftheriou and D. Falconer, "Tracking properties and steady-state performance of RLS adaptive filter algorithms," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 1097–1109, Oct. 1986.
[11] J. M. Cioffi, "Limited precision effects in adaptive filtering," *IEEE Trans. Circuits Syst.*, vol. CAS-34, pp. 821–833, July 1987.
[12] G. Stewart and J. Sun, *Matrix Perturbation Theory*. New York: Academic, 1990.
[13] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Baltimore, MD: Johns Hopkins Univ. Press, 1989.
[14] N. Higham, *Accuracy and Stability of Numerical Algorithms*, Philadelphia, PA: SIAM, 1996.

Athanasios P. Liavas was born in Pyrgos, Greece, in 1966. He received the diploma and the Ph.D. degrees in computer engineering from the University of Patras, Patras, Greece, in 1989 and 1993, respectively.

He is currently a Research Fellow with the Département Signal et Image, Institut National des Télécommunications, Evry, France, under the framework of the Training and Mobility of Researchers (TMR) program of the European Commission. His research interests include adaptive signal processing algorithms, blind system identification, and biomedical signal processing.

Dr. Liavas is a member of the Technical Chamber of Greece.

Phillip A. Regalia (SM'96) was born in Walnut Creek, CA, in 1962. He received the B.Sc. (highest honors), M.Sc., and Ph.D. degrees in electrical and computer engineering from the University of California, Santa Barbara, in 1985, 1987, and 1988, respectively, and the Habilitation à Diriger des Recherches degree from the University of Paris, Orsay, France, in 1994.

He is presently a Professor with the Institut National des Télécommunications, Evry, France, with research interests focused in adaptive signal processing.

Dr. Regalia serves as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING, and as an editor for the *International Journal of Adaptive Control and Signal Processing*.