Aktivitet Frsrjning(ADJ DTJ EXT HSK LEK LIU STM UBB USL VR V96
Aspects on Accelerated Convergence in Stochastic Approximation Schemes

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

December 12, 1996

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 and approximate 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.

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., [9], [7], [6].

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

\[
y(t) = \theta^T(t)\varphi(t) + e(t)
\]

\[
\theta(t) = \theta(t-1) + w(t)
\]

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

\[
\hat{\theta}(t) = \hat{\theta}(t-1) + \mu_t P(t)\varphi(t)(y(t) - \hat{\theta}^T(t-1)\varphi(t))
\]

The choices of step size \(\mu_t\) and modifying matrix \(P(t)\) has been the subject to 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

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

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 \( \varphi \) 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 [8], [3] is as follows:

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

\[
\hat{\theta}(t) = \frac{1}{t} \sum_{k=1}^{t} \hat{\theta}(k)
\]

Then \( \hat{\theta}(t) \) will have the same optimal asymptotic accuracy as the choice (3) would give, but at a considerably lower computational cost.

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 very simple and approximate calculations, making use of rather sophisticated averaging results in a pragmatic way - without checking the conditions for their applicability. These calculations will also provide some insights into how the averaging like (4) “thinks and works”.

### 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

\[
\mu_t P(t) = \frac{S(t-1)}{R_2 + \varphi^T(t)S(t-1)\varphi(t)}
\]
\[ S(t) = S(t-1) + R_1 - \frac{S(t-1)\phi(t)\phi^T(t)S(t-1)}{R_2 + \phi^T(t)S(t-1)\phi(t)} \]  
(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).

For “small” matrices \( R_1 \) (slowly varying systems) we can approximately describe this solution as follows. Let

\[ R_1 = \gamma^2 \hat{R}_1 \]

Then

\[ \mu = \gamma \]  
(7)

\[ P(t) \approx \hat{P} \cdot \frac{1}{R_2} \]  
(8)

with

\[ \hat{R}_1 = \frac{1}{R_2} \hat{P} Q \hat{P} \]  
(9)

where

\[ Q = E\phi(t)\phi^T(t) \]  
(10)

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

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

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

\[ PQ\Pi + \PiQP = \mu R_2 PQP + \frac{\gamma^2}{\mu} \hat{R}_1 \]

(see e.g. [6]). Minimizing this expression with respect to \( P \) and \( \mu \) gives (of course) the solution (7)-(9).

3 A Basic Relationship

Consider the following recursion formula:

\[ x(t) = (I - \mu A)x(t-1) + \mu w(t) \]  
(12)

(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

\[ z(t) = (1 - \rho)z(t-1) + \rho x(t) \]  
(13)

Equation (13) means that

\[ z(N) = \rho \sum_{t=1}^{N} (1 - \rho)^{N-t} x(t) \]  
(14)
The equally weighted average (4) can be seen as the limit as \( \rho \to 0 \). Formally it corresponds to the time varying choice \( \rho = \rho(t) = 1/t \).

Now, solving (12) gives for \( x(0) = 0 \)

\[
x(t) = \sum_{k=1}^{t} (I - \mu A)^{t-k} \mu w(k)
\]  

(15)

which inserted into (14) yields

\[
z(N) = \rho \sum_{i=1}^{N} \sum_{k=1}^{t} (1 - \rho)^{N-1} (I - \mu A)^{t-k} \mu w(k) = \]

\[
= \rho \sum_{k=1}^{N} \left[ \sum_{l=k}^{N} (1 - \rho)^{N-1} (I - \mu A)^{t-k} \right] \mu w(k)
\]

(16)

Let us consider the inner sum:

\[
\sum_{l=k}^{N} (1 - \rho)^{N-l} (I - \mu A)^{t-k} =
\]

\[
= (1 - \rho)^{N} (I - \mu A)^{-k} \left[ \sum_{l=k}^{N} \left( \frac{I - \mu A}{1 - \rho} \right)^{l} \right] =
\]

\[
= (1 - \rho)^{N} (I - \mu A)^{-k} \left[ \left( \frac{I - \mu A}{1 - \rho} \right)^{-1} \times \left( \left( \frac{I - \mu A}{1 - \rho} \right)^{k} - \left( \frac{I - \mu A}{1 - \rho} \right)^{N+1} \right) \right]
\]

For the moment, denote

\[
f(A, \rho, \mu) = \mu (I - \frac{I - \mu A}{1 - \rho})^{-1}
\]

(17)

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 (16) gives

\[
z(N) = f(A, \rho, \mu) \sum_{k=1}^{N} (1 - \rho)^{N-k} \rho w(k) = \]

\[
+ f(A, \rho, \mu) \frac{\rho}{\mu} \cdot \frac{(I - \mu A)}{1 - \rho} x(N)
\]

(18)
For $f(A, \rho, \mu)$ we find that
\[
\mu \left( I - \frac{I - \mu A}{1 - \rho} \right)^{-1} = (1 - \rho)(A - \frac{\rho}{\mu} I)^{-1} = \\
= A^{-1}(1 - \rho) \left( I - \frac{\rho}{\mu} A^{-1} \right)^{-1} = \\
A^{-1} + \rho + \frac{\rho}{\mu} A^{-1} + \text{ higher order terms}
\]

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

**Lemma 3.1.** Let $x(t)$ and $z(t)$ be given by (12) and (13), respectively. Let
\[
\hat{z}(t)
\]
be given by
\[
\hat{z}(t) = (1 - \rho)\hat{z}(t - 1) + \rho A^{-1}w(t)
\]
Then, assuming that $\rho < \mu$ and $\frac{\rho}{\mu} A^{-1}$ we have
\[
|z(t) - \hat{z}(t)| \leq \frac{\rho}{\mu} \| A^{-1} \| \left[ \frac{|\hat{z}(t)|}{1 - \rho} + |x(t)| \right]
\]

### 4 Tracking algorithms with a second round of averaging

Now, let us go back to the tracking case. The basic algorithm we discuss is
\[
\hat{\theta}(t) = \hat{\theta}(t - 1) + \mu P\varphi(t)(y(t) - \varphi^T(t)\hat{\theta}(t - 1))
\]
($P = I$ 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
\[
\hat{\theta}(t) = (1 - \rho)\hat{\theta}(t - 1) + \rho \hat{\theta}(t)
\]
The idea is that $\rho << \mu$, so that some real averaging takes place.

First rewrite (20) as
\[
\hat{\theta} = (I - \mu P\varphi(t)\varphi^T(t))\hat{\theta}(t - 1) + \mu P\varphi(t)y(t)
\]
For slow adaptation, small $\mu$, the term $(I - \mu P\varphi(t)\varphi^T(t))$ behaves like its time or ensemble average
\[
I - \mu PQ \sim (I - \mu P\varphi(t)\varphi^T(t))
\]
(Recall that $Q = E[\varphi(t)\varphi^T(t)]$). This is at the heart of all stochastic averaging results. See, e.g. the discussion in [6], Section 4.3. Such results are established, e.g. in [1], [4], [2], [3]. We shall in this heuristic analysis simply do the replacement (23) without further ado, and consider
\[
\hat{\theta}(t) = (I - \mu PQ)\hat{\theta}(t - 1) + \mu P\varphi(t)y(t)
\]
bearing in mind that averaging theory guarantees that
\[ \hat{\theta}(t) \approx \tilde{\theta}(t) \]

We have that, from (21)
\[ \hat{\theta}(t) = \tilde{\theta}(t) + \hat{\theta}_c(t) \]
where
\[ \tilde{\theta}(t) = (1 - \rho)\tilde{\theta}(t - 1) + \rho\tilde{\theta}(t) \]
and \( \hat{\theta}_c(t) \) will be an average of \( \hat{\theta}(t) - \tilde{\theta}(t) \), which we will neglect. From Lemma 3.1 we have that
\[ \hat{\theta}(t) \approx \hat{\theta}_*(t) \]
where, with \( A = PQ \)
\[ \hat{\theta}_*(t) = (I - \rho)\hat{\theta}_*(t) + \rho Q^{-1}\varphi(t)y(t) \quad (25) \]
The consequence thus is that the averaged estimate (21) behaves like (25). We note that (25) is independent of \( \mu \) and \( P \) in (20). It only depends on \( \rho \), the averaging factor in the second round and on \( Q \), the covariance matrix of the regression vector. We note also that (25) is a special case of (24) corresponding to \( \mu = \rho \) and \( P = Q^{-1} \).

Hence The averaged estimate (21) behaves like the estimate we would have obtained in the original algorithm with \( P = Q^{-1} \) and a gain/stepsize \( \mu = \rho \).

In this asymptotic and heuristic context this in turn is the same as using
\[ P(t) = \left[ \rho \sum_{k=1}^{t}(1 - \rho)^{t-k}\varphi(k)\varphi^T(k) \right]^{-1} \quad (26) \]
which is close to \( Q^{-1} \) for small \( \rho \).

Finally, (26) inserted into (3) with \( \mu = \rho \) is exactly the Recursive Least Squares algorithm. We have thus established the same result as in [8] and in Kushner and Yang (1993): Averaging leads to optimal accuracy for estimating time invariant parameters. (Although we have not “proved” it).

However, it also follows that in the tracking case, where the covariance matrix \( R_1 \) of \( \omega(t) \) in (2) is non-zero, we do no better than recursive least squares. The optimal tracking algorithm depends on \( R_1 \), but the second round of averaging cannot take this information into account. Recursive least squares is not necessarily better than the gradient algorithm as a parameter tracker - it all depends on how the parameters move (the covariance matrix \( R_1 \)). The beneficial effects of the second averaging round (21) are therefore not as obvious as in the constant parameter case.
5 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, since this need not be optimal in the tracking case it does not follow that the second round of averaging in the accelerated scheme has any beneficial effect on the tracking error.

References


