Adaptive Filtering with Averaging*

G. Yin†

Abstract

Adaptive filtering algorithms are considered in this work. The main effort is devoted to improve the performance of such algorithms. Two classes of algorithms are given. The first one uses averaging in the approximation sequence obtained via slowly varying gains, and the second one utilizes averages in both the approximation sequence and the observed signals. Asymptotic properties–convergence and rate of convergence are developed. Analysis to one of the algorithms is presented. It is shown that the averaging approach gives rise to asymptotically optimal performance and results in asymptotically efficient procedures.

Keywords. adaptive filtering, averaging, asymptotic optimality.

1991 Mathematics subject classification. 93E11, 93E25, 60G35, 60F05.

1 Introduction

The purpose of this work is to study two classes of stochastic recursive algorithms, which can be utilized in a wide range of applications in adaptive signal processing and many other related fields. The main effort is placed on improving the asymptotic performance of the algorithms.

The problem under consideration is to recursively update an approximating sequence to the vector $\theta \in \mathbb{R}^r$ that minimizes the estimation error of a random signal, $y \in \mathbb{R}$, from an observation vector $\varphi \in \mathbb{R}^r$. The calculations are done without knowing the statistics of y and φ, on the basis of a sequence of observations $\{(\varphi_n, y_n)\}$. Throughout the paper, we shall assume the sequence $\{(\varphi_n, y_n)\}$ to be stationary and

$$E\varphi_n\varphi_n' = R > 0,\ E\varphi_n y_n = q, \tag{1.1}$$

where $R > 0$ means that the matrix R is symmetric positive definite. it is easily seen that θ is the unique solution of the Wiener-Hopf equation $R\theta = q$.

*This research was supported in part by the National Science Foundation under Grant DMS-9224372, in part by the IMA with fund provided by the National Science Foundation, and in part by Wayne State University.

†Department of Mathematics, Wayne State University, Detroit, MI 48202.

A standard algorithm for approximating θ is of the form:

$$\theta_{n+1} = \theta_n + a_n \varphi_n (y_n - \varphi_n' \theta_n), \tag{1.2}$$

where $\{a_n\}$ is a sequence of positive scalars satisfying $\sum_n a_n = \infty$, $a_n \xrightarrow{n} 0$, and z' denotes the transpose of z.

Many algorithms for adaptive filtering, adaptive array processing, adaptive antenna systems (cf. [1] and the references therein), adaptive equalization (cf. [2]), adaptive noise cancellation (cf. [1]), pattern recognition and learning (cf. [3]) etc. have been or can be recast into the same form as (1.2), with only signal, training sequence and/or reference signals varying from applications to applications. An extensive list of references on the applications mentioned above can be found for example in [1], [4] etc. For related problems in adaptive systems, consult [5], [6] among others.

Algorithm (1.2) and its variations have been studied extensively for many years, various results of convergence and rates of convergence have emerged, and numerous successful applications have been reported (cf. [7], [8], [9], [10] and the references therein).

In contrast with these developments, the efficiency issue (asymptotic optimality) is the main focus here. Our primary concern is to design asymptotically efficient and easily implementible algorithms with asymptotically optimal convergence speed so as to improve the performance of the algorithm.

The rest of the work is arranged as follows. Discussions on asymptotic optimality is given next. The precise problem formulation is presented in Section 3. Then Section 4 is devoted to the convergence and asymptotic normality of the algorithm, from which the asymptotic optimality is obtained. A number of further remarks are made in Section 5.

2 Asymptotic optimality

It was shown in the literature that under appropriate conditions, $\theta_n \xrightarrow{n} \theta$ with probability one or weakly, and $(1/\sqrt{a_n})(\theta_n - \theta)$ converges in distribution to a normal random vector with covariance Σ. The scaling factor $\sqrt{a_n}$ together with the covariance Σ is a measure of rate of convergence.

It has been a long time effort to improve the rate of convergence and reduce the variance in the adaptive estimation problems. The investigation of obtaining asymptotic optimality can be traced back to the early 50's. As was noted in [11], this is closely linked to an optimization problem.

To review the development in this direction, we digress a little, and begin with a related problem. Consider the following one dimensional, stochastic approximation algorithm

$$x_{n+1} = x_n + \frac{\Gamma}{n}(f(x_n) + \tilde{\zeta}_n), \tag{2.1}$$

where $\{\tilde{\zeta}_n\}$ is a sequence of random disturbances, and Γ is a parameter to be specified later. Under appropriate conditions, it can be shown that $x_n \to x^0$ w.p.1 (where x^0 is such that

$f(x^0) = 0$) and $\sqrt{n}(x_n - x^0) \sim N(0, \Sigma)$ with the asymptotic variance given by

$$\Sigma = \Sigma(\Gamma) = \frac{\Gamma^2 \Sigma_0}{2\Gamma H + 1}, \tag{2.2}$$

where $H = f_x(x^0) < 0$. Eq. (2.2) reveals the fact that the asymptotic variance depends on the parameter Γ. As a function of Γ, $\Sigma(\Gamma)$ is well behaved. Minimizing Σ w.r.t. Γ leads to the choice of $\Gamma^* = -1/H$ and the optimal variance is given by $\Sigma^* = \Sigma_0/H^2$.

A first glance may make one believe that the problem is completely solved. Nevertheless, H is very unlikely to be know to start with. Therefore, much work has been devoted to design efficient algorithms in order to achieve the asymptotic optimality. One of the approaches is the adaptive stochastic approximation method. The essence of such an approach is that in lieu of Γ, a sequence of estimates $\{\Gamma_n\}$ is constructed and (2.1) is replaced by

$$x_{n+1} = x_n + \frac{\Gamma_n}{n}(f(x_n) + \tilde{\zeta}_n). \tag{2.3}$$

The emphasis is then placed on designing the algorithm such that $\Gamma_n \to -H^{-1}$ and $x_n \to x^0$. Moreover, it is desired to have that $\sqrt{n}(x_n - x^0) \sim N(0, \Sigma^*)$, where $\Sigma^* = H^{-1}\Sigma_0(H^{-1})'$.

The aforementioned approach can be adopted to treat adaptive filtering problems. In this case, the algorithm takes the form

$$\theta_{n+1} = \theta_n + \frac{\Gamma_n}{n}\varphi_n(y_n - \varphi_n'\theta_n).$$

Similar to the argument above, it can be shown that $\Gamma_n \to R^{-1}$, $\theta_n \to \theta$, and $\sqrt{n}(\theta_n - \theta) \sim N(0, \Sigma^*)$, where $\Sigma^* = R^{-1}\Sigma_0 R^{-1}$ and Σ_0 is the covariance of the signals involved. Further discussion on this matter and related problems (with the corresponding approaches in adaptive filtering like algorithms) can be found in [12] and the references therein. While this approach does give us the consistency of $\{\Gamma_n\}$ and $\{x_n\}$ or $\{\theta_n\}$, and the desired optimality, it is computationally intensive for multidimensional problems. If a multidimensional problem is encountered, a sequence of matrix-valued estimates must be constructed, i.e., the estimate of every entry of the gradient matrix or the matrix R must be obtained.

Now, coming back to algorithm (1.2), take $a_n = a/n^\gamma$, for $0 < \gamma \leq 1$ and some $a > 0$. A moment of reflection reveals that as far as the scaling factor is concerned, $\gamma = 1$ leads to the best order due to the central limit theorem. In order to implement adaptive filtering procedures, one wishes the iterates move to a neighborhood of the true parameter θ reasonably fast. Rapid decreasing sequence a_n often yields poor results in the initial phase of computation. Therefore, one might wish to choose large step size a_n, i.e., $\gamma < 1$. Nevertheless, larger step size will result in slower rate of convergence. Therefore, there seems to be a dilemma.

Very recently, some new methods were proposed and suggested for stochastic approximation methods in [13], [14] and [15]. In these new developments, arithmetic averaging is used

in an essential way. The procedures are multi-step iterative schemes. Two of the notable algorithms are

$$\begin{aligned} x_{n+1} &= x_n + a_n(f(x_n) + \tilde{\zeta}_n) \\ \bar{x}_n &= \frac{1}{n}\sum_{i=1}^{n} x_i, \end{aligned} \tag{2.4}$$

and

$$\begin{aligned} x_{n+1} &= \bar{x}_n + a_n n\bar{y}_n \\ \bar{x}_n &= \frac{1}{n}\sum_{i=1}^{n} x_i,\ \bar{y}_n = \frac{1}{n}\sum_{i=1}^{n}(f(x_i) + \tilde{\zeta}_i), \end{aligned} \tag{2.5}$$

where $\{a_n\}$ is a sequence of 'slowly' varying gain (slow with respect to $1/n$). Some amazing things happen. It turns out that for both algorithms, $\{\bar{x}_n\}$ is an asymptotically optimal convergent sequence of estimates.

Algorithm (2.4) was suggested independently in [13] and [14], respectively, whereas (2.5) was initially studied in [15] in the context of application to sequential estimation of LD_{50}, which is a measure of toxicity defined as the dose level that would produce a death rate of 50% in a given population of animals.

In treating algorithm (2.4), independent, identically distributed (i.i.d.) noise and martingale difference type of processes were considered in [13] and [14]. It was shown in [13],

$$E(\bar{x}_n - x^0)(\bar{x}_n - x^0)' = \frac{1}{n}\Sigma^* + o(1/n),$$

whereas asymptotic normality was obtained in [16]. φ-mixing type of noise was dealt with in [17]. Further extensions were provided in [18]. Algorithms with state feedback were proposed in [19]. As for (2.5), some interesting heuristic argument was given in [15]; one dimensional linear problem with i.i.d. random processes was considered in [20], whereas much more general situation was studied in [21].

The use of the averaging approach allows the iterates to get to a vicinity of θ faster, mean while, it keeps the best possible order of rate of convergence and makes the asymptotic covariance to be the optimal one. It produces a "squeezing effect" forcing the iterates get to a vicinity of θ faster without paying the price of increasing the asymptotic covariance matrix or slowing down the convergence speed.

It should be noted that one of the crucial requirements is that the step size a_n is slowly varying with respect to $1/n$. We shall return to this point in Section 5. Motivated and inspired by the approaches mentioned above, two classes of adaptive filtering algorithms will be studied in the sequel.

3 Two classes of algorithms with averaging

In this section, two classes of adaptive filtering type of algorithms with averaging are presented. Conditions needed in the subsequent study are given. For simplicity, the slowly

varying gain is taken to be of the form $a_n = 1/n^\gamma$, $1/2 < \gamma < 1$. Algorithms with more general gain sequences can be treated. For related work in stochastic approximation, we refer to [16], [18] and [19] among others. In 3.1, adaptive algorithm with averaging in the trajectories is given and in 3.2, another algorithm with averaging in both trajectories and observed signals is presented.

3.1 Algorithm I: averaging in the iterates

The following algorithm is inspired by the averaging approach suggested in [13] and [14]. The idea here is to generate a sequence of rough estimates using slowly varying gain first, and then take arithmetic averages of the resulting iterates. Consider the algorithm

$$\begin{aligned} \theta_{n+1} &= \theta_n + \frac{1}{n^\gamma}\varphi_n(y_n - \varphi_n'\theta_n),\ 1/2 < \gamma < 1 \\ \bar{\theta}_n &= \frac{1}{n}\sum_{j=1}^{n}\theta_j. \end{aligned} \tag{3.1}$$

Notice that the averaging here creates no additional burden since it can be recursively updated as

$$\bar{\theta}_{n+1} = \bar{\theta}_n - \frac{1}{n+1}\bar{\theta}_n + \frac{1}{n+1}\theta_{n+1}.$$

3.2 Algorithm II: averaging in both iterates and observations

Motivated by the work [15] (cf. also [20] and [21]), another class of adaptive filtering algorithm which uses averaging in both trajectories and observations, is suggested in this paper. In addition to the advantages mentioned at the last section, the algorithm with averaging in both iterates and signals appears to be more stable in the initial period, whereas for the algorithms studied in [13], [14], [17] and [18], the averaging normally should be carried out after the iterations have passed the transient period, i.e., in implementing the algorithm given in (3.1), one normally needs to wait for a while until the sequence $\{\theta_n\}$ has 'settled down', then to start the averaging procedure since taking averages in the first a few iterations may result in poor performance and create large errors. Apparently, to improve the initial performance of the algorithm is an important task. This leads us to consider the following algorithm

$$\begin{aligned} \theta_{n+1} &= \bar{\theta}_n + \frac{1}{n^\gamma}\sum_{i=1}^{n}\varphi_i(y_i - \varphi_i'\theta_i),\ 1/2 < \gamma < 1 \\ \bar{\theta}_n &= \frac{1}{n}\sum_{j=1}^{n}\theta_j. \end{aligned} \tag{3.2}$$

It appears that this algorithm works better in the initial computation period in that the averaging can be executed from beginning without producing large burst of errors. The reason stems from the fact that it is the averaged signal instead of the signal itself is used in the iteration, i.e., the random processes are smoothed out in this procedure and used in the iteration.

We close this section by making the following remarks regarding to the literature. The two algorithms suggested above fall into the category of multistep algorithms. Early attempts and investigations in this direction can be found in [23], where ideas from numerical analysis for improving approximation to solution of ordinary differential equations were utilized. In addition, the work of [24] and [25] are worth mentioning.

4 Convergence and rates of convergence

This section is concentrated on the asymptotic optimality issues. Algorithm II is analyzed. Section 4.1 states the main conditions and hypotheses; Section 4.2 deals with almost sure convergence and Section 4.3 to Section 4.5 are on asymptotic normality. First, a stability theorem is obtained for θ_n; then some asymptotic equivalency results are established; finally asymptotic distribution is derived.

4.1 Assumptions

The following assumptions will be used throughout.

(A1) $\{\varphi_n, y_n\}$ is a stationary sequence such that (1.1) holds. In addition,

$$E|\varphi_n|^{4+\delta} < \infty, \ E|y_n|^{4+\delta} < \infty \text{ for some } \delta > 0. \tag{4.1}$$

(A2) $\{\varphi_n\varphi_n' - R\}$ and $\{\varphi_n y_n - q\}$ are moving average sequences of order m, i.e.,

$$\begin{aligned} \varphi_n\varphi_n' - R &= \sum_{i=0}^{m} C_i\kappa_{n-i} \\ \varphi_n y_n - q &= \sum_{i=0}^{m} D_i\nu_{n-i}, \end{aligned} \tag{4.2}$$

where C_i, D_i, $i \leq m$ are matrices with appropriate dimension, and $\{\kappa_n\}$ and $\{\nu_n\}$ are stationary martingale difference sequences.

Remark: By virtue of (4.1),

$$E|\kappa_n|^{2+\tilde{\delta}} < \infty \text{ and } E|\nu_n|^{2+\tilde{\delta}} < \infty \text{ for some } \tilde{\delta} > 0.$$

Much more general conditions can be incorporated in the problem formulation. We refer to [21] for additional references. Although the assumptions stated here are not the most general one, they do allow us to give a simpler presentation. It seems to be more instructive to present the main idea without going through complicated technical details. Owing to these reasons, we choose these relatively simple conditions.

4.2 Convergence of Algorithm II

Theorem 4.1. *Suppose that the conditions (A1) and (A2) hold. Then*

$$\begin{aligned}&\sup_n |\theta_n| < \infty \text{ w.p.1, and } \sup_n |\bar{\theta}_n| < \infty \text{ w.p.1;}\\ &\theta_n \xrightarrow{n} \theta \text{ w.p.1, and } \bar{\theta}_n \xrightarrow{n} \theta \text{ w.p.1.}\end{aligned}$$

To obtain the desired convergence property, we make use of the well-known ordinary differential equation methods (cf. [7] and [8]). A comparison technique will be used and an auxiliary sequence for which the convergence is easily established will be constructed. Throughout the rest of the paper, K will denote a generic positive constant. Its values may change for different usage.

Proof: First, notice that by virtue of the local martingale convergence theorem,

$$\begin{aligned}&\sum_i \frac{1}{i^\gamma}(\varphi_i y_i - q) \text{ converges w.p.1 and}\\ &\sum_i \frac{1}{i^\gamma}(\varphi_i \varphi_i' - R) \text{ converges w.p.1.}\end{aligned} \tag{4.3}$$

Define $\xi_i = \varphi_i y_i - \varphi_i \varphi_i' \theta$ for each i. It follows from (4.3),

$$\sum_i \frac{1}{i^\gamma}\xi_i \text{ converges w.p.1.} \tag{4.4}$$

Hence by Kronecker's Lemma,

$$\frac{1}{n^\gamma}\sum_{i=1}^{n} \xi_i \xrightarrow{n} 0 \text{ w.p.1.} \tag{4.5}$$

Rewrite (3.2) as

$$\begin{aligned}\theta_{n+1} = \theta_n &+ \frac{1}{n^\gamma}\varphi_n(y_n - \varphi_n'\theta_n) + \frac{1-\gamma}{(n-1)^\gamma n}\sum_{i=1}^{n-1}\varphi_i(y_i - \varphi_i'\theta_i)\\ &+ \frac{1}{(n-1)^\gamma}\eta_n \sum_{i=1}^{n-1}\varphi_i(y_i - \varphi_i'\theta_i), \text{ for } n > 1;\\ \theta_2 &= \theta_1 + (\varphi_1 y_1 - \varphi_1\varphi_1'\theta_1),\\ \text{where} &\eta_n = O\left(\tfrac{1}{n^2}\right).\end{aligned} \tag{4.6}$$

To obtain the desired result, define an auxiliary sequence $\{u_n\}$ as follows.

$$\begin{aligned}&u_{n+1} = u_n + \frac{1}{n^\gamma}\varphi_n(y_n - \varphi_n' u_n), \text{ for } n > 1;\\ &u_1 = \theta_1, \ u_2 = \theta_2.\end{aligned} \tag{4.7}$$

The sequence $\{u_n\}$ is essentially generated by a standard adaptive filtering algorithm. By virtue of an argument as in [26] Section IV (E), $\sup_n |u_n| < \infty$ w.p.1 and $u_n \xrightarrow{n} \theta$ w.p.1.

To proceed, set $e_n = \theta_n - u_n$. Direct computation yields that

$$\begin{aligned} e_{n+1} = e_n &- \frac{1}{n^\gamma}\varphi_n\varphi_n' e_n - \frac{1-\gamma}{(n-1)^\gamma n}\sum_{i=1}^{n-1}\varphi_i\varphi_i' e_i - \frac{1}{(n-1)^\gamma}\eta_n\sum_{i=1}^{n-1}\varphi_i\varphi_i' e_i \\ &- \frac{1-\gamma}{(n-1)^\gamma n}\sum_{i=1}^{n-1}\varphi_i(y_i - \varphi_i' u_i) - \frac{1}{(n-1)^\gamma}\eta_n\sum_{i=1}^{n-1}\varphi_i(y_i - \varphi_i' u_i), \text{ for } n > 1; \\ e_1 = e_2 &= 0. \end{aligned} \tag{4.8}$$

Let

$$\pi_n = -\left(\frac{1-\gamma}{n}\sum_{i=1}^{n-1}\varphi_i(y_i - \varphi_i' u_i) + \eta_n\sum_{i=1}^{n-1}\varphi_i(y_i - \varphi_i' u_i)\right).$$

In view of the definition of $\{u_n\}$,

$$u_{n+1} = u_1 + \sum_{i=1}^{n}\frac{1}{i^\gamma}\varphi_i(y_i - \varphi_i' u_i).$$

Since $\sup_n |u_n| < \infty$ w.p.1,

$$\sum_{i=1}^{n}\frac{1}{i^\gamma}\varphi_i(y_i - \varphi_i' u_i) \text{ converges w.p.1, and } \frac{1}{n^\gamma}\sum_{i=1}^{n}\varphi_i(y_i - \varphi_i' u_i) \xrightarrow{n} 0 \text{ w.p.1}$$

by Kronecker's lemma. As a result $\pi_n \xrightarrow{n} 0$ w.p.1.

Let

$$B_{nk} = \begin{cases} \prod_{i=k+1}^{n}(I - \varphi_i\varphi_i'/i^\gamma), & k < n; \\ I, & k = n. \end{cases}$$

It follows from (4.8),

$$\begin{aligned} e_{n+1} = (\gamma-1)&\sum_{j=2}^{n}\frac{B_{nj}}{(j-1)^\gamma j}\sum_{i=1}^{j-1}\varphi\varphi_i' e_i \\ &- \sum_{j=2}^{n-1}\frac{B_{nj}}{(j-1)^\gamma}\eta_j\sum_{i=1}^{j-1}\varphi_i\varphi_i' e_i + \sum_{j=2}^{n-1}\frac{B_{nj}}{(j-1)^\gamma}\pi_j. \end{aligned}$$

Consequently, by interchanging the order of summations,

$$|e_{n+1}| \le K\sum_{i=1}^{n}\left(\sum_{j=i}^{n}\frac{|B_{nj}|}{j^{1+\gamma}}\right)|\varphi_i\varphi_i'||e_i| + K\sum_{i=1}^{n}\frac{|B_{ni}|}{i^\gamma}|\pi_i|. \tag{4.9}$$

It can be verified that

$$\sum_{j=1}^{n}\frac{|B_{nj}|}{j^\gamma} = O(1) \text{ and } \sum_{j=1}^{n}\frac{|B_{nj}|}{j^{1+\gamma}} = O(1/n).$$

Since $\pi_n \xrightarrow{n} 0$ w.p.1,

$$\tilde{\pi}_n = \sum_{i=1}^{n}\frac{|B_{ni}|}{i^\gamma}|\pi_i| \xrightarrow{n} 0 \text{ w.p.1.}$$

Applying the Gronwall's inequality to (4.9), we arrive at

$$|e_{n+1}| \leq K\tilde{\pi}_n \exp\left(\frac{1}{n}\sum_{i=1}^{n}|\varphi_i\varphi_i'|\right) \xrightarrow{n} 0 \text{ w.p.1} \tag{4.10}$$

by the boundedness of $(1/n)\sum_{i=1}^{n}|\varphi_i\varphi_i'|$. Therefore, with probability one, $\{\theta_n\}$ is bounded uniformly in n, $\lim_n \theta_n$ exists and is equal to $\lim_n u_n = \theta$. The convergence of $\{\theta_n\}$ is thus established. Finally, since $\bar{\theta}_n$ is the arithmetic average of θ_i, $i \leq n$, it is also bounded w.p.1 and $\bar{\theta}_n \xrightarrow{n} \theta$ w.p.1. □

4.3 A stability result

To carry out the analysis in the sequel, we need to make sure that a scaled sequence of the estimation error $\{\theta_n - \theta\}$ is bounded (tight) in some appropriate sense. As a preparation for further study, first an order of magnitude estimate or a stability result of $\{\theta_n\}$ is proved. In studying dynamical systems, Liapunov functions are very helpful. Let $V(\theta) = (1/2)\theta'\theta$. $V(\cdot)$ is a Liapunov function. A stability result in terms of $V(\cdot)$ will be given below.

Proposition 4.2. *Under conditions of Theorem 4.1,*

- $V(\theta_n - \theta) = O(n^{-\gamma})$ *for n sufficiently large;*
- $\{n^{\gamma/2}(\theta_n - \theta)\}$ *is tight.*

Proof: We derive the order of magnitude estimate first. The argument to follow can be applied to more general correlated signals as well.

Define $\tilde{\theta}_n = \theta_n - \theta$. Owing to Theorem 4.1, (4.6) can be rewritten as

$$\tilde{\theta}_{n+1} = \tilde{\theta}_n - \frac{1}{n^\gamma}R\tilde{\theta}_n + \frac{1}{n^\gamma}(R - \varphi_n\varphi_n')\tilde{\theta}_n + \frac{1}{n^\gamma}\xi_n + \frac{1}{n^\gamma}\zeta_n \tag{4.11}$$

where

$$\zeta_n = \frac{1-\gamma}{n}\sum_{i=1}^{n-1}\varphi_i(y_i - \varphi_i'\theta_i) + \eta_n\sum_{i=1}^{n-1}\varphi_i(y_i - \varphi_i'\theta_i),$$

and $\xi_n = \varphi_n y_n - \varphi_n\varphi_n'\theta$. From Theorem 4.1, it can be seen that $\zeta_n \xrightarrow{n} 0$ w.p.1.

To proceed, for some $\Delta > 0$, let

$$p(j,\Delta) = \max\left\{k;\ \sum_{l=j}^{k} E^{1/2}\,|\zeta_l|^2 \leq \Delta\right\}.$$

For some $M > 0$ sufficiently large, consider the partition

$$M = \tau_0 < \tau_1 = p(\tau_0,\Delta) < \tau_2 = p(\tau_1,\Delta) < \cdots < \tau_\nu = p(\tau_{\nu-1},\Delta).$$

It is now clear that

$$\sum_{l=\tau_i}^{\tau_{i+1}} E^{1/2}\,|\zeta_l|^2 \leq \Delta, \quad \text{for each } i < \nu.$$

For any n satisfying $\tau_i \le n < \tau_{i+1}$, we have

$$V(\tilde{\theta}_{n+1}) - V(\tilde{\theta}_n) = \tilde{\theta}'_n \left(-\frac{1}{n^\gamma} R\tilde{\theta}_n - \frac{1}{n^\gamma}(\varphi_n\varphi'_n - R)\tilde{\theta}_n + \frac{1}{n^\gamma}\xi_n + \frac{1}{n^\gamma}\zeta_n \right) + \rho_n + O(n^{-2\gamma})(1 + V(\tilde{\theta}_n)),$$

where $E\rho_n = O(n^{-2\gamma})$.

The technique of perturbed Liapunov function method (cf. [27] and the references therein) will be employed. Due to the fact that $(1/n)\sum_{j=1}^n \xi_j$ is the 'effective' noise process, direct adoption of the approach in [27] will not work. In the following proof, we first prove the desired result on a subsequence via a perturbed Liapunov function approach. Using the estimate on the subsequence as a bridge, the result then is established for any n large enough.

A number of perturbed Liapunov functions are introduced. These perturbations are small in magnitude and result in desired cancellations.

Define

$$V_1(\theta, n) = \sum_{j=n}^{\tau_{i+1}} E_n \frac{1}{j^\gamma}\theta'\xi_j$$

where E_n denotes conditioning on the data up to n, i.e., conditioning on the σ-algebra $\mathcal{F}_n = \sigma\{(\varphi_j, y_j);\ j \le n\}$. It can be seen that

$$\begin{aligned} &E|V_1(\theta, n)| = O(n^{-\gamma})(1 + V(\theta)) \text{ for each } \theta \\ &V_1(\tilde{\theta}_{n+1}, n+1) - V_1(\tilde{\theta}_n, n) = \tilde{\rho}_n - \frac{1}{n^\gamma}\tilde{\theta}'_n\xi_n, \end{aligned} \tag{4.12}$$

where $E\tilde{\rho}_n = O(n^{-2\gamma})$.

Define

$$\begin{aligned} V_2(\theta, n) &= \sum_{j=n}^{\tau_{i+1}} \frac{1}{j^\gamma} E_n\theta'(R - \varphi_j\varphi'_j)\theta \\ V_3(\theta, n) &= \sum_{j=n}^{\tau_{i+1}} \frac{1}{j^\gamma}\theta'\zeta_j. \end{aligned}$$

Similar as above, it can be shown that

$$\begin{aligned} &E|V_2(\theta, n)| = O(n^{-\gamma})(1 + V(\theta)) \text{ for each } \theta \\ &V_2(\tilde{\theta}_{n+1}, n+1) - V_2(\tilde{\theta}_n, n) = \hat{\rho}_n - \frac{1}{n^\gamma}\tilde{\theta}'_n(R - \varphi_n\varphi'_n)\tilde{\theta}_n, \end{aligned} \tag{4.13}$$

where $E\hat{\rho}_n = O(n^{-2\gamma})$;

$$\begin{aligned} &E|V_3(\theta, n)| = O(n^{-\gamma})(1 + V(\theta)) \text{ for each } \theta \\ &V_3(\tilde{\theta}_{n+1}, n+1) - V_3(\tilde{\theta}_n, n) = \varpi_n - \frac{1}{n^\gamma}\tilde{\theta}'_n\zeta_n, \end{aligned} \tag{4.14}$$

where $E\varpi_n = O(n^{-2\gamma})$.

Let

$$\tilde{V}(\theta, n) = V(\theta) + \sum_{j=1}^{3} V_j(\theta, n).$$

Detailed computation leads to

$$E\tilde{V}(\tilde{\theta}_{n+1}, n+1) - E\tilde{V}(\tilde{\theta}_n, n) \leq -\frac{\hat{\lambda}}{n^\gamma} V(\tilde{\theta}_n) + O(n^{-2\gamma})(1 + V(\tilde{\theta}_n)$$

for some $\hat{\lambda} > 0$. Owing to (4.12), (4.13) and (4.14),

$$E\tilde{V}(\tilde{\theta}_{n+1}, n+1) - E\tilde{V}(\tilde{\theta}_n, n) \leq -\frac{\lambda}{n^\gamma} \tilde{V}(\tilde{\theta}_n, n) + O(n^{-2\gamma})$$

for some $\lambda > 0$. It then follows

$$E\tilde{V}(\tilde{\theta}_{n+1}, n+1) \leq \Phi_{n|\tau_i - 1} E\tilde{V}(\tilde{\theta}_{\tau_i}, \tau_i) + K \sum_{j=\tau_i}^{n} \Phi_{n|j} \frac{1}{j^{2\gamma}},$$

where

$$\Phi_{n|k} = \begin{cases} \prod_{j=k+1}^{n} (1 - \lambda/j^\gamma), & k < n; \\ 1, & k = n. \end{cases}$$

By virtue of a summation by parts,

$$\sum_{j=1}^{n} \Phi_{n|j} \frac{1}{j^{2\gamma}} = O(n^{-\gamma}).$$

Thus,

$$E\tilde{V}(\tilde{\theta}_{n+1}, n+1) \leq \Phi_{n|\tau_i - 1} E\tilde{V}(\tilde{\theta}_{\tau_i}, \tau_i) + K/n^\gamma.$$

Owing to the estimates in (4.12), (4.13) and (4.14), we also have

$$EV(\tilde{\theta}_{n+1}) \leq \Phi_{n|\tau_i - 1} EV(\tilde{\theta}_{\tau_i}) + K/n^\gamma.$$

To proceed, we derive an upper bound when the subsequence $\{\tau_i\}$ is used as the iteration number. Let

$$\alpha_n = (n-1)^\gamma EV(\tilde{\theta}_n).$$

It then follows

$$\alpha_{\tau_{i+1}} \leq \left(\frac{\tau_{i+1} - 1}{\tau_i - 1} \right)^\gamma \Phi_{\tau_{i+1} - 1 | \tau_i - 1} \alpha_{\tau_i} + K.$$

Notice that

$$0 < \left(\frac{\tau_{i+1} - 1}{\tau_i - 1} \right)^\gamma \Phi_{\tau_{i+1} - 1 | \tau_i - 1} \leq c < 1,$$

for some $c \in \mathbb{R}$. Solving the difference inequality above yields

$$\alpha_{\tau_{i+1}} \leq c^{i+1} \alpha_{\tau_0} + K \frac{1 - c^{i+1}}{1 - c}.$$

As a result, $\sup_i \alpha_{\tau_i} \leq K < \infty$. It in turn implies that $EV(\tilde{\theta}_{\tau_i}) = O(\tau_i^{-\gamma})$.

Finally, passing the result from the subsequence $\{\tau_i\}$ to $\{n\}$ leads to that for any $T < \infty$, and any n satisfying $M \leq n < T$, there must exist an integer j, such that $\tau_j \leq n < \tau_{j+1}$. Similar estimates as above yield that

$$\alpha_{n+1} \leq \sup_j \alpha_{\tau_j} + K, \quad \text{and} \quad \sup_n \alpha_n < \infty.$$

To prove the second statement, notice that the Liapunov function is quadratic. For any $\varepsilon > 0$, choose $K_\varepsilon = [1/\varepsilon]$, where $[1/\varepsilon]$ denotes the largest integral part of $1/\varepsilon$. By virtue of the Markov inequality and the first part of the proposition,

$$P\left(\frac{V(\theta_n - \theta)}{n^\gamma} \geq K_\varepsilon\right) \leq \frac{EV(\theta_n - \theta)}{K_\varepsilon n^\gamma} \leq \frac{K}{K_\varepsilon} \leq K\varepsilon.$$

The tightness thus follows, and the proof of the proposition is completed. □

4.4 Asymptotic equivalency

Noticing that the desired asymptotic properties is on the sequence $\{\bar{\theta}_n\}$, first rewrite Algorithm II in an appropriate form. Since

$$\theta_{n+1} = (n+1)(\bar{\theta}_{n+1} - \bar{\theta}_n) + \bar{\theta}_n,$$

(3.2) yields that

$$\bar{\theta}_{n+1} = \bar{\theta}_n + \frac{1}{n^\gamma(n+1)} \sum_{i=1}^{n} \varphi_i(y_i - \varphi_i'\theta_i), \ 1/2 < \gamma < 1. \tag{4.15}$$

Using the definition of ξ_n and putting $\hat{\theta}_n = \bar{\theta}_n - \theta$, (4.15) can be further written as

$$\begin{aligned} \hat{\theta}_{n+1} = \hat{\theta}_n - \frac{R}{n^\gamma}\hat{\theta}_n + \frac{R}{n^\gamma(n+1)}\hat{\theta}_n \\ + \frac{1}{n^\gamma(n+1)} \sum_{i=1}^{n} (R - \varphi_i\varphi_i')\tilde{\theta}_i + \frac{1}{n^\gamma(n+1)} \sum_{i=1}^{n} \xi_i. \end{aligned} \tag{4.16}$$

Define

$$A_{nk} = \begin{cases} \prod_{i=k+1}^{n}(I - R/i^\gamma), & k < n; \\ I, & k = n. \end{cases}$$

Solution to (4.16) gives us

$$\begin{aligned} \sqrt{n}\hat{\theta}_{n+1} = \sqrt{n}A_{n0}\hat{\theta}_1 + \sqrt{n}\sum_{k=1}^{n} \frac{1}{k^\gamma(k+1)} A_{nk} R\hat{\theta}_k \\ + \sqrt{n}\sum_{k=1}^{n} \frac{1}{k^\gamma(k+1)} A_{nk} \sum_{i=1}^{k} (R - \varphi_i\varphi_i')\tilde{\theta}_i \\ + \sqrt{n}\sum_{k=1}^{n} \frac{1}{k^\gamma(k+1)} A_{nk} \sum_{i=1}^{k} \xi_i. \end{aligned} \tag{4.17}$$

Our effort in this subsection is devoted to prove that the first three terms on the right-hand side of the equality above are asymptotically unimportant, and the last term is asymptotically equivalent to $(R^{-1}/\sqrt{n})\sum_{i=1}^{n} \xi_i$. More precise statements is provided below.

Proposition 4.3. *Under the conditions of Proposition 4.2,*

$$\sqrt{n}\bar{\theta}_{n+1} = \frac{R^{-1}}{\sqrt{n}} \sum_{i=1}^{n} \xi_i + o(1),$$

where $o(1) \xrightarrow{n} 0$ *in probability.*

Proof: In what follows, we examine each of the terms in (4.17) separately.

First, since $\sup_n |\hat{\theta}_n| < \infty$ w.p.1, for some $\mu > 0$,

$$|\sqrt{n} A_{n0} \hat{\theta}_1| \leq \sqrt{n} |A_{n0}| |\hat{\theta}_1| \leq K \sqrt{n} \exp(-\mu \sum_{i=1}^{n} 1/i^{\gamma}) \xrightarrow{n} 0 \text{ w.p.1.} \tag{4.18}$$

As for the second term, by virtue of Proposition 4.2,

$$\begin{aligned} & E|\sqrt{n} \sum_{k=1}^{n} \frac{1}{k^{\gamma}(k+1)} A_{nk} R \hat{\theta}_k| \\ & \quad \leq K \sqrt{n} \sum_{k=M}^{n} \frac{1}{k^{1+\gamma}} |A_{nk}| E|\hat{\theta}_k| + \sqrt{n} \sum_{k=1}^{M-1} \frac{1}{k^{1+\gamma}} |A_{nk}| |R| E|\hat{\theta}_k| \\ & \quad \leq K n^{-\frac{1+\gamma}{2}} + K \sqrt{n} |A_{nM}| \xrightarrow{n} 0. \end{aligned} \tag{4.19}$$

Therefore, the second term also tends to 0 in probability.

To proceed, we examine the last term in (4.17). By using a partial summation,

$$\begin{aligned} \sqrt{n} \sum_{k=1}^{n} \frac{1}{k^{\gamma}} A_{nk} \frac{1}{k} \sum_{i=1}^{k} \xi_i &= \left(\sum_{k=1}^{n} \frac{1}{k^{\gamma}} A_{nk} \right) \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i \right) \\ &+ \sqrt{n} \sum_{k=1}^{n-1} \left(\sum_{i=1}^{k} \frac{1}{i^{\gamma}} A_{ni} \right) \left(\frac{1}{k} \sum_{i=1}^{k} \xi_i - \frac{1}{k+1} \sum_{i=1}^{k+1} \xi_i \right). \end{aligned} \tag{4.20}$$

Owing to the property of A_{nk},

$$A_{nk} - A_{n,k-1} = \frac{R}{k^{\gamma}} A_{nk}.$$

This then yields that

$$\sum_{k=1}^{n} \frac{1}{k^{\gamma}} A_{nk} = R^{-1}(I - A_{n0}).$$

Due to the fact that $(1/\sqrt{n}) \sum_{i=1}^{n} \xi_i$ is bounded in probability and $A_{n0} \xrightarrow{n} 0$,

$$\left(\sum_{k=1}^{n} \frac{1}{k^{\gamma}} A_{nk} \right) \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i \right) = \frac{R^{-1}}{\sqrt{n}} \sum_{i=1}^{n} \xi_i + o(1), \tag{4.21}$$

where $o(1) \xrightarrow{n} 0$ in probability. Likewise,

$$\begin{aligned} & \sqrt{n} \sum_{k=1}^{n-1} \left(\sum_{i=1}^{k} \frac{1}{i^{\gamma}} A_{ni} \right) \left(\frac{1}{k} \sum_{i=1}^{k} \xi_i - \frac{1}{k+1} \sum_{i=1}^{k+1} \xi_i \right) \\ & \quad = \sqrt{n} R^{-1} \sum_{k=1}^{n-1} (A_{nk} - A_{n0}) \left(\frac{1}{(k+1)k} \sum_{i=1}^{k} \xi_i - \frac{1}{k+1} \xi_{k+1} \right) \\ & \quad \xrightarrow{n} 0 \text{ in probability.} \end{aligned} \tag{4.22}$$

A few details are omitted.

Owing to (4.20)–(4.22) and noticing

$$\begin{aligned}&\sqrt{n}\sum_{k=1}^{n}\frac{1}{k^{\gamma}(k+1)}A_{nk}\sum_{i=1}^{k}\xi_i\\&\quad=\sqrt{n}\sum_{k=1}^{n}\frac{1}{k^{1+\gamma}}A_{nk}\sum_{i=1}^{k}\xi_i-\sqrt{n}\sum_{k=1}^{n}\frac{1}{k^{1+\gamma}(k+1)}A_{nk}\sum_{i=1}^{n}\xi_i.\end{aligned}$$

It is easily seen that the last term above goes to 0 in probability. This together with (4.20) yields

$$\begin{aligned}&\sqrt{n}\sum_{k=1}^{n}\frac{1}{k^{\gamma}(k+1)}A_{nk}\sum_{i=1}^{k}\xi_i\\&\quad=\frac{R^{-1}}{\sqrt{n}}\sum_{i=1}^{n}\xi_i+o(1),\ \text{where } o(1)\xrightarrow{n}0 \text{ in probability.}\end{aligned}\tag{4.23}$$

Finally, we come back to the next to the last term in (4.17). Using similar arguments as above, it can be shown that

$$\sqrt{n}\sum_{k=1}^{n}\frac{1}{k^{\gamma}(k+1)}A_{nk}\sum_{i=1}^{k}(R-\varphi_i\varphi_i')\tilde{\theta}_i=\frac{R^{-1}}{\sqrt{n}}\sum_{i=1}^{n}(R-\varphi_i\varphi_i')\tilde{\theta}_i+o(1),$$

where $o(1)\xrightarrow{n}0$ in probability.

Now,

$$\begin{aligned}&E|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(R-\varphi_i\varphi_i')\tilde{\theta}_i|^2\\&\quad=\frac{1}{n}\sum_{i=1}^{n}E\tilde{\theta}_i'(R-\varphi_i\varphi_i')'(R-\varphi_i\varphi_i')\tilde{\theta}_i\\&\qquad+\frac{2}{n}\sum_{i=1}^{n}\sum_{j>i}E\tilde{\theta}_i'(R-\varphi_i\varphi_i')'(R-\varphi_j\varphi_j')\tilde{\theta}_j.\end{aligned}\tag{4.24}$$

Recall the assumptions (A1) and (A2). For ease of presentation, set $m=1$. (For more general cases, the proof is the same except more complex notations are needed). The first term on the right side of the equality in (4.24) tends to 0 by the fact $\tilde{\theta}_n\xrightarrow{n}0$ w.p.1, $\sup_n|\tilde{\theta}_n|<\infty$ w.p.1 and $E|R-\varphi_n\varphi_n'|^2<\infty$.

As for the second term, notice that the signals involved are m-dependent (with $m=1$ in this case). First, when $j=i+1$, since $\tilde{\theta}_{i+1}$ is $\mathcal{F}_i$ measurable,

$$E\tilde{\theta}_i'(R-\varphi_i\varphi_i')'E_i(R-\varphi_{i+1}\varphi_{i+1}')\tilde{\theta}_{i+1}=E\tilde{\theta}_i'(R-\varphi_i\varphi_i')'C_1\kappa_i\tilde{\theta}_{i+1}.$$

For $j\geq i+2$,

$$\begin{aligned}&E\tilde{\theta}_i'(R-\varphi_i\varphi_i')'(R-\varphi_j\varphi_j')\tilde{\theta}_j\\&\quad=E\tilde{\theta}_i'(R-\varphi_i\varphi_i')'E_iE_{j-2}E_{j-1}(C_0\kappa_j+C_1\kappa_{j-1})\tilde{\theta}_j\\&\quad=E\tilde{\theta}_i'(R-\varphi_i\varphi_i')'C_1E_iE_{j-2}\kappa_{j-1}\left\{\tilde{\theta}_{j-1}+\frac{1}{(j-1)^{\gamma}}[\varphi_{j-1}(y_{j-1}-\varphi_{j-1}'\theta_{j-1})+\zeta_{j-1}]\right\}\\&\quad=E\tilde{\theta}_i'(R-\varphi_i\varphi_i')'C_1E_i\frac{\kappa_{j-1}}{(j-1)^{\gamma}}[\varphi_{j-1}(y_{j-1}-\varphi_{j-1}'\theta_{j-1})+\zeta_{j-1}].\end{aligned}$$

Thus

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n}\sum_{j>i}E\tilde{\theta}_i'(R-\varphi_i\varphi_i')'(R-\varphi_j\varphi_j')\tilde{\theta}_j\right| \\
&\quad\le \frac{K}{n}\sum_{i=1}^{n}E|\tilde{\theta}_i||R-\varphi_i\varphi_i'|E_i\left|\sum_{j>i}\frac{1}{(j-1)^\gamma}[\varphi_{j-1}(y_{j-1}-\varphi_{j-1}'\theta_{j-1})+\zeta_{j-1}]\right| \\
&\qquad +\frac{K}{n}\sum_i E|\tilde{\theta}_i||R-\varphi_i\varphi_i'||\kappa_i||\tilde{\theta}_{i+1}|.
\end{aligned} \tag{4.25}$$

Since $\sup_n|\tilde{\theta}_n|<\infty$ w.p.1, $\tilde{\theta}_n\to 0$ w.p.1, $E|R-\varphi_i\varphi_i'|^2<\infty$, and $E|\kappa_i|^2<\infty$,

$$\frac{K}{n}\sum_i E|\tilde{\theta}_i||R-\varphi_i\varphi_i'||\kappa_i||\tilde{\theta}_{i+1}|\le \xrightarrow{n}0.$$

We have already demonstrated that

$$\sum_j\frac{1}{j^\gamma}\varphi_j(y_j-\varphi_j'\theta_j)\text{ converges w.p.1}$$

in the previous section. In view of Proposition 4.2,

$$\frac{1}{n}\sum_{i=1}^{n}E|\tilde{\theta}_i||R-\varphi_i\varphi_i'|E_i\left|\sum_{j>i}\frac{1}{(j-1)^\gamma}\varphi_{j-1}(y_{j-1}-\varphi_{j-1}'\theta_{j-1})\right|\le Kn^{-\gamma/2}\xrightarrow{n}0.$$

Notice that since

$$\begin{aligned}
&\sum_i\frac{1}{i^\gamma}\varphi_i(y_i-\varphi_i'\theta_i)\text{ converges w.p.1,}\\
&\frac{1}{n^\gamma}\sum_{i=1}^{n-1}\varphi_i(y_i-\varphi_i'\theta_i)\xrightarrow{n}0\text{ w.p.1.}
\end{aligned}$$

Owing to the definition of $\{\zeta_n\}$, a very rough estimate gives us

$$\left|\sum_{j>i}\frac{1}{(j-1)^\gamma}\zeta_{j-1}\right|\le K\ln n\text{ uniform in }i,$$

and as a result, Proposition 4.2 yields

$$\frac{1}{n}\sum_{i=1}^{n}E|\tilde{\theta}_i||R-\varphi_i\varphi_i'|\left|\sum_{j>i}\frac{1}{(j-1)^\gamma}\zeta_{j-1}\right|\le Kn^{-\gamma/2}\ln n\xrightarrow{n}0.$$

Thus, the right-hand side of (4.25) goes to 0. The asymptotic equivalence is established. □

Remark: For more general random signals, (e.g., certain mixing processes etc.), the asymptotic equivalency still holds. The basic idea and proof of the proposition remain to be the same. Nevertheless, to account the stochastic effect, the idea of 'fixed θ process' (cf. [27] and the references therein), which indicates that for n large enough, the random signals evolves as though θ never changes, needs to be used.

4.5 Limiting distribution

Asymptotic distribution of a suitably scaled sequence of the estimation error $\bar{\theta}_n - \theta$ is considered in this subsection. With the preparation of previous developments, in view of the central limit theorem for martingale difference sequences, and noticing that

$$\sqrt{n}(\bar{\theta}_n - \theta) = \sqrt{n}(\bar{\theta}_{n+1} - \theta) + o(1)$$

where $o(1) \xrightarrow{n} 0$ in probability, we are now in a position to state the following result on asymptotic distribution.

Theorem 4.4. *Under the conditions of Proposition 4.3,*

$$\sqrt{n}(\bar{\theta}_n - \theta) \sim N(0, R^{-1}\Sigma_0 R^{-1}),$$

where Σ_0 is the covariance of the signals, i.e.,

$$\Sigma_0 = E\xi_1\xi_1' + \sum_{i=2}^{m} E\xi_1\xi_i' + \sum_{i=2}^{m} E\xi_i\xi_1'.$$

From the theorem, it follows that Algorithm II is asymptotically optimal in that it has the optimal rate of convergence with the best covariance possible.

5 Further discussions

Algorithm II was analyzed in this paper. Similar approach can be taken to study the asymptotic properties of Algorithm I (cf. [22]). In what follows, several issues are discussed. In Section 5.1, a few remarks are given regarding to the questions of different gain sequences, the noise processes and constrained version of the algorithms etc.; Section 5.2 is concerned with functional limit theorems; the connection between the averaging algorithms and some singularly perturbed stochastic systems is studied in Section 5.3; continuous parameter problems are treated in Section 5.4.

5.1 A few remarks

By examining the result obtained, one may wonder if the gain sequence is changed to $a_n = a/n^\gamma$ what will change in the outcome. It is certainly interesting to see if there is a contribution in the asymptotic covariance from the constant a. It turns out that the answer is negative. No matter what constant a is placed in a_n, a will be cancelled eventually in the process of averaging. Thus, we conclude that the optimality cannot be improved by placing a constant in the gain.

As was mentioned before, several extensions are possible. Only moving average type of noise processes are treated in this paper. For more general random processes, we refer to [17], [18], [19] and [21] for the stochastic approximation counter part.

Adaptive beam forming algorithms, which is an array processing of the adaptive filtering type with an additional constraint can also be treated in the light of the averaging procedures discussed in this work. Let $\theta \in \mathbb{R}^{r\times o}$, $\varphi_n, C \in \mathbb{R}^{r\times l}$, $y_n, \Phi \in \mathbb{R}^{o\times l}$. The basic problem is to find a recursive algorithm asymptotically converging to θ, the minimizer of

$$E|\theta'\varphi_n - y_n|^2$$
$$\text{subject to the constraint } \theta' C = \Phi.$$

A necessary and sufficient condition for the constraint to hold is

$$\Phi C^\dagger C = \Phi,$$

where $z^\dagger$ denotes the pseudo-inverse of z. Two types of averaging algorithms are devised as follows:

$$\begin{aligned} \theta_{n+1} &= C^{\dagger'}\Phi' + P\left(\theta_n + \frac{1}{n^\gamma}(\varphi_n y_n' - \varphi_n\varphi_n'\theta_n)\right), \quad 1/2 < \gamma < 1 \\ \bar{\theta}_n &= \frac{1}{n}\sum_{i=1}^{n}\theta_i \text{ with } \theta_1 = C^{\dagger'}\Phi', \end{aligned} \tag{5.1}$$

and

$$\theta_{n+1} = C^{\dagger'}\Phi' + P\left(\bar{\theta}_n + \frac{1}{n^\gamma}\sum_{i=1}^{n}(\varphi_i y_i' - \varphi_i\varphi_i'\theta_i)\right), \quad 1/2 < \gamma < 1, \tag{5.2}$$

where $P = I - CC^\dagger$. By considering certain vector spaces, and carrying out appropriate decompositions, these equations can further be written in a more convenient and manageable form (cf. [28] and the references therein); the asymptotic properties can then be studied.

Various projection and truncation algorithms can be designed in conjunction with the averaging approaches. Furthermore, adaptive filtering with averaging can be adopted and used in the framework of using multiprocessors and parallel processing (cf. [29] and the references therein) methods.

5.2 Functional limit theorems

The asymptotic optimality obtained in the previous section can be strengthened. Far reaching functional limit theorems can be established. In lieu of examining $\sqrt{n}(\bar{\theta}_n - \theta)$, let

$$w_n(t) = \frac{[nt]}{\sqrt{n}}(\bar{\theta}_{[nt]+1} - \theta), \text{ for } t \in [0,1]$$

where $[z]$ denotes the largest integral part of z. Then $w_n(\cdot) \in D^r[0,1]$ the space of functions defined on $[0,1]$, that are right continuous, and have left-hand limits endowed with the Skorokhod topology (cf. [30] and the references therein). The pertinent notion of convergence is in the sense of weak convergence (cf. [30]). Using similar arguments as in the previous section, it can be proved that

$$w_n(t) = \frac{R^{-1}}{\sqrt{n}}\sum_{i=1}^{[nt]}\xi_i + o(1),$$

where $o(1) \xrightarrow{n} 0$ in probability uniformly in t. A functional central limit theorem for the sequence $\frac{1}{\sqrt{n}} \sum_{i=1}^{[nt]} \xi_i$ together with the Slutsky's lemma yields that $w_n(\cdot)$ converges weakly to a Brownian motion $w(\cdot)$ with the 'optimal' covariance $t\Sigma$ where $\Sigma = R^{-1}\Sigma_0 R^{-1}$.

5.3 Singularly perturbed systems

In [18], the connection of stochastic approximation algorithms with averaging and some singularly perturbed systems are exploited. This in turn gives clear explanation on why the averaging idea works and why it is important to use slowly varying gains. It will be seen in the sequel that Algorithm II discussed in this work also has a natural connection with a singularly perturbed system.

We begin with the recursion defined by (3.2) and the equation for $\bar{u}_n = \sqrt{n}(\bar{\theta}_n - \theta)$. As in [18], put them in the same time scale, we have

$$\begin{aligned} \frac{1}{n^{1-\gamma}}(\theta_{n+1} - \theta_n) &= \frac{1}{n}(\varphi_n y_n - \varphi_n \varphi_n' \theta_n) + \frac{1}{n}\zeta_n \\ \bar{u}_{n+1} - \bar{u}_n &= -\frac{1}{2n}\bar{u}_n(1 + O(1/n)) + \frac{1}{\sqrt{n+1}}\theta_{n+1}. \end{aligned} \tag{5.3}$$

Except the extra term ζ_n, the equations above have the same form as that of [18]. Eq. (5.3) can be viewed as a multiple time scale adaptive filtering algorithm, which has a close connection to a singularly perturbed system (cf. [18] Eq. (2.8)) of the form

$$\begin{aligned} \varepsilon dz^\varepsilon &= A_{11} z^\varepsilon dt + dw_1 \\ dx^\varepsilon &= A_{22} x^\varepsilon dt + A_{12} z^\varepsilon dt + dw_2. \end{aligned}$$

In addition, notice that the requirement of slowly varying gain is crucial. For example, if $\gamma = 1$, the structure of the singularly perturbed system will be destroyed.

5.4 Continuous time analogue

In addition to the mathematical interest, the reasons for considering continuous version algorithm stem from the fact that the continuous problems are good approximation to discrete ones when the sampling is taken rather frequently. It is important to establish that everything works well if the sampling rate becomes very high.

Continuous time analog of the algorithms discussed here are given below. Corresponding to Algorithm I, we have

$$\begin{aligned} \dot{\theta}_t &= \frac{1}{t^\gamma}\varphi_t(y_t - \varphi_t'\theta_t), \quad 1/2 < \gamma < 1 \\ \bar{\theta}_t &= \frac{1}{t}\int_0^t \theta_s ds; \end{aligned} \tag{5.4}$$

corresponding to Algorithm II,

$$\theta_t = \frac{1}{t}\int_0^t \theta_s ds + \frac{1}{t^\gamma}\int_0^t \varphi_s(y_s - \varphi_s'\theta_s)ds, \ 1/2 < \gamma < 1. \tag{5.5}$$

To study the asymptotic properties, we first prove $\theta_t \to \theta$ w.p.1. Then define

$$B_t(\tau) = \tau\sqrt{t}(\bar{\theta}_{t\tau} - \theta) \text{ for each } \tau \in [0,1],$$

and

$$w_t(\tau) = \frac{1}{\sqrt{t}} \int_0^{t\tau} \xi_s ds \quad \text{for } \tau \in [0,1].$$

Then it can be shown as $t \to \infty$,

$$w_t(\cdot) \text{ converges weakly to a Brownian motion } w(\cdot) \text{ and}$$
$$B_t(\cdot) \text{ converges weakly to a Brownian motion } B(\cdot)$$

such that $B(\tau) = R^{-1}w(\tau)$. The discussion on optimality can be carried out similar to that of Section 2. In view of the limiting Brownian motion $B(\cdot)$, the continuous version of the adaptive filtering algorithms are also asymptotic optimal.

References

[1] B. Widrow and S. Stearns, *Adaptive Signal Processing*, Prentice-Hell, Englewood Cliffs, NJ, 1985.

[2] A. Gersho, Adaptive equalization of highly dispersive channels for data transmission, *Bell Syst. Tech. J.* **48** (1969), 55-70.

[3] S. Lakshmivarahan, *Learning Algorithms and Applications*, Springer Verlag, New York, 1981.

[4] G.C. Goodwin and K.S. Sin, *Adaptive Filtering Prediction and Control*, Prentice-Hall, Englewood Cliffs, NJ, 1984.

[5] K.L. Åström, *Introduction to Stochastic Control*, Academic Press, New York, 1970.

[6] P.R. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification, and Adaptive Control*,Prentice Hall, New Jersey, 1986.

[7] L. Ljung, Analysis of recursive stochastic algorithms, *IEEE Trans. Automatic Control*, **AC-22** (1977), 551-575.

[8] H.J. Kushner and D.S. Clark, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer-Verlag, 1978.

[9] E. Eweda and O. Macchi, Quadratic and almost sure convergence of unbounded stochastic approximation algorithms with correlated observations, *Ann. Institut. Henri Poincare* **19** (1983), 235-255.

[10] H.J. Kushner and A. Shwartz, Weak convergence and asymptotic properties of adaptive filters with constant gains, *IEEE Trans. Inform. Theory* **IT-30** (1984), 177-182.

[11] K.L. Chung, On a stochastic approximation method, *Ann. Math. Statist.* **25** (1954), 463-483.

[12] A. Benveniste, M. Metivier and P. Priouret, *Adaptive Algorithms and Stochastic Approximation*, Springer-Verlag, Berlin, 1990.

[13] B.T. Polyak, New method of stochastic approximation type, *Automat. Remote Control* **51** (1990), 937-946.

[14] D. Ruppert, Stochastic approximation, in *Handbook in Sequential Analysis*, B.K. Ghosh and P.K. Sen Eds., Marcel Dekker, 503-529, New York, 1991.

[15] J.A. Bather, Stochastic approximation: A generalization of the Robbins-Monro procedure, in *Proc. Fourth Prague Symp. Asymptotic Statist.*, P. Mandl and M. Hušková Eds., 1989, 13-27.

[16] B. Polyak and A. Juditsky, Acceleration of stochastic approximation by averaging, *SIAM J. Control Optim.* **30** (1992), 838-855.

[17] G. Yin, On extensions of Polyak's averaging approach to stochastic approximation, *Stochastics* **36** (1991), 245-264.

[18] H.J. Kushner and J. Yang, Stochastic approximation with averaging of the iterates: optimal asymptotic rate of convergence for general processes, Technical Report, LCDS #91-9, Brown Univ., 1991, also to appear in *SIAM J. Control Optim.*

[19] H.J. Kushner and J. Yang, Stochastic approximation with averaging and feedback: rapidly convergent "on line" algorithms, and applications to adaptive systems, Technical Report, LCDS #92-8, Brown Univ., 1992.

[20] R. Schwabe, Stability results for smoothed stochastic approximation procedures, Fachbereich Mathematik, Series A, Mathematik, Preprint Nr. A-92-14, Freie Universität Berlin, 1992.

[21] G. Yin and K. Yin, Asymptotically optimal rate of convergence of smoothed stochastic recursive algorithms, to appear in *Stochastics*.

[22] G. Yin, Asymptotic optimal rate of convergence for an adaptive estimation procedure, *Lecture Notes in Control Inform.* 184, *Stochastic Theory and Adaptive Control*, T. Duncan and B.Pasik-Duncan Eds., Springer-Verlag, 1992, 480-489.

[23] Ya. Z. Tsypkin, *Adaptation and Learning in Automatic Systems*, Academic Press, New York, 1971.

[24] S.V. Shil'man and A.I. Yastrebov, Convergence of a class of multistep stochastic adaptation algorithms, *Avtomatikha i Telemekhanika* **38** (1976), 111-118.

[25] A.P. Korostelev, Multistep procedures of stochastic optimization, *Automat. Remote Control* **43** (1982), 621-627.

[26] M. Métivier and P. Priouret, Applications of a Kushner and Clark lemma to general classes of stochastic algorithms, *IEEE Trans. Inform.* **IT-30** (1984), 140-150.

[27] H.J. Kushner, *Approximation and Weak Convergence Methods for Random Processes, with applications to Stochastic Systems Theory*, MIT Press, 1984.

[28] G. Yin, Asymptotic properties of an adaptive beam former algorithm, *IEEE Trans. Inform. Theory* **IT-35** (1989), 859-867.

[29] H.J. Kushner and G. Yin, Asymptotic properties of distributed and communicating stochastic approximation algorithms, *SIAM J. Control Optim.* **25** (1987), 1266-1290.

[30] S.N. Ethier and T.G. Kurtz, *Markov Processes, Characterization and Convergence*, Wiley, New York, 1986.