Proceedings of the 28th Conference
on Decision and Control
Tampa, Florida • December 1989

Non-Asymptotic Results for Finite-Memory WLS Filters

Maciej Niedźwiecki* and Lei Guo**

* Institute of Computer Science, Technical University of Gdańsk
ul. Majakowskiego 11/12, Gdańsk, Poland

** Institute of Systems Science, Academia Sinica
Beijing 100080, People's Republic of China

Abstract

The paper presents, what we believe to be, the first non-asymptotic analysis of properties of weighted least squares (WLS) adaptive filters used for identification of time-varying systems. We show that the problem of mean-square boundedness of WLS estimates is closely related to the problem of invertibility - in the mean sense - of the corresponding regression matrix. We discuss conditions under which such invertibility is guaranteed. Based on that, a number of results are derived paralleling those already obtained for least mean square (LMS) filters and the problem of "statistical robustness" of the WLS estimator is discussed.

1 Introduction

Consider the following time-varying stochastic system

$$y(t) = \alpha^T(t)\phi(t) + n(t) \tag{1}$$

where $\phi(t) = [u_1(t), \ldots, u_r(t)]^T$ is the measurable input vector, $\alpha(t) = [a_1(t), \ldots, a_r(t)]^T$ is the unknown (time dependent) parameter vector and $\{n(t)\}$ denotes the unobservable (scalar) measurement noise. We will assume that

(A1) the noise process $\{n(t)\}$ is a sequence of zero-mean independent and identically distributed random variables and $E[n^2(t)] = \rho_0$.

(A2) the input process $\{\phi(t)\}$, independent of $\{n(t)\}$, is a sequence of indentically distributed m-dependent random vectors (i.e. $\exists m$ such that $\forall t$ sequences $\{\phi(i), i \le t\}$ and $\{\phi(i), i \ge t+m\}$ are independent) and $E[\phi(t)\phi^T(t)] = R_0 > 0$.

(A3) time-varying parameters form a sequence $\{\alpha(t)\}$, independent of $\{\phi(t)\}$ and $\{n(t)\}$, which is bounded in the mean square sense, i.e.

$$E[\|\alpha(t)\|^2] \le A < \infty \qquad \forall t$$

Assumption about m-dependence of the input sequence is not critical and will be relaxed to include weaker mixing (asymptotic independence) and covariance conditions later on.
Note that in the case where

$$u_i(t) = u(t-i), \qquad i = 1, \ldots, r \tag{2}$$

(1) specializes to the dynamic finite impulse response (FIR) model widely used in adaptive filtering, e.g. for the purpose of adaptive equalization of communication channels.
If parameters in (1) vary sufficiently slowly with time the method of weighted least squares (WLS) can be used for the purpose of tracking of $\alpha(t)$. Let $\{w(t)\}$ denote the nonnegative and nonincreasing weighting sequence, such that

$$\sum_{t=0}^{\infty} w(t) = 1 \tag{3}$$

(the normalization constraint (3) is not essential for our analysis and was introduced for the sake of notational convenience). Assuming, for convenience, that the infinite observation history is available at the instant t, the WLS estimator can be defined in the following way

$$\hat{\alpha}(t) = \arg\min_{\alpha} \sum_{i=0}^{\infty} w(i)[y(t-i) - \alpha^T\phi(t-i)]^2$$

$$= \Big(\sum_{i=0}^{\infty} w(i)\phi(t-i)\phi^T(t-i)\Big)^{-1}\Big(\sum_{i=0}^{\infty} w(i)y(t-i)\phi(t-i)\Big) = \tilde{R}^{-1}(t)\tilde{S}(t) \tag{4}$$

with obvious definitions of $\tilde{R}(t)$ and $\tilde{S}(t)$.
In practice the requirement that the WLS estimator should be recursively computable limits our choice of $w(t)$ to several standard windows. If, for example, the exponential window is used $(w(t) = (1-\lambda)\lambda^t,\ 0 < \lambda < 1)$ one can replace (4) by the following recursive algorithm [1]

$$\begin{aligned} \hat{\alpha}(t) &= \hat{\alpha}(t-1) + D(t)\phi(t)\epsilon(t) \\ \epsilon(t) &= y(t) - \hat{\alpha}^T(t-1)\phi(t) \end{aligned} \tag{5}$$

where the matrix $D(t)$ can be updated using the well-known formula

$$D(t) = \frac{1}{\lambda}\Big[D(t-1) - \frac{D(t-1)\phi(t)\phi^T(t)D(t-1)}{\lambda + \phi^T(t)D(t-1)\phi(t)}\Big] \tag{6}$$

Similar (but basically two-step) algorithm can be derived for the sliding rectangular window $(w(t) = 1/N$ for $t < N$ and $= 0$ for $t \ge N)$ - see e.g. [1]. Various fast versions of the WLS algorithm are also available but they usually require "safety jacketing" because of possible numerical ill-conditioning - see Cioffi [2].
If the data-dependent adaptation matrix $D(t)$ in (5) is replaced by a small adaptation gain μ one arrives at the so-called least mean square (LMS) algorithm

$$\hat{\alpha}(t) = \hat{\alpha}\ (t-1) + \mu\phi(t)\epsilon(t) \tag{7}$$

Although computationally less demanding than the WLS algorithm, the LMS algorithm may suffer from a very slow initial convergence - a disadvantageous effect if rapid adaptation is required. Despite this difference both algorithms have very similar parameter tracking properties - see e.g. Eleftheriou and Falconer [3].
While the statistical properties of LMS filters seem to be well-explored and documented the situation is less clear for the WLS filters. So far all the analyses were based on asymptotic arguments, i.e. strictly speaking, they dealt with the case where the effective length of the window tended to infinity. Almost no precise results seem to exist for strictly finite-length windows - the recent paper of Macchi and Eweda [19] being the only noticeable exception[1]. However, even the results presented in [19] rely critically on the assumption about inveribility of a certain stochastic regression matrix (assumption (A2) in [19]) which is postulated but is very difficult to verify.
In this paper we present, what we believe to be, the first non-asymptotic analysis of properties of the WLS estimator based on realistic and verifiable assumptions. We show that the problem of the mean-square boundedness of $\hat{\alpha}(t)$ is closely related to the problem of invertibility - in the mean sense - of the regression matrix $\tilde{R}(t)$ in (4). We discuss conditions under which such invertibility is guaranteed if sufficiently strong mixing (asymptotic independence) and/or covariance conditions are imposed on $\{\phi(t)\}$. Based on that, a number of results can be derived paralleling those obtained for LMS estimators by Macchi and Eweda [4],[5] and the problem of "statistical robustness" of the WLS estimator can be properly addressed.

2 Review of Known Results

From among a large number of results on stability and tracking bounds for the LMS algorithm we would like to point to a sequence of insightful papers by Macchi and Eweda [4]-[7].
Assuming that the input sequence is stationary and m-dependent the authors were able to show that, for sufficiently small but non-zero gain μ

$$E[\|\hat{\alpha}(t) - \alpha(t)\|^2] \le C(\mu) < \infty \tag{8}$$

In the constant parameter case $(\alpha(t) = \alpha_0)$ we have [6]

$$C(\mu) = C_1\mu \tag{9}$$

that is, the bound on random fluctuations in steady state decreases with the stepsize in a linear way.
If system parameters vary with time the choice of μ becomes a trade-off between the steady-state accuracy and tracking ability of the estimation algo-

[1] This paper was brought to our attention during the revision process

CH2642-7/89/0000-1785$1.00 © 1989 IEEE

rithm. Assuming, for example, that $\{\alpha(t)\}$ evolves according to the random walk model it is possible to show that [6]

$$C(\mu) = C_1\mu + \frac{C_2}{\mu} \tag{10}$$

which illustrates the need for compromise mentioned above.
Denote by ι the equivalent width of the window $\{w(t)\}$ (equivalent number of observations)

$$\iota = 1/\sum_{t=0}^{\infty} w^2(t) \tag{11}$$

deciding upon the "memory" of the WLS filter [8]. One can argue that the quantity $1/\iota$ determines the adaptation gain of the WLS algorithm, i.e. it plays exactly the same role as the stepsize μ in the LMS filter.
Basically, two different approaches were used to analyse properties of WLS estimators:

- the approach based on Taylor series approximations (for any weighting sequence), see e.g. Niedźwiecki [8], [9]. [10], [11]
- the approach based on ODE approximations (for exponential weighting), see e.g. Benveniste [12], [13], Kushner and Huang [14] and Ljung [15].

In both cases the derived results hold only asymptotically, that is for $\iota \to \infty$. Almost no results seem to exist if ι is finite and fixed. For example, for the constant parameter case only a considerably weaker version of (8)-(9) is available. According to Eweda and Macchi [7] for arbitrarily small $\varepsilon > 0$ the estimation error $\|\hat{\alpha}(t) - \alpha_0\|^2$ has an upper bound proportional to $1/\iota$ *with probability* $1 - \varepsilon$.
The finite mean square tracking bound established subsequently in [19] rests on an *implicit* assumption that there exist an integer N_0 and a constant $0 < c < \infty$ such that $\forall\ N \geq N_0, \forall\ t$

$$E\left\{\left[\lambda_{min}\left(\sum_{i=0}^{N-1}\phi(t-i)\phi^T(t-i)\right)\right]^{-8}\right\} < c \tag{12}$$

It turns out, however, that verification of assumptions similar to (12) is far from being obvious and constitutes the very core of the tracking assessment problem. We will consider this issue in more detail in section 4.

3 "Idealized" WLS Estimator

Since under (A2) we have

$$\tilde{R}(t) = \sum_{i=0}^{\infty} w(i)\phi(t-i)\phi^T(t-i) \underset{\iota\to\infty}{\longmapsto} R_0 \tag{13}$$

where convergence takes place either in the mean square sense or with probability one [9], for sufficiently large ι one can attempt to replace the regression matrix $\tilde{R}(t)$ in (4) by its expectation. The resulting "idealized" WLS estimator

$$\hat{\bar{\alpha}}(t) = R_0^{-1}\tilde{S}(t) \tag{14}$$

is analytically easy to handle. Moreover, provided that the difference

$$E[\|\hat{\alpha}(t) - \hat{\bar{\alpha}}(t)\|^2] \tag{15}$$

is sufficiently small one can infer about properties of the WLS estimator $\hat{\alpha}(t)$ by analysing properties of its "idealized" counterpart - that was the line of thinking in [9], [10]. One of the points behind studying properties of (15) is that, via the inequality

$$E[\|\hat{\alpha}(t) - \alpha(t)\|^2] \leq 2E[\|\hat{\alpha}(t) - \hat{\bar{\alpha}}(t)\|^2] + 2E[\|\hat{\bar{\alpha}}(t) - \alpha(t)\|^2] \tag{16}$$

boundedness of (15) implies boundedness of the mean square tracking error (under (A1) - (A3) boundedness of the second term on the right hand side of (16) can be shown quite easily).
Not surprisingly, the problem of boundedness of (15) can be related to the problem of invertibility, in the mean sense, of the matrix $\tilde{R}(t)$. In particular, we have the following:

Lemma 1. *Under assumptions (A1) - (A3) we have*

$$E[\|\hat{\alpha}(t) - \hat{\bar{\alpha}}(t)\|^2] = 0(\mathrm{tr}\{E[\Delta(t)]\}) \tag{17}$$

where $\Delta(t) = \tilde{R}^{-1}(t) - R_0^{-1}$.

Proof

$$\Delta\alpha(t) = \hat{\alpha}(t) - \hat{\bar{\alpha}}(t) = \Delta(t)\tilde{S}(t) = \sum_{i=0}^{\infty} x(i)v(i)$$

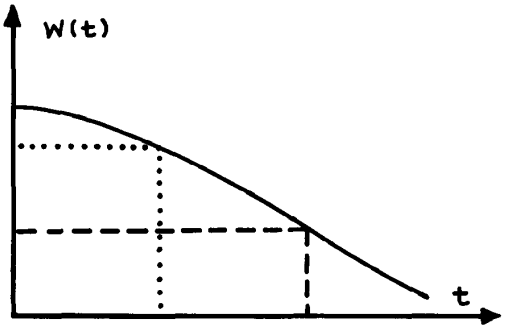

Fig.1 Rectangular windows "inscribed" in $\{w(i)\}$

where

$$x(i) = \sqrt{w(i)}\Delta(t)\phi(t-i), \quad v(i) = \sqrt{w(i)}\left[\phi^T(t-i)\alpha(t-i) + n(t-i)\right]$$

Using the Schwartz inequalty one gets

$$E[\|\Delta\alpha(t)\|^2] \leq E\Big[\sum_{i=0}^{\infty}\|x(i)\|^2\Big]E\Big[\sum_{i=0}^{\infty}v^2(i)\Big]$$

Observe that

$$E\Big[\sum_{i=0}^{\infty}\|x(i)\|^2\Big] = E\Big[\ \mathrm{tr}\ \Big\{\sum_{i=0}^{\infty}x(i)x^T(i)\Big\}\Big]$$
$$= \mathrm{tr}\ \{E[\Delta(t)\tilde{R}(t)\Delta(t)]\} = \mathrm{tr}\ \{E[\Delta(t)]\}$$

Similarly

$$E\Big[\sum_{i=0}^{\infty}v^2(i)\Big] \leq \sum_{i=0}^{\infty}E[\|\alpha(t-i)\|^2]E[\mathrm{tr}\{w(i)\phi(t-i)\phi^T(t-i)\}] + \rho_0$$
$$\leq A\mathrm{tr}\{R_0\} + \rho_0 = 0(1)$$

4 Invertibility of the Regression Matrix

4.1 Preliminary considerations

According to Lemma 1, proving boundedness of the mean square tracking error amounts to finding conditions under which

$$E[\tilde{R}^{-1}(t)] < \infty$$

i.e. under which the stochastic regression matrix $\tilde{R}(t)$ is invertible in the mean sense. Let $\{w'(t)\}$:

$$w'(t) = \begin{cases} c & \text{for} \quad t < N \\ 0 & \text{for} \quad t \geq N. \end{cases} \tag{18}$$

denote a rectangular window "inscribed" in $\{w(t)\}$ - see Fig. 1. Since $\tilde{R}(t) \geq cR(t)$ where

$$R(t) = \sum_{i=0}^{N-1}\phi(t-i)\phi^T(t-i)$$

the overbounding technique can be used, i.e. the boundedness results for general windows are implied by the corresponding results for the "inscribed" rectangular wndows. However, even under uniform weighting the problem of invertibility of the stochastic regression matrix is conspicuously absent from the statistical literature. Observe that

$$\left(R^{-1}(t)\right)_{ij} = \frac{\left(\mathrm{adj}\ R(t)\right)_{ij}}{\det R(t)}$$

and hence, using Hölder inequality, one gets ($k > 1$)

$$E[(R^{-1}(t))_{ij}] \leq \left[E[(\det R(t))^{-k}]\right]^{\frac{1}{k}}\left[E\left[(\mathrm{adj}R(t))_{ij}^{\frac{k}{k-1}}\right]\right]^{\frac{k-1}{k}} \tag{19}$$

Imposing some moment conditions on $u_i(t)$, $i = 1, \ldots, r$ namely

(A4) $\exists\varepsilon > 0$ such that

$$E[(u_i(t))^{2(r-1)+\varepsilon}] < \infty, \quad \forall t, i = 1, \ldots, r$$

and choosing k sufficiently large so that $2(r-1)/(k-1) \leq \varepsilon$ one can guarantee boundedness of the second factor on the right hand side of (19). Therefore to prove invertibility of $R(t)$ in the mean it suffices to find conditions under which the determinant of $R(t)$ is invertible in kth-moment

$$E\big[(\det R(t))^{-k}\big] < \infty \tag{20}$$

for sufficiently large k.

4.2 Need for additional constraints

Quite obviously, in order to satisfy (20) one needs

$$P\big(\det R(t) = 0\big) = 0 \tag{21}$$

which can not be guaranteed by imposing only moment conditions on $\{\phi(t)\}$, such as

$$E[\phi(t)\phi^T(t)] > 0 \qquad \forall t \tag{22}$$

We will clarify this point by means of the following example:

Example
Consider the case where (2) holds and $\{u(t)\}$ is an i.i.d. sequence taking only two values: $+1$ and -1 with probabilities p $(0 < p < 1)$ and $1-p$, respectively. Then it is straightforward to show that for any finite N (21) is not fulfilled even though (22) is. ■

Quite clearly, additional constraint is needed to rule out cases such as the one above. The following result, which can be thought of as a refinement of (21), will be very useful for our further purposes.

Lemma 2. *The determinant of $R(t)$ is invertible in k-th moment provided that $\exists x_0$, $\eta > 0$, $L \geq k+1$ such that $\forall x : x_0 > x > 0$ and $\forall t$ it holds*

$$P\big(\det R(t) \leq x\big) \leq \eta x^L \tag{23}$$

Proof
Using (23) one gets

$$E\big[(\det R(t))^{-k}\big] = \int_0^\infty x^{-k} dP\big(\det R(t) \leq x\big)$$

$$= k\int_0^\infty x^{-k-1} P\big(\det R(t) \leq x\big)dx \leq k\int_0^{x_0} \eta x^{L-k-1}dx + x_0^{-k} < \infty$$

4.3 Explicit invertibility condition

We will show that the implicit condition (23) can be met if N in (18) is sufficiently large and

(A5) $\exists \gamma$, δ, $x_0 > 0$ such that $\forall x : x_0 > x > 0$, and $\forall t$

$$\sup_{\|\beta\|=1} P\big((\beta^T\phi(t))^2 < x\big) \leq \gamma x^\delta \tag{24}$$

In order to simplify the analysis we will derive the main result under the assumption that $\{\phi(t)\}$ is an i.i.d. sequence and then we will extend it to more general cases.

Lemma 3 (key technical lemma) . *Suppose that $\{\phi(t)\}$ is an i.i.d. sequence obeying assumption (A5). Then the condition (23) of Lemma 2 is fulfilled and the number L in (23) can be made arbitrarily large by increasing N.*

Proof
For convenience take $N = rM$. We have

$$R(t) = \sum_{j=0}^{M-1} H_j(t), \quad H_j(t) = \sum_{k=0}^{r-1} \phi(t-jr-k)\phi^T(t-jr-k)$$

since for nonnegative definite matrices A and B we have $\det(A+B) \geq \det A + \det B$ it holds

$$\det R(t) \geq \sum_{j=0}^{M-1} \det H_j(t)$$

and consequently

$$P\big(\det R(t) \leq x\big) \leq P\Big(\sum_{j=0}^{M-1} \det H_j(t) \leq x\Big) \leq \big[P(\det H_0(t) \leq x)\big]^M$$

- since the matrices $H_j(t)$ are mutually independent. Observe that $\det H_0(t) = \big(\det W(t)\big)^2$ where

$$W(t) = \begin{bmatrix} u_1(t) & \cdots & u_1(t-r+1) \\ \vdots & & \vdots \\ u_r(t) & \cdots & u_r(t-r+1) \end{bmatrix}$$

and

$$\det W(t) = \sum_{i=1}^{r} (-1)^{i-1} u_i(t) z_i(t) \triangleq \phi^T(t) z(t)$$

where $z(t)$ is a collection of corresponding minors of the matrix $W(t)$. Note that

$$z_i^2(t) = \det\Big(\sum_{j=1}^{r-1} \phi_{[i]}(t-j)\phi_{[i]}^T(t-j)\Big)$$

$$\phi_{[i]}(t) = [u_1(t), \ldots, u_{i-1}(t), u_{i+1}(t), \ldots u_r(t)]^T$$

Rewrite $\det H_0(t)$ in the form $\det H_0(t) = \|z(t)\|^2\big(\beta^T(t)\phi(t)\big)^2$ where $\beta(t) = z(t)/\|z(t)\|$ for all $z(t) \neq 0$. Using simple calculations based on conditional probabilities one gets, for any $x < x_0$:

$$P(\det H_0(t) \leq x) = P\big(\det H_0(t) \leq x \,|\, \|z(t)\|^2 \geq \sqrt{x}\big) P(\|z(t)\|^2 \geq \sqrt{x})$$

$$+ P\big(\det H_0(t) \leq x \,|\, \|z(t)\|^2 < \sqrt{x}\big) P\big(\|z(t)\|^2 < \sqrt{x}\big)$$

$$\leq P\Big(\big(\beta^T(t)\phi(t)\big)^2 \leq \sqrt{x} \,|\, \|z(t)\|^2 \geq \sqrt{x}\Big) + P\big(\|z(t)\|^2 < \sqrt{x}\big) \tag{25}$$

where it was assumed that for any positive x : $P(\|z(t)\|^2 < \sqrt{x}) > 0$ so that the corresponding conditional probability is well-defined (if not, a simple modification can be introduced).
Owing to the fact that the regression vector $\phi(t)$ is independent of $z(t) = f\big(\phi(t-1), \ldots, \phi(t-r+1)\big)$ and that $\|\beta(t)\| = 1$ we get (c.f. (A5)):

$$P\Big(\big(\beta^T(t)\phi(t)\big)^2 \leq \sqrt{x} \,|\, \|z(t)\|^2 \geq \sqrt{x}\Big) \leq \sup_{\|\beta\|=1} P\Big(\big(\beta^T\phi(t)\big)^2 \leq \sqrt{x}\Big) \leq \gamma x^{\delta/2} \tag{26}$$

Consider, in turn, the second term on the right hand side of (25)

$$P\big(\|z(t)\|^2 < \sqrt{x}\big) = P\Big(\sum_{i=1}^{r} z_i^2(t) < \sqrt{x}\Big) \leq \sum_{i=1}^{r} P\big(z_i^2(t) < \sqrt{x}\big) \tag{27}$$

We shall prove, by induction, that under conditions implied by (A5) the following proposition is true.

Proposition. *$\exists \xi_r$, $\tau_r > 0$ such that $\forall x : x_0 > x > 0$ and $\forall t$*

$$P\big(\det H_0(t) \leq x\big) = P\Big(\det\Big[\sum_{j=0}^{r-1} \phi(t-j)\phi^T(t-j)\Big] \leq x\Big) \leq \xi_r x^{\tau_r} \tag{28}$$

Actually, suppose that the proposition is true for $r-1$, i.e. for all $(r-1)$-dimensional subvectors of $\phi(t)$ there exist constants ξ_{r-1}, τ_{r-1} such that

$$P\big(z_i^2(t) \leq x\big) = P\Big(\det\Big[\sum_{j=1}^{r-1} \phi_{[i]}(t-j)\phi_{[i]}^T(t-j)\Big] \leq x\Big)$$

$$\leq \xi_{r-1} x^{\tau_{r-1}}, \qquad i = 1, \ldots, r$$

Now, combining this with (27) one gets $P(\|z(t)\|^2 < \sqrt{x}) \leq r\xi_{r-1} x^{\frac{\tau_{r-1}}{2}}$ and consequently, using (25), (26) and the bound obtained above one has

$$P\big(\det H_0(t) \leq x\big) \leq \gamma x^{\frac{\delta}{2}} + r\xi_{r-1} x^{\frac{\tau_{r-1}}{2}} \leq \xi_r x^{\tau_r}$$

for $x \in (0, 1]$, $\tau_r = \min\Big(\frac{\delta}{2}, \frac{\tau_{r-1}}{2}\Big)$ and appropriately chosen ξ_r.
Since proposition stems immediately from (A5) in the case where $r = 1$, it is also true in the general case. Finally, observe that

$$P\big(\det R(t) \leq x\big) \leq [P(\det H_0(t) \leq x)]^M \leq (\xi_r)^M x^{M\tau_r} = \eta x^L$$

where $L = M\tau_r$ can be made arbitrarily large by increasing M (i.e. N). ■

Remark 1
It is known that distribution of any r-dimensional random vector ϕ can be factored as (Lebesgue decomposition theorem)

$$F(\phi) = \mu_c F_c(\phi) + \mu_d F_d(\phi) + \mu_s F_s(\phi)$$

where F_c, F_d, F_s are continuous discrete and singular distributions, respectively and μ_c, μ_d, μ_s are nonnegative constants such that $\mu_c + \mu_d + \mu_s = 1$. What (A5) effectively says is that $F(\phi)$ should be free of discrete and singular (supported on hyperplanes) components. Additionally it rules out "almost discrete" and "almost singular" components in the continuous distribution. We note that (A5) admits a very large class of continuous distributions, e.g. all ones characterized by bounded probability density functions (such as Gaussian, uniform etc.).

Remark 2
Note that the dependence structure of $\{\phi(t)\}$ was not used when we derived the moment condition (A4). If the input sequence is m-dependent existence of second-order moments (implied by (A2)) is sufficient to prove boundedness of the second term on the right hand side of (20).

4.4 Extension to weaker mixing and covariance conditions

The requirement that the sequence of regression vectors $\{\phi(t)\}$ should be white (as stated in conditions of Lemma 3) is, quite clearly, very inconvenient. Note, for example, that it is never met for FIR models (2), even if the input sequence is white ! (since successive regression vectors share $r-1$ components).
Basically, the results of Lemma 3 can be extended in two different directions - to weaker mixing (asymptotic independence) conditions and weaker covariance (rate of decorrelation) conditions.
As far as mixing is concerned relaxation of i.i.d. assumption to $m-$dependence (consistent e.g. with (2) under white noise excitation) is straightforward. Suppose that $\{\phi(t)\}$ is an identically distributed and m-dependent sequence obeying (A5). Then $\{\phi(tm)\}$ is an i.i.d. sequence and hence, for suitably large N

$$E\Big[R^{-1}(t)\Big] \le E\Big[(\sum_i \phi(t-im)\phi^T(t-im))^{-1}\Big] < \infty$$

Extension to weaker mixing conditions is also possible. In particular, denote by $\mathcal{F}_i^t$ the sigma-algebra generated by the $\{\phi(i);\ t \le i \le s\}$. If the following mixing (asymptotic independence) condition is fulfilled:

$$|P(AB) - P(A)P(B)| \le \psi(n)P(A)P(B)$$

for any events $A \in \mathcal{F}_{-\infty}^t$ and $B \in \mathcal{F}_s^\infty$, where $n = s-t$ and $\psi(n) \to 0$ for $n \to \infty$ (the sequence $\{\phi(t)\}$ is called super-uniformly mixing or ψ-mixing). Then all previous results can be easily extended to such sequences.

From the practical point of view much more interesting results can be obtained by means of relaxing covariance conditions imposed on $\{\phi(t)\}$. Actually, consider the case where $\phi(t)$ is the output of the state space model

$$\begin{aligned} x(t+1) &= Ax(t) + B\eta(t) \\ \phi(t) &= Cx(t) + D\eta(t) \end{aligned} \tag{29}$$

Then we have the following result

Theorem 1. *Lemma 3 holds if the model (29) is output reachable and $\{\eta(t)\}$ is an i.i.d. sequence obeying (A5).*

Outline of proof
The proof is based on the following basic inequality valid for output reachable state space models (see e.g. [17], [18])

$$\lambda_{min}\Big[\sum_{i=0}^{N-1}\phi(t-i)\phi^T(t-i)\Big] \ge c\lambda_{min}\Big[\sum_{i=0}^{N+\nu-1}\bar{\eta}(t-i)\bar{\eta}^T(t-i)\Big], \quad \forall t \tag{30}$$

where $\bar{\eta}(t) = [\eta^T(t), \dots, \eta^T(t-\nu)]^T$, ν is the McMillan degree of the system (29) and $c > 0$.
The following example will illustrate the main steps in the proof of Theorem 1

Example
Let $\{u(t)\}$ be generated from the following $AR(p)$ model:

$$u(t) + a_1u(t-1) + \dots + a_pu(t-p) = \epsilon(t)$$

where $\{\epsilon(t)\}$ is an i.i.d. sequence satisfying $P(|\epsilon(t)| \le x) \le \gamma x^\delta$, for some $\gamma > 0$, $\delta > 0$. Then Lemma 3 also holds with $\phi(t) = [u(t-1), \dots, u(t-r)]^T$.

Proof
Let us denote

$$A(q^{-1}) = a_0 + a_1q^{-1} + \dots + a_pq^{-p}$$

where $a_0 = 1$ and q^{-1} is the backwards shift operator, and define

$$\psi(t) = A(q^{-1})\phi(t), \quad R_1(t) = \sum_{i=0}^{N-p-1}\psi(t-i)\psi^T(t-i)$$

Then by the Schwarz inequality it is seen that for any vector $\alpha \in R^r$,

$$\alpha^T R_1(t)\alpha = \sum_{i=0}^{N-p-1}[\alpha^T\psi(t-i)]^2 = \sum_{i=0}^{N-p-1}\Big[\sum_{j=0}^{p}a_j\alpha^T\phi(t-i-j)\Big]^2$$

$$\le \sum_{j=0}^{p}a_j^2\sum_{j=0}^{p}\sum_{i=0}^{N-p-1}[\alpha^T\phi(t-i-j)]^2 \le (p+1)\sum_{j=0}^{p}a_j^2\alpha^T R(t)\alpha$$

Consequently, by the arbitrariness of α,

$$\lambda_{\min}[R(t)] \ge \frac{\lambda_{\min}[R_1(t)]}{(p+1)\sum_{j=0}^{p}a_j^2}$$

Hence the desired result follows by observing that

$$\lambda_{\min}[R_1(t)] \ge \frac{\det R_1(t)}{\{\lambda_{\max}[R_1(t)]\}^{r-1}}$$

and that $\psi(t) = [\epsilon(t-1), \dots, \epsilon(t-r)]$, is an r-dependent sequence. ■
Theorem 1 extends the invertibility result to a very general class of stationary signals with rational spectra. Extension to a limited class of nonstationary signals is also possible using the same approach (c.f. [17]).

Remark
Generally speaking, the higher is dimension r of regression vector and the weaker is mixing condition imposed on $\{\phi(t)\}$, the larger should be N in order to guarantee sufficiently large value of L in (23). We note however that the lower bound on L resulting from our analysis is a *deterministic quantity* obtained without referring to any asymptotic arguments.

5 Results for Sliding Window LS Estimators - the Gaussian Case

5.1 Tighter bounds for $R^{-1}(t)$

Much stronger results can be obtained if we assume that the sequence $\{\phi(t)\}$ is normally distibuted. Assume, for convenience that $N = mK$. Then we have the following

Lemma 4. *If the sequence $\{\phi(t)\}$ is stationary, Gaussian and m-dependent then*

$$\frac{R_0^{-1}}{N} \le E[R^{-1}(t)] \le \frac{R_0^{-1}}{N - m(r+1)} \qquad \forall t \tag{31}$$

Proof
Observe that

$$R(t) = \sum_{j=1}^{m}G_j(t) \tag{32}$$

where

$$G_j(t) = \sum_{i=0}^{K-1}\phi(t-j-im+1)\phi^T(t-j-im+1)$$

Since the sequences $\{\phi(t-j-im+1),\ i \ge 0\}$ are i.i.d. and Gaussian, the matrices $G_j(t)$ are Wishart-distributed with K degrees of freedom

$$G_j(t) \sim W(KR_o, K) \tag{33}$$

Hence, using properties of the inverted Wishart distribution [16]

$$E[G_j^{-1}(t)] = \frac{R_0^{-1}}{K-r-1} \tag{34}$$

Using the inequality (see Appendix 1)

$$\Big(\sum_{j=1}^{m}G_j(t)\Big)^{-1} \le \frac{1}{m^2}\Big(\sum_{j=1}^{m}G_j^{-1}(t)\Big) \tag{35}$$

and combining it with (32)-(34) one obtains

$$E[R^{-1}(t)] \le \frac{R_0^{-1}}{m(K-r-1)}$$

which is nothing but the upper bound in (31). The lower bound in (31) stems from the fact that (see Appendix 2)

$$E[R^{-1}(t)] \ge [E[R(t)]]^{-1} \tag{36}$$

(the matrix variant of the Jensen inequality for inverses).

5.2 Evaluation of parameter tracking bounds

Several conclusions can be drawn from (31) for the sliding window LS estimators. First, observe that for the rectangular window

$$\tilde{R}(t) = \frac{1}{N}R(t)$$

and hence using (31) and (36) one gets

$$0 \le E[\Delta(t)] \le \frac{m(r+1)}{N - m(r+1)}R_0^{-1}$$

Consequently, for $N > m(r+1)$ we have (c.f. Lemma 1)

$$E[\|\hat{\alpha}(t) - \hat{\alpha}(t)\|^2 \le \frac{b}{N} \tag{37}$$

where b - we emphasise this fact strongly - is a *determinstic* constant not depending on N and t. We will look for the bound on the mean square parameter tracking error in the form

$$E[\|\hat{\alpha}(t) - \alpha(t)\|^2] \leq D(N)$$

First, we consider the case of time-invariant parameters: $\alpha(t) = \alpha_0$. One can show that under (A1)-(A3) $\hat{\alpha}(t)$ is an unbiased estimate of α_0 and

$$\mathrm{cov}[\hat{\alpha}(t)] = \rho_0 E[R^{-1}(t)]$$

Consequently

$$D(N) = \frac{D_1}{N} \tag{38}$$

which parallels (9), the result derived by Macchi and Eweda for LMS filters. The counterpart of (10) can be derived using the inequality (16). Note that

$$E[\|\hat{\hat{\alpha}}(t) - \alpha(t)\|^2] = E[\|\hat{\hat{\alpha}}(t) - \bar{\alpha}(t)\|^2] + E[\|\bar{\alpha}(t) - \alpha(t)\|^2] \tag{39}$$

where $\{\bar{\alpha}(t)\}$ denotes the average path of parameter estimates

$$\bar{\alpha}(t) = E[\hat{\hat{\alpha}}(t)|A(t)] = \sum_{i=0}^{\infty} w(i)\alpha(t-i) = \frac{1}{N}\sum_{i=0}^{N-1} \alpha(t-i)$$

$$A(t) = \{\alpha(t), \alpha(t-1), \ldots\}$$

and observe that there is no cross-coupling term on the right-hand side of (39) due to orthogonality of $\hat{\hat{\alpha}}(t) - \bar{\alpha}(t)$ and $\bar{\alpha}(t) - \alpha(t)$.
Assuming that the true parameter trajectory can be modelled as random walk in sufficiently long but finite time interval $T = [t_1, t_2]$, $t_2 - t_1 >> N$, one can show that $(t \in T)$: $E[\|\hat{\hat{\alpha}}(t) - \bar{\alpha}(t)\|^2] = 0(1)$, $E[\|\bar{\alpha}(t) - \alpha(t)\|^2] = 0(N)$ resulting in

$$D(N) = \frac{D_1}{N} + D_2 N \tag{40}$$

We note however, that (40) is - unlike (10) - a *local* result, valid for finite, though possibly very long time intervals. Extension to infinite time intervals is forbidden under (A3) (only the mean square bounded parameter trajectories can be analyzed in the present framework).

5.3 Extension to the case of non-uniform weighting and non-Gaussian regressors

By applying the central limit theorem to the properly normalized elements of the matrix $\tilde{R}(t) - R_0$ and using the appropriate truncation technique one can show that

$$E[\|\hat{\alpha}(t) - \hat{\hat{\alpha}}(t)\|^2] = 0\left(\frac{1}{\iota}\right)$$

in the case of non-uniform weighting and non-Gaussian regressors. This, however, is based on asymptotic theory we were trying to avoid so far. Hence, it holds *only* as long as $\iota \mapsto \infty$.
A more conservative but non-asymptotic bound can be obtained using the Schwartz inequality. Observe that

$$\mathrm{tr}\{E[\Delta(t)]\} = \mathrm{tr}\{E[\tilde{R}^{-1}(t)(R_0 - \tilde{R}(t))R_0^{-1}]\} = E[\mathrm{tr}\{\Delta_1(t)\Delta_2(t)\}]$$

where

$$\Delta_1(t) = R_0^{-1}\tilde{R}^{-1}(t), \quad \Delta_2(t) = R_0 - \tilde{R}(t)$$

Let $\|\Delta(t)\|^2 = \mathrm{tr}\{\Delta(t)\Delta^T(t)\}$. Using Schwartz inequality and (36) one gets

$$0 \leq \mathrm{tr}\{E[\Delta(t)]\} \leq \left(E[\|\Delta_1(t)\|^2]\right)^{\frac{1}{2}}\left(E[\|\Delta_2(t)\|^2]\right)^{\frac{1}{2}}$$

By a similar argument to that used in section 5.1 $E[\|\Delta_1(t)\|^2] = 0(1)$. In the case of i.i.d. regressors

$$E[\|\Delta_2(t)\|^2 = E\left[\|\sum_{i=0}^{\infty} w(i)[\phi(t-i)\phi^T(t-i) - R_0]\|^2\right] = d\sum_{i=0}^{\infty} w^2(i)$$

where $d = E[\|\phi(t)\|^4] - \left(E[\|\phi(t)\|^2]\right)^2$. More generally, one can show that for m-dependent regressors $E[\|\Delta_2(t)\|^2] = 0(1/\iota)$ and hence, combining all the results given above with (17) one gets

$$E[\|\hat{\alpha}(t) - \hat{\hat{\alpha}}(t)\|^2] \leq \frac{c}{\sqrt{\iota}} \tag{41}$$

where c is a deterministic constant not depending on ι and t. Combining this with (39) a bound analogous to (40) (but expressed in terms of ι) can also be derived for an arbitrary WLS estimator.
The bound (41) was derived for a system with time-varying coefficients. If the system is time-invariant, i.e. $\alpha(t) = \alpha_0$ the problem is much easier to handle. Due to the mutual independence of the processes $\{\phi(t)\}$ and $\{n(t)\}$, implied by assumption (A2), one has

$$\mathrm{cov}[\hat{\alpha}(t)] = \rho_0 E\left[\tilde{R}^{-1}(t)\left(\sum_{i=0}^{\infty} w^2(i)\phi(t-i)\phi^T(t-i)\right)\tilde{R}^{-1}(t)\right]$$

$$\leq \frac{\rho_0}{\kappa} E\left[\tilde{R}^{-1}(t)\right] \tag{42}$$

where $\kappa = 1/(\max_t w(t))$ is the quantity usually called the effective width of the window (effective number of observations). One can easily show that for a sequence of windows of the same shape but increasing width it holds $\kappa \propto \iota$, i.e. both measures of the window size differ merely by a constant multiplier. According to (42) the fluctuations of WLS parameter estimates $E[\|\hat{\hat{\alpha}}(t) - \alpha_0\|^2] = tr\{\mathrm{cov}[\hat{\hat{\alpha}}(t)]\}$ are, under stationary conditions, inversly proportional to the size of the applied window which is a further generalization of (38).

6 Important Special Case - Exponentially Weighed LS Estimators

Quite clearly, if the window used in the method of WLS is strictly finite-length (i.e. if $w(t) = 0\ \forall t > t_0$) our technical assumption (A5) is practically unavoidable. This is also a limitation of all results obtained using the concept of "inscribed" window. Using a slightly different technique we will show that (A5) is not needed any more if exponential weighting is applied. Rewrite the expression for the exponentially weighted LS estimator in the form

$$\hat{\alpha}(t) = R^{-1}(t)S(t) \tag{43}$$

where now $(0 < \lambda < 1)$:

$$R(t) = \sum_{i=0}^{t-1} \lambda^i \phi(t-i)\phi^T(t-i), \quad S(t) = \sum_{i=0}^{t-1} \lambda^i y(t-i)\phi(t-i)$$

We note that (43) can be recursively updated using (5)-(6) (note: $D(t) = R^{-1}(t)$ if the exact initialization is used, i.e. if $D(t_0)$ is set to $R^{-1}(t_0)$ for sufficiently large t_0). Note also that

$$R(t) = \lambda R(t-1) + \phi(t)\phi^T(t) \tag{44}$$

Exact initialization corresponds to taking $R(0) = 0$ in (44). However, in practice, recursion (6) is started using $D(0) = D_0 > 0$ which amounts to taking

$$R(0) > 0 \tag{45}$$

in (44) and which will play a crucial role in our analysis. We are ready to prove the following:

Lemma 5. *Suppose that $\{\phi(t)\}$ is an i.i.d. sequence such that $E[\phi(t)\phi^T(t)] = R_0 > 0$. Then the condition (23) of Lemma 2 is fulfilled and the number L in (23) can be made arbitrarily large by increasing $\iota = (1+\lambda)/(1-\lambda)$.*

Proof
Note that $R(t) \geq \lambda^r R(t-r) + \lambda^{r-1}Q(t)$ where $Q(t) = \sum_{i=0}^{r-1} \phi(t-i)\phi^T(t-i)$ and

$$\det R(t) \geq \det R'(t) + \det Q'(t) \tag{46}$$

where $R'(t) = \lambda^r R(t-r)$, $Q'(t) = \lambda^{r-1}Q(t)$.
We have

$$\begin{aligned} P(\det R(t) \leq x) &= P(\det R(t) \leq x | \det Q'(t) < x_0) P\left(\det Q'(t) < x_0\right) \\ &+ P(\det R(t) \leq x | \det Q'(t) \geq x_0) P\left(\det Q'(t) \geq x_0\right) \end{aligned}$$

Using (46) and the fact that the matrices $Q'(t)$ and $R'(t)$ are independent we obtain $P(\det R(t) \leq x | \det Q'(t) \geq x_0) = 0$ for all $x < x_0$, and

$$P(\det R(t) \leq x | \det Q'(t) < x_0) \leq P(\det R'(t) \leq x)$$

which results in

$$P(\det R(t) \leq x) \leq P(\det R'(t) \leq x) p_0' \tag{47}$$

where

$$p_0' = P(\det Q'(t) < x_0) = \frac{p_0}{\lambda^{r(r-1)}} \tag{48}$$

and

$$p_0 = P(\det Q(t) < x_0) < 1 \tag{49}$$

(since $E[\phi(t)\phi^T(t)] > 0$).
We will use inductive reasoning. Suppose that our assertion is true for $R(t-r)$, i.e. $P(\det R(t-r) \leq x) \leq \eta x^L$.
Then $\forall x < x_0$:

$$P(\det R'(t) \leq x) = P(\lambda^{r^2} \det R(t-r) \leq x) \leq \frac{\eta}{\lambda'} x^L, \quad \lambda' = \lambda^{Lr^2}$$

and hence, according to (47) $P(\det R(t) \leq x) \leq (p_0'\eta/\lambda')x^L \leq \eta x^L$, i.e. our

assertion is true for $R(t)$ provided that $\lambda \geq \lambda_0$ such that

$$\ell n \lambda_0 = \frac{\ell n p_0}{(L+1)r^2 - r}$$

Since $R(t) \geq \lambda^r R(0), \quad t = 1, \ldots, r$ we get $P(\det R(t) \leq x) = 0 \leq \eta x^L \quad t = 1, \ldots, r$ for all $x < x_0 = \det(\lambda^r R(0))$ and arbitrarily large L. Our assertion is therefore true for any t.
Finally, note that arbitrarily large value of L can be guaranteed in (23) provided that the forgetting constant λ is sufficiently close to 1. ■

Extension of Lemma 5 to ψ-mixing sequences (which includes m-dependence as a special case) and weaker covariance conditions is straightforward. Therefore, only assumptions (A1)-(A3) are needed to guarantee boundedness of the mean square parameter tracking error if the method of exponential weighting is used!

7 Statistical Robustness

On the qualitative level the results obtained in previous sections raise several important issues which can be easily overlooked if a mechanical, "bookkeeping" approach towards certain mathematical details is adopted.
First of all, one should realize that results of Section 4 indicate certain non-robustness properties - as far as statistical analysis of WLS filters is concerned - of strictly finite-length windows. Assumption (A5) admits a large class of continuous distributions but rules out all discrete ones. Is it a serious limitation? In a way it is. In the world of computers and digital processing, random variables with continuous distributions belong in mathematical "science fiction". Any form of quantization turns a continuous random variable into a discrete one. Hence, results of Section 4 are not robust against quantization. The situation is essentially different if exponential weighting is applied. Let

$$p_{\min} = P(\det Q(t) = 0)$$

If $p_{\min} = 0$ one can make p_0, given by (49), arbitrarily small by decreasing x_0. This corresponds to the case where there is no discrete or singular component in $F(\phi)$. Presence of such components, however, does not destroy invertibility of $R(t)$ which was the case for finite length windows. Instead, it sets a lower bound on the forgetting constant λ

$$\ell n \lambda_{\min} = \frac{\ell n p_{\min}}{(L+1)r^2 - r}$$

i.e. the minimum equivalent width of the windows for which invertibility is guaranteed.
The related question is that of practical significance of the results based on (A5). If regression vector $\phi(t)$ is subject to quantization - as it always happens in practice - the expected value of the mean square parameter tracking error is, at least theoretically, infinite and all the results laboriously derived in Sections 4 and 5 have hardly any meaning. Or have they not? The point is that even under very crude quantization (like the one considered in our example in Section 4) the probability $P(\det R(t) = 0)$ will take extremely small values for typical window sizes. One can therefore argue that the insights provided by our analysis can be applied also, quite safely, to the case where (A5) is formally not valid. In particular, if one is not scared of Maxwell's demon, one can still explain properties of the WLS filter in terms of properties of its "idealized" version without being wrong once in one billion years!

Acknowledgement
The first author would like to thank Prof. P. Hall from the Department of Statistics, A.N.U. for his comments on the invertibility problem. We would also like to acknowledge the many helpful remarks of Dr. O. Macchi, which improved the readability of this paper.

References

[1] Goodwin, G. and K. Sin, *"Adaptive Filtering, Prediction and Control"*, Prentice Hall, 1984.

[2] Cioffi, J., "Limited-precision effects in adaptive filtering", *IEEE Trans. Circ. and Syst.*, Vol. CAS-34, pp. 821-833, 1987.

[3] Eleftheriou, E. and D.D. Falconer, "Tracking properties and stady state performance of RLS adaptive filter algorithms", *IEEE Trans. ASSP, Vol. ASSP-34*, pp. 1097-1111, 1986.

[4] Macchi, O. and E.Eweda, "Second order convergence analysis of stochastic adaptive linear filtering", *IEEE Trans. Auto. Control*, Vol. AC-28, pp. 76-85, 1983.

[5] Eweda, E. and O. Macchi, "Tracking error bounds of adaptive nonstationary filtering", *Automatica*. Vol. 21, pp. 293-302, 1985.

[6] Macchi, O., "Optimization of adaptive identification for time-varying filters", *IEEE Trans. Auto. Control*, Vol. AC-31, pp. 283-287, 1986.

[7] Eweda, E. and O. Macchi "Convergence of the RLS and LMS adaptive filters", *IEEE Trans. Circuits and Systems*, Vol. CAS-34, pp. 799-803, 1987.

[8] Niedźwiecki, M. "On the localized estimators and generalized Akaike's criteria", *IEEE Trans. Auto. Control*, Vol. AC-29, pp. 970-983, 1984.

[9] Niedźwiecki, M. "First-order tracking properties of weighted least squares estimators", *IEEE Trans. Auto. Control*, Vol. AC-33, pp. 94-96, 1988.

[10] Niedźwiecki, M. "On tracking characteristics of weighted least squares estimators applied to non-stationary system identification", *IEEE Trans. Auto. Control*, Vol. AC-33, pp. 96-98, 1988.

[11] Niedźwiecki, M. "Optimization of the window shape in weighted least squares identification of a class of nonstationary systems", *Proc. 7th Conf. on Analysis and Optimization of Systems*, Antibes, France, 1986.

[12] Benveniste, A. and G. Ruget, "A measurement of tracking capability of recursive algorithms with constant gaims", *IEEE Trans. Auto. Control*, Vol. AC-27, pp. 639-649, 1982.

[13] Benveniste, A., "Design of adaptive algorithms for the tracking of time-varying systems", *Int. Journ. of Adaptive Contr. and Signal Proc.*, Vol. 1, pp. 3-29, 1987.

[14] Kushner, H.J. and H. Huang, "Asymptotic properties of stochastic approximations with constant gains", *SIAM J. Contr.*, Vol. 19, pp. 87-105, 1981.

[15] Ljung, L., "Adaptation and tracking in system identification", *Proc. IFAC Symp. on Identification and System Parameter Estimation*, Beijing, Vol. 1, pp. 1-10, 1988.

[16] Das Gupta, S., "Some aspects of discrimination function coefficients", *Sankhyā*, Vol. A-30, pp. 387-400, 1968.

[17] Green, M. and J.B. Moore, "Persistence of excitation in linear systems", *Systems & Control Letters*, Vol. 7, pp.351-360, 1986; see also *Preprints of ACC*, 1985, pp. 412-417, Boston.

[18] Chen, H.F. and L. Guo, "Adaptive Control via consistent estimation for deterministic systems", *Int. J. Contr.* Vol. 45, No. 6, 1987, pp. 2183-2202.

[19] Macchi, O. and E. Eweda, "Compared speed and accuracy of the RLS and LMS algorithms with constant forgetting factors," *Traitement du signal*, Vol. 22, pp. 255-267, 1988.

Appendix 1 (derivation of (35))

The inequality can be easily proved by induction using the following proposition:

Proposition. *For any two positive-definite matrices A and B and any integer m*

$$m^2 A^{-1} + B^{-1} \geq (m+1)^2 (A+B)^{-1}$$

Proof
Proof is straightfoward in the scalar case. The multivariate case can be converted into the scalar one by performing the simultaneous diagonalization of matrices A and B (note: there exists a real matrix Q and a positive diagonal matrix Λ such that: $Q^T A Q = \Lambda$ and $Q^T B Q = I$). ■

Suppose that (35) holds for a certain m. Then, using the result of the proposition above, one gets

$$\sum_{j=1}^{m+1} G_j^{-1}(t) = \sum_{j=1}^{m} G_j^{-1}(t) + G_{m+1}^{-1}(t) \geq m^2 \Big(\sum_{j=1}^{m} G_j(t) \Big)^{-1} + G_{m+1}^{-1}(t)$$

$$\geq (m+1)^2 \Big(\sum_{j=1}^{m+1} G_j(t) \Big)^{-1}$$

i.e. (35) is also true for $m+1$. Since it is also true for $m = 1$ (c.f. proposition above) it remains valid for any m.

Appendix 2 (proof of (36))

Observe that for $R_0 = E[\tilde{R}(t)]$, $\tilde{R}(t) = R(t)/N$: $E[\tilde{R}^{-1}(t) - R_0^{-1}] =$
$= E\big[(\tilde{R}^{-1}(t) - R_0^{-1})\tilde{R}(t)(\tilde{R}^{-1}(t) - R_0^{-1})\big] \geq 0$ which is nothing but (36).