

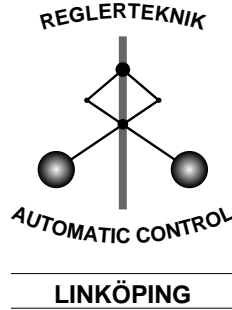
Recursive Least Squares and Accelerated Convergence in Stochastic Approximation Schemes

Lennart Ljung

Department of Electrical Engineering
Linköping University, S-581 83 Linköping, Sweden

WWW: <http://www.control.isy.liu.se>

Email: ljung@isy.liu.se



Report no.: LiTH-ISKY-R-2291

For The the Special Issue of *Adaptive Control and Signal
Processing* in Honor of Ya.Z. Tsypkin

Technical reports from the Automatic Control group in Linköping are available by anonymous ftp at the address [ftp.control.isy.liu.se](ftp://ftp.control.isy.liu.se). This report is contained in the pdf file 2291.pdf.

Recursive Least Squares and Accelerated Convergence in Stochastic Approximation Schemes

Lennart Ljung
Department of Electrical Engineering
Linköping University
S-581 83 Linköping, Sweden
e-mail: `ljung@isy.liu.se`

October 4, 2000

Abstract

So called accelerated convergence is an ingenious idea to improve the asymptotic accuracy in stochastic approximation (gradient based) algorithms. The estimates obtained from the basic algorithm are subjected to a second round of averaging, which leads to optimal accuracy for estimates of time-invariant parameters. In this contribution some simple calculations are used to get some intuitive insight into these mechanisms. Of particular interest is to investigate the properties of accelerated convergence schemes in tracking situations. It is shown that a second round of averaging leads to the recursive least squares algorithm with a forgetting factor. This also means that in case the true parameters are changing as a random walk, accelerated convergence does not, typically, give optimal tracking properties.

1 Introduction

Tracking of time varying parameters is a basic problem in many applications, and there is a considerable literature on this problem. See, among many references, e.g. [7], [5], [3].

A typical set-up is as follows: Suppose observed data $\{y(t), \varphi(t), t = 1, \dots\}$ are generated by the linear regression structure

$$\begin{aligned} y(t) &= \theta^T(t) \varphi(t) + e(t) \\ \theta(t) &= \theta(t-1) + w(t) \end{aligned} \tag{1}$$

The generic algorithm for estimating $\theta(t)$ in (1) is

$$\hat{\theta}(t) = \hat{\theta}(t-1) + \mu_t P(t) \varphi(t) (y(t) - \hat{\theta}^T(t-1) \varphi(t)) \tag{2}$$

The choices of step size μ_t and modifying matrix $P(t)$ has been the subject of extensive discussion and analysis, which we will not dwell upon here. We merely remark that in case $\{e(t)\}$ is white noise with time-invariant covariance and $w(t) \equiv 0$ (i.e. the parameter vector $\theta(t)$ is indeed constant) then the choice

$$\mu_t P(t) = \left[\sum_{k=1}^t \varphi(k) \varphi^T(k) \right]^{-1} \quad (3)$$

leads to the least squares estimate $\hat{\theta}(t)$, which indeed has the optimal accuracy for this case. That is, the covariance matrix of the asymptotic distribution of $\hat{\theta}(t)$ meets the Cramer-Rao bound.

This optimal choice (3) may require a substantial amount of calculations, if the dimension of φ is large. Partly because of this simpler choices of $P(t)$ in (2) have been attractive. The LMS algorithm uses $P(t) = I$ (identity matrix) which is a gradient based update algorithm. This gives an order of magnitude less calculations. The disadvantage with this choice is that the accuracy of the estimate (or “the convergence rate”) could be much worse. The rule of thumb is that the worse conditioned the matrix (3) is, the worse convergence rate.

Now, the ingenious observation and analysis of [6], [2] is as follows:

1. Use (2) with $P(t) = I$ and μ_t a sequence that decays slower than $1/t$.
2. Average the estimates $\hat{\theta}(t)$ obtained from (2):

$$\hat{\theta}(t) = \frac{1}{t} \sum_{k=1}^t \hat{\theta}(k) \quad (4)$$

Then $\hat{\theta}(t)$ will have the same optimal asymptotic accuracy as the choice (3) would give, but at a considerably lower computational cost. The approach has been termed “accelerated convergence”.

So far we only have discussed the time-invariant parameter case: $w(t) = 0$ in (2). In applications, the most important use of adaptive algorithms like (3) is really to deal with time-varying properties. It is therefore interesting to look into what accelerated convergence schemes - “second round of averaging” - like (4) will do for the tracking case. It is the purpose of this contribution to do that. It will be done using simple and essentially algebraic calculations for the case where the regressors φ are white. These calculations will also provide some insights into how the averaging like (4) “thinks and works”. For the case of general regressors we will use also results from [1] and [4].

2 Optimal Tracking Algorithms

What is the best choice of $\mu_t P(t)$ in (2) for time-varying parameter $\theta(t)$? In case $\{e(t)\}$ and $\{w(t)\}$ in (1) - (2) are white Gaussian noises, it is well known

that the optimal tracking algorithm is provided by the Kalman filter, which uses

$$\mu_t P(t) = \frac{S(t-1)}{R_2 + \varphi^T(t)S(t-1)\varphi(t)} \quad (5)$$

$$S(t) = S(t-1) + R_1 - \frac{S(t-1)\varphi(t)\varphi^T(t)S(t-1)}{R_2 + \varphi^T(t)S(t-1)\varphi(t)} \quad (6)$$

Here R_1 is the covariance matrix of $w(t)$ and R_2 is the variance of $e(t)$ (Here assumed to be a scalar).

It may be of interest to interpret this solution in a pragmatic and simplified way, see also [3]. (This interpretation is not necessary for the main result of this paper, and the reader may skip directly to Section 3.): For “small” matrices R_1 (slowly varying systems) we can approximately describe this solution as follows. Let

$$R_1 = \gamma^2 \bar{R}_1$$

Then

$$\mu = \gamma \quad (7)$$

$$P(t) \approx \bar{P} \cdot \frac{1}{R_2} \quad (8)$$

with

$$\bar{R}_1 = \frac{1}{R_2} \bar{P} Q \bar{P} \quad (9)$$

where

$$Q = E \varphi(t) \varphi^T(t) \quad (10)$$

The matrix is then also the value of the (optimal) covariance matrix of the error

$$\Pi = E \left(\hat{\theta}(t) - \theta(t) \right) \left(\hat{\theta}(t) - \theta(t) \right)^T \quad (11)$$

For an arbitrary choice of μ and $P(t) \equiv P$ in (3) the same type of calculations show that the error covariance matrix Π in (11) is obtained as the solution to

$$PQ\Pi + \Pi QP = \mu R_2 P Q P + \frac{\gamma^2}{\mu} \bar{R}_1$$

(see e.g. [3]). Minimizing this expression with respect to P and μ gives (of course) the solution (7)-(9).

3 Tracking algorithms with a second round of averaging

Now, the optimal tracking algorithm (2), (5), (6) requires knowledge of R_2 . One of the most common *ad hoc* choices of algorithms is instead to use a *forgetting factor recursive least squares* algorithm: This is obtained by

$$\begin{aligned}\hat{\theta}_{\text{ls}}(t) &= \hat{\theta}_{\text{ls}}(t-1) + \frac{S(t-1)\varphi(t)}{(1-\rho) + \varphi^T(t)S(t-1)\varphi(t)}(y(t) - \varphi^T(t)\hat{\theta}_{\text{ls}}(t-1)) \\ S(t) &= \left[S(t-1) - \frac{S(t-1)\varphi(t)\varphi^T(t)S(t-1)}{(1-\rho) + \varphi^T(t)S(t-1)\varphi(t)} \right] / (1-\rho)\end{aligned}\quad (12)$$

Here $1-\rho$ is the forgetting factor. The estimate can also be explicitly expressed as (with suitable initial conditions $\hat{\theta}_{\text{ls}}$ and $S(0)$)

$$\hat{\theta}_{\text{ls}}(t) = \left[\sum_{k=1}^t (1-\rho)^{t-k} \varphi(k)\varphi^T(k) \right]^{-1} \sum_{k=1}^t (1-\rho)^{t-k} \varphi(k)y(k) \quad (13)$$

The estimate (12) could be costly to implement for large dimension of φ . It is therefore of interest to investigate what a second round of averaging would do in this case. The tracking analog of (2), (4) would be as follows. First form

$$\hat{\hat{\theta}}(t) = \hat{\hat{\theta}}(t-1) + \mu P \varphi(t)(y(t) - \varphi^T(t)\hat{\hat{\theta}}(t-1)) \quad (14)$$

($P = I$ would be the stochastic gradient algorithm). To apply the averaging idea (accelerated convergence) in the tracking case would be to form a time-weighted average

$$\hat{\theta}(t) = (1-\rho)\hat{\theta}(t-1) + \rho\hat{\hat{\theta}}(t) \quad (15)$$

The idea is that $0 < \rho \ll \mu$, so that some real averaging takes place.

We shall show that the estimates $\hat{\theta}_{\text{ls}}$ and $\hat{\theta}$ are close. The objective with this proof is two-fold. One is to establish the counterpart of accelerated convergence in the tracking case. The other is to use simple, mostly algebraic calculations, which will give some insights into the mechanisms of how the second round of averaging ((4) and (15)) works. This could be of value also for the time invariant case.

We start by a result that shows how the matrix inversion in (3) and (13) is accomplished from the sum of a geometric series in (15).

4 A Basic Relationship

Consider the following recursion formula:

$$x(t) = (I - \mu A)x(t-1) + \mu w(t) \quad (16)$$

(Clearly, this corresponds to a typical error propagation equation for adaptive algorithms, see the next section). Let us then average the sequence $\{x(t)\}$ by

$$z(t) = (1 - \rho)z(t-1) + \rho x(t) \quad (17)$$

Equation (17) means that

$$z(N) = \rho \sum_{t=1}^N (1 - \rho)^{N-t} x(t) \quad (18)$$

The equally weighted average (4) can be seen as the limit as $\rho \rightarrow 0$. Formally it corresponds to the time varying choice $\rho = \rho(t) = 1/t$. Let us also introduce

$$\hat{z}(N) = \rho \sum_{t=1}^N (1 - \rho)^{N-t} A^{-1} w(t) \quad (19)$$

We shall prove that z and \hat{z} are close when ρ/μ is small.

Now, solving (16) gives for $x(0) = 0$

$$x(t) = \sum_{k=1}^t (I - \mu A)^{t-k} \mu w(k) \quad (20)$$

which inserted into (18) yields

$$\begin{aligned} z(N) &= \rho \sum_{t=1}^N \sum_{k=1}^t (1 - \rho)^{N-1} (I - \mu A)^{t-k} \mu w(k) \\ &= \rho \sum_{k=1}^N \left[\sum_{t=k}^N (1 - \rho)^{N-1} (I - \mu A)^{t-k} \right] \mu w(k) \end{aligned} \quad (21)$$

Let us consider the inner sum:

$$\begin{aligned} \sum_{t=k}^N (1 - \rho)^{N-t} (I - \mu A)^{t-k} &= (1 - \rho)^N (I - \mu A)^{-k} \left[\sum_{t=k}^N \left(\frac{I - \mu A}{1 - \rho} \right)^t \right] \\ &= (1 - \rho)^N (I - \mu A)^{-k} \left[\left(I - \frac{I - \mu A}{1 - \rho} \right)^{-1} \times \left(\left(\frac{I - \mu A}{1 - \rho} \right)^k - \left(\frac{I - \mu A}{1 - \rho} \right)^{N+1} \right) \right] \end{aligned}$$

For the moment, denote

$$f(A, \rho, \mu) = \mu \left(I - \frac{I - \mu A}{1 - \rho} \right)^{-1} \quad (22)$$

The inner sum is thus given by

$$\frac{1}{\mu} f(A, \rho, \mu) \cdot (1 - \rho)^{N-k} + \frac{1}{\mu} f(A, \rho, \mu) \cdot \frac{(I - \mu A)}{1 - \rho} \cdot (I - \mu A)^{N-k}$$

Inserting this into (21) gives

$$\begin{aligned}
z(N) &= f(A, \rho, \mu) \sum_{k=1}^N (1 - \rho)^{N-k} \rho w(k) + \\
&\quad + f(A, \rho, \mu) \frac{\rho}{\mu} \cdot \frac{(I - \mu A)}{1 - \rho} x(N)
\end{aligned} \tag{23}$$

From (19) we find that

$$\begin{aligned}
z(N) - \hat{z}(N) &= (f(A, \rho, \mu) - A^{-1}) A \hat{z}(N) + f(A, \rho, \mu) \frac{\rho}{\mu} \cdot \frac{(I - \mu A)}{1 - \rho} x(N) \\
&= \frac{\rho}{\mu} \left(I - \frac{\rho}{\mu} A^{-1} \right)^{-1} (A^{-1} - \mu I) (\hat{z}(N) + x(N))
\end{aligned} \tag{24}$$

where we in the second step inserted the definition of f .

We can sum up these simple algebraic relationships as a lemma:

Lemma 1. Let $x(t)$ and $z(t)$ be given by (16) and (17), respectively, which ρ and μ positive. Let $\hat{z}(t)$ be given by

$$\hat{z}(t) = (1 - \rho) \hat{z}(t-1) + \rho A^{-1} w(t) \tag{25}$$

Let $\|A^{-1}\| = \alpha$. Then, assuming that $\frac{\rho}{\mu} < 1/\alpha$ we have

$$|z(t) - \hat{z}(t)| \leq \frac{\rho}{\mu} \frac{\alpha + \mu}{1 - \alpha \frac{\rho}{\mu}} [|\hat{z}(t)| + |x(t)|]$$

Note that the lemma describes an algebraic relationship, and does not depend on the particular sequence w or the choice of A (as long as it is invertible). The lemma shows how the geometric series, inherent in the second round of averaging provides the matrix inversion that is crucial for obtaining the optimal estimates.

5 Connections to Recursive Least Squares

We shall now show that the estimate $\hat{\theta}$ defined by (14)–(15) will be arbitrarily close to the recursive least squares estimate (12). To focus on the basic mechanisms we first give a simple and direct proof for the case where the regressors φ are independent:

Theorem 1: Let $\varphi(t)$ be a sequence of independent random vectors with $E\varphi(t)\varphi^T(t) = Q > 0$ and $E|\varphi(t)|^4 \leq \beta$. Let $y(t)$ be a sequence of random variables with variances bounded by κ , and independent of future φ . ($y(t)$ is otherwise arbitrary and need not be subject to (1).) Let $\hat{\theta}$ be defined by

$$\begin{aligned}
\hat{\hat{\theta}}(t) &= \hat{\hat{\theta}}(t-1) + \mu P \varphi(t) (y(t) - \varphi^T(t) \hat{\hat{\theta}}(t-1)); \quad \hat{\hat{\theta}}(0) = 0 \\
\hat{\theta}(t) &= (1 - \rho) \hat{\theta}(t-1) + \rho \hat{\hat{\theta}}(t); \quad \hat{\theta}(0) = 0.
\end{aligned} \tag{26}$$

with $0 < \rho < 1$, μ positive and $P > 0$. Moreover, let $\hat{\theta}_{\text{ls}}$ be defined by (13), or in recursive form with suitable initial conditions:

$$\begin{aligned}\hat{\theta}_{\text{ls}}(t) &= \hat{\theta}_{\text{ls}}(t-1) + \frac{S(t-1)\varphi(t)}{(1-\rho) + \varphi^T(t)S(t-1)\varphi(t)}(y(t) - \varphi^T(t)\hat{\theta}_{\text{ls}}(t-1)) \\ S(t) &= \left[S(t-1) - \frac{S(t-1)\varphi(t)\varphi^T(t)S(t-1)}{(1-\rho) + \varphi^T(t)S(t-1)\varphi(t)} \right] / (1-\rho)\end{aligned}\quad (27)$$

Let $\alpha = \|(PQ)^{-1}\|$, and assume that $\mu < \|PQ\|$ and $\rho/\mu < 1/\alpha$. Also assume that $E\|\hat{\theta}(t)\|^2 \leq L$. Then

$$\hat{\theta}(t) - \hat{\theta}_{\text{ls}}(t) = C_1\sqrt{\mu}\xi_1(t) + C_2\frac{\rho}{\mu}\xi_2(t) + C_3\sqrt{\rho}\xi_3(t)\hat{\theta}_{\text{ls}}(t) \quad (28)$$

where ξ_k are random vectors and matrices with variances norm-bounded by 1. The constants C_i can be derived from the assumptions as

$$C_1 = \sqrt{\beta\alpha L}\|P\|, \quad C_2 = \frac{\alpha + \mu}{1 - \alpha\rho/\mu}\sqrt{\kappa\|Q\|(\|Q^{-1} + \alpha\|P\|)}, \quad C_3 = \sqrt{\beta}\|Q^{-1}\|$$

Remark 1: The first term $\sqrt{\mu}$ is conservative, and it should be possible to improve that to μ .

Remark 2: The conditions of the theorem guarantee that $I - \mu PQ$ is a stable matrix. Under the independence assumptions of the theorem it can be shown, in a straightforward manner, that $E\|\hat{\theta}\|^2$ is bounded, so this condition does not have to be stated as an assumption. It has been included only to make the proof less technical.

Proof: In addition to the estimates $\hat{\theta}$, $\hat{\theta}$, and $\hat{\theta}_{\text{ls}}$ defined in the text, let us introduce

$$x(t) = (I - \mu PQ)x(t-1) + \mu P\varphi(t)y(t) \quad (29)$$

$$z(t) = (1 - \rho)z(t-1) + \rho x(t) \quad (30)$$

$$\hat{z}(t) = (1 - \rho)\hat{z}(t-1) + \rho Q^{-1}\varphi(t)y(t) \quad (31)$$

We shall write

$$\hat{\theta} - \hat{\theta}_{\text{ls}} = \hat{\theta} - z + z - \hat{z} + \hat{z} - \hat{\theta}_{\text{ls}} \quad (32)$$

which will account for the three terms in (28).

Let us start with the last one: We have

$$\hat{z}(t) = Q^{-1} \sum_{k=1}^t (1 - \rho)^{t-k} \rho \varphi(k)y(k)$$

Let

$$P(t) = \rho \sum_{k=1}^t (1 - \rho)^{t-k} \varphi(k) \varphi^T(k)$$

Then, ignoring an exponentially decaying term $(1 - \rho)^t Q$ we have

$$P(t) - Q = \rho \sum_{k=1}^t (1 - \rho)^{t-k} (\varphi(k) \varphi^T(k) - Q) \quad (33)$$

This is a weighted sum of zero mean, and independent random variables, and it is immediate to verify (using the geometric series) that the variance of the matrix elements is bounded by

$$\rho^2 \sum_{k=1}^t (1 - \rho)^{2(t-k)} \beta \approx \frac{\beta}{2 - \rho}$$

Now, using (13) we find that

$$\hat{z}(t) - \hat{\theta}_{\text{ls}}(t) = (Q^{-1} - P^{-1}(t))P(t)\hat{\theta}_{\text{ls}}(t) = Q^{-1}(P(t) - Q)\hat{\theta}_{\text{ls}}(t) \quad (34)$$

This shows the third term of (28).

For $z(t) - \hat{z}(t)$ we apply Lemma 1. Note that the variances of $\hat{z}(t)$ and $x(t)$ (being weighted averages of $Q^{-1}\varphi(k)y(k)$ and of $P\varphi(k)y(k)$) are bounded by the variances of its terms.

It now only remains to consider $\hat{\theta}(t) - z(t)$. This quantity obeys

$$\hat{\theta}(t) - z(t) = (1 - \rho)(\hat{\theta}(t-1) - z(t-1)) + \rho(\hat{\hat{\theta}}(t) - x(t))$$

This means that the variance of $\hat{\theta}(t) - z(t)$ is bounded by the variance of $h(t) = (\hat{\hat{\theta}}(t) - x(t))$. (This is where the conservativeness mentioned in Remark 1 enters. The term $h(t)$ is zero mean, and some variance reduction when averaged over is to be expected.)

Let us therefore consider $h(t)$. It obeys

$$h(t) = (I - \mu PQ)h(t-1) + \mu P(Q - \varphi(t)\varphi^T(t))\hat{\hat{\theta}}(t-1) \quad (35)$$

To compute the variance of h , multiply each side by its transpose and take expectation. Since, by assumption, $\varphi(t)$ is independent of the past, no cross terms will make any contribution and we get with $\Pi(t) = Eh(t)h^T(t)$

$$\Pi(t) = (I - \mu PQ)\Pi(t-1)(I - \mu PQ)^T + \mu^2 PKPR(t) \quad (36)$$

where K contains the fourth moment terms of φ and $R(t)$ contains the second moment terms of $\hat{\hat{\theta}}(t-1)$, which are bounded. This means that the second moment of h is bounded by μ , which concludes the proof of the theorem.

The theorem shows that averaging the simple estimate $\hat{\hat{\theta}}$ can give an estimate arbitrarily close to the recursive least squares estimate. Selecting, e.g., $\mu = \rho^{2/3}$ gives a deviation between the two estimates of order of magnitude $\rho^{1/3}$.

Non-white Regressors

The assumption of white regressors makes certain technical aspects easier, and the proof was not obscured by issues that have been extensively dealt with elsewhere.

The assumption of white regressors was used in two places:

- In the calculation of the variance of (33). It is elementary to show that the same result holds (with another C_3) as soon as the correlation among the φ decays sufficiently fast.
- When ignoring the cross term in (35). Detailed analysis of this cross term in the non-white case has been carried out in, e.g., [4], for the case that φ is ϕ -mixing with function $\phi(M)$. In Equation (19) of that paper this cross term is called $\alpha(t)$. Its effect is that the result still holds with $\sqrt{\mu}$ replaced by $\sigma(\sqrt{\mu})$, where σ is defined by

$$\sigma(\mu) = \min_{M>0} (\phi(M) + \mu M)$$

where $\phi(M)$ is the ϕ -mixing function. See also [1] for related results.

6 Conclusions

The question asked in this paper was what the effect of “accelerated convergence” schemes for stochastic approximation would achieve in a tracking situation.

The basic accelerated scheme will then consist of two averaging algorithms with constant, and different step sizes. The first one uses larger steps and is typically a stochastic gradient (LMS) scheme. The second one performs exponential smoothing of the estimates obtained from the first step.

Simple calculations (asymptotic in the step sizes) show that what is obtained in this way is the recursive least squares estimate, corresponding to a forgetting factor given by the second step’s exponential forgetting. The step size and update direction in the first algorithm do not affect the resulting estimate (asymptotically).

Thus, the accelerated scheme will be a cheap way to obtain asymptotically the recursive least squares estimate. However, this also means that the accelerated convergence scheme does not give optimal tracking properties, since that would require the algorithm (2), (5), (6).

References

- [1] L. Guo and L. Ljung. Performance analysis of general tracking algorithms. *IEEE Trans. Automatic Control*, AC-40:1388–1402, August 1995.

- [2] H. J. Kushner and J. Yang. Stochastic approximation with averaging of the iterates: Optimal asymptotic rate of convergence for general processes. *SIAM Journal of Control and Optimization*, 31(4):1045–1062, 1993.
- [3] L. Ljung and S. Gunnarsson. Adaptive tracking in system identification - a survey. *Automatica*, 26(1):7–22, 1990.
- [4] L. Ljung and P. Priouret. A result of the mean square error obtained using general tracking algorithms. *Int. J. of Adaptive Control and Signal Processing*, 5(4):231–250, 1991.
- [5] L. Ljung and T. Söderström. *Theory and Practice of Recursive Identification*. MIT press, Cambridge, Mass., 1983.
- [6] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAIM J. Control Optim.*, 30:838–855, 1992.
- [7] B. Widrow and S. Stearns. *Adaptive Signal Processing*. Prentice-Hall, Englewood-Cliffs, 1985.