COMBINATION COEFFICIENTS FOR FASTEST CONVERGENCE OF DISTRIBUTED LMS ESTIMATION

Kevin T. Wagner

Naval Research Laboratory
Radar Division
Washington, DC 20375, USA

Miloš I. Doroslovački

The George Washington University
Department of Electrical and Computer Engineering
Washington, DC 20052, USA

ABSTRACT

Diffusion strategies for learning across networks which minimize the transient regime mean-square deviation across all nodes are presented. The problem of choosing combination coefficients which minimize the mean-square deviation at all given time instances results in a quadratic program with linear constraints. The implementation of the optimal procedure is based on the estimation of weight deviation vectors for which an algorithm is proposed. Additionally, the optimization that uses relaxed constraints is considered. The proposed methods were validated through simulations for different estimation distribution strategies and input signals. The results show a potential for significant improvement of the convergence speed.

Index Terms— Adaptive filtering, convergence, distributed algorithms, least mean square algorithms.

1. INTRODUCTION

Recently work has been performed on diffusion strategies for adaptation and learning over networks [1]. Previous efforts have focused on choosing combination coefficients such that the sum of the mean square deviations (MSD) across all nodes is minimized once steady-state has been reached. Diffusion strategies such as combine-then-adapt (CTA) and adapt-then-combine (ATC) have been shown to achieve smaller MSD in steady-state then the MSD achieved by a noncooperative network. In this work, two new approaches are proposed on how to efficiently exchange information across a network. The proposed approaches offer faster transient regime convergence as well as optimize the steady-state performance because the mean square deviation is minimized at each time instance.

2. DISTRIBUTED ESTIMATION PROBLEM FRAMEWORK

2.1. Notation

Let $\mathbf{C}$ denote an arbitrary complex-valued matrix. The conjugate, transpose, and conjugate transpose of the matrix $\mathbf{C}$ are denoted by $\mathbf{C}^*$, $\mathbf{C}^T$, and $\mathbf{C}^H$; respectively. These operators have the same meanings when employed on vectors and scalars where applicable. All signals are assumed to be complex in this presentation, unless stated otherwise.

2.2. Network Model

Assume a connected network with N nodes, where each node has an agent. The neighborhood of any particular agent, k, is denoted by $\mathcal{N}_k$ and consists of all nodes that are connected to node k. Node k is assumed to be in $\mathcal{N}_k$. Two nodes are said to be connected if they can share information directly with each other. Additionally it is assumed that if agent l is a neighbor of agent k then agent k is a neighbor of agent l. Let $\{a_{kl}(i), a_{lk}(i)\}$ represent the set of nonnegative combination coefficients between agents k and l at time i. The scalar $a_{lk}(i)$ is used by agent k to scale the estimate it receives from agent l while $a_{kl}(i)$ is used by agent l to scale estimate received from agent k. Define

$$\mathbf{a}_k(i) = [a_{1k}(i), a_{2k}(i), \ldots, a_{Nk}(i)]^T.$$

If the l^{th} and k^{th} nodes are not connected at time i then $a_{lk}(i) = a_{kl}(i) = 0$. It is assumed that $\mathbf{a}_k(i)^T\mathbf{1} = 1$ and $a_{lk} \geq 0 \; \forall \; l \in 1, 2, \ldots, N$, where $\mathbf{1}$ is the vector of all ones.

2.3. Problem Framework

Consider the situation in which the N agents are attempting to estimate the unknown impulse response vector $\mathbf{w}_o$ which has length L. Each attempts to estimate the impulse response using the least mean square (LMS) algorithm [2]. Each agent k interrogates the unknown impulse response at time i by sending an input sequence $\mathbf{x}_k(i)$ where $\mathbf{x}_k(i) = [x_k(i), x_k(i-1), \ldots, x_k(i-L+1)]^T$. Let the received response be represented by $d_k(i) = \mathbf{x}_k^H(i)\mathbf{w}_o + v_k(i)$ where $v_k(i)$ is the measurement noise. The impulse response of the system is estimated with the adaptive filter coefficient vector, $\mathbf{w}_k(i)$, which has length L also. The output of the adaptive filter is given by $y_k(i) = \mathbf{x}_k^H(i)\mathbf{w}_k(i-1)$. The error signal $e_k(i)$ is equal to difference of the measured output, $d_k(i)$ and the output of the adaptive filter $y_k(i)$. Define the weight deviation at time i for the k^{th} node as

$$\mathbf{z}_k(i) = \mathbf{w}_o - \mathbf{w}_k(i). \tag{1}$$

2.4. Model Assumptions

The input signal $\mathbf{x}_k(i)$ is a zero-mean stationary random process with covariance matrix given by $E\{\mathbf{x}_k(i)\mathbf{x}_k(i)^H\} = \mathbf{R}$ for all nodes. The noise process is white with variance $E\{v_k(i)\; v_k^*(i)\} = \sigma_{v,k}^2$, independent of the input signal at all nodes, and independent of all measurement noises at nodes $l \neq k$. The input signals are independent of each other at different nodes implying $E\{\mathbf{x}_k(i)\mathbf{x}_l^H(i)\} = 0$ for $l \neq k$. All agents employ the same step size μ in their LMS algorithms. The step size μ is sufficiently small such that the weight deviation can be considered independent of the input signal for all times and nodes.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

3. DERIVATION OF OPTIMAL COMBINATION COEFFICIENTS

Begin by defining the mean square deviation at time i for the k^{th} node as

$$\text{MSD}_k(i) = E\{\mathbf{z}_k^H(i)\mathbf{z}_k(i)\}; \quad (2)$$

respectively.

The goal is to choose combination coefficients such that the sum of the MSD across all nodes at time i is minimized. This minimization problem can be written as

$$\begin{aligned}\min_{\mathbf{A}(i)} \sum_{k=1}^{N} E\{\mathbf{z}_k(i)^H\mathbf{z}_k(i)\} &= \min_{\mathbf{a}_1(i)} E\{\mathbf{z}_1^H(i)\mathbf{z}_1(i)\} \\ &+ \ldots + \min_{\mathbf{a}_N(i)} E\{\mathbf{z}_N^H(i)\mathbf{z}_N(i)\} \quad (3)\end{aligned}$$

where

$$\mathbf{A}(i) \equiv [\mathbf{a}_1(i), \ldots, \mathbf{a}_N(i)]\,. \quad (4)$$

Hence minimizing the MSD across all nodes is equivalent to minimizing the MSD of each node individually as shown in (3). The relationship between $\mathbf{z}_k(i)$ and $\mathbf{a}_k(i)$ depends on the type of distributed algorithm used.

3.1. CTA Optimal Combination Coefficients

The CTA diffusion strategy is given by

$$\begin{aligned}\boldsymbol{\psi}_k(i-1) &= \sum_{l\in\mathcal{N}_k} a_{lk}(i)\mathbf{w}_l(i-1) \\ \mathbf{w}_k(i) &= \boldsymbol{\psi}_k(i-1) + \mu\mathbf{x}_k(i)[d_k(i) - \mathbf{x}_k^H(i)\boldsymbol{\psi}_k(i-1)]. \quad (5)\end{aligned}$$

Substituting (5) into (1) allows the weight deviation at time i and node k to be rewritten as

$$\mathbf{z}_k(i) = \left[\mathbf{I} - \mu\mathbf{x}_k(i)\mathbf{x}_k^H(i)\right]\mathbf{Z}(i-1)\mathbf{a}_k(i) - \mu\mathbf{x}_k(i)v_k(i) \quad (6)$$

where

$$\mathbf{Z}(i-1) \equiv [\mathbf{z}_1(i-1), \ldots, \mathbf{z}_N(i-1)]\,. \quad (7)$$

Next the MSD at node k and time i can be calculated as

$$\begin{aligned}E\{\mathbf{z}_k^H(i)\mathbf{z}_k(i)\} &= \mathbf{a}_k^H(i)E\{\mathbf{Z}^H(i-1)\mathbf{Q}\mathbf{Z}(i-1)\}\mathbf{a}_k(i) \\ &+ \mu^2\sigma_v^2\text{Tr}(\mathbf{R}) \quad (8)\end{aligned}$$

where

$$\mathbf{Q} = E\{\left[\mathbf{I} - \mu\mathbf{x}_k(i)\mathbf{x}_k^H(i)\right]^2\} = \mathbf{I} - 2\mu\mathbf{R} + \mu^2\mathbf{R}\text{Tr}(\mathbf{R}) + \mu^2\mathbf{R}^2$$

for complex circular input signals and

$$\mathbf{Q} = \mathbf{I} - 2\mu\mathbf{R} + \mu^2\mathbf{R}\text{Tr}(\mathbf{R}) + 2\mu^2\mathbf{R}^2$$

for real input signals.

Examining (8) it is clear that the term $\mu^2\sigma_v^2\text{Tr}(\mathbf{R})$ is not a function of $\mathbf{a}_k(i)$, therefore the minimization of the MSD of the k^{th} node can be rewritten as

$$\min_{\mathbf{a}_k(i)} \mathbf{a}_k^H(i)E\{\mathbf{Z}^H(i-1)\mathbf{Q}\mathbf{Z}(i-1)\}\mathbf{a}_k(i). \quad (9)$$

The minimization given in (9) is a quadratic program with linear constraints $\mathbf{a}_k^T(i)\mathbf{1} = 1$ and $a_{lk}(i) \geq 0\ \forall\ l \in 1, 2, \ldots, N$. Quadratic programs of this type have been studied extensively and iterative methods, such as the active set method, exist [3] [4] [5] [6].

3.2. ATC Optimal Combination Coefficients

The ATC diffusion strategy is given by

$$\begin{aligned}\boldsymbol{\psi}_k(i) &= \mathbf{w}_k(i-1) + \mu\mathbf{x}_k(i)[d_k(i) - \mathbf{x}_k^H(i)\mathbf{w}_k(i-1)] \\ \mathbf{w}_k(i) &= \sum_{l\in\mathcal{N}_k} a_{lk}(i)\boldsymbol{\psi}_l(i). \quad (10)\end{aligned}$$

Substituting (10) into (1) and defining

$$\boldsymbol{\Psi}(i) = [\boldsymbol{\psi}_1(i), \boldsymbol{\psi}_2(i), \ldots, \boldsymbol{\psi}_N(i)] \quad (11)$$

allows the weight deviation at time i and node k to be rewritten as

$$\mathbf{z}_k(i) = [\mathbf{W}_o - \boldsymbol{\Psi}(i)]\,\mathbf{a}_k(i) \text{ where } \mathbf{W}_o \equiv [\mathbf{w}_o, \ldots, \mathbf{w}_o]\,. \quad (12)$$

Defining

$$\mathbf{Y}(i) \equiv [\mathbf{y}_1(i), \mathbf{y}_2(i), \ldots, \mathbf{y}_N(i)] = [\mathbf{W}_o - \boldsymbol{\Psi}(i)]$$

allows the MSD at the k^{th} node can be written as

$$E\{\mathbf{z}_k^H(i)\mathbf{z}_k(i)\} = \mathbf{a}_k(i)^H E\{\mathbf{Y}^H(i)\mathbf{Y}(i)\}\mathbf{a}_k(i). \quad (13)$$

Hence in order to minimize the MSD at the k^{th} node the following problem needs to be solved

$$\min_{\mathbf{a}_k(i)} \mathbf{a}_k(i)^H E\{\mathbf{Y}^H(i)\mathbf{Y}(i)\}\mathbf{a}_k(i). \quad (14)$$

Again this minimization problem is a quadratic program and can be solved using the methods previously referenced.

4. ALGORITHM IMPLEMENTATIONS

In order to solve the minimization problems given in (9) and (14) the quantities $E\{\mathbf{Z}^H(i-1)\mathbf{Q}\mathbf{Z}(i-1)\}$ and $E\{\mathbf{Y}^H(i)\mathbf{Y}(i)\}$ must be estimated, respectively. We will present an approach for estimating these quantities.

4.1. Estimation Techniques

4.1.1. Performance Bounds

Let us substitute

$$E\{\mathbf{Z}^H(i-1)\mathbf{Q}\mathbf{Z}(i-1)\} \text{ with } \mathbf{Z}^H(i-1)\mathbf{Q}\mathbf{Z}(i-1)$$

and

$$E\{\mathbf{Y}^H(i)\mathbf{Y}(i)\} \text{ with } \mathbf{Y}^H(i)\mathbf{Y}(i).$$

These substitutions require knowledge of the true impulse response and therefore are not feasible. This substitution allows an upper bound on performance to be found.

4.1.2. Feasible Implementation

Estimation of the term $E\{\mathbf{Y}^H(i)\mathbf{Y}(i)\}$ found in the ATC algorithm will be presented here, however the same techniques can be extended to estimation of $E\{\mathbf{Z}^H(i-1)\mathbf{Q}\mathbf{Z}(i-1)\}$ in the CTA algorithm. Initially, time averaging was attempted to estimate the quantity $E\{\mathbf{Y}^H(i)\mathbf{Y}(i)\}$, however due to the combination step given in (10) the time-averaged estimates were poor and resulted in poor algorithm performance.

Instead the following approach was adopted. Define the a posteriori error as

$$\epsilon_k(i) \equiv d_k(i) - \mathbf{x}_k^H(i)\boldsymbol{\psi}_k(i).$$

Next the relationship

$$\begin{aligned}\sum_{i=mP}^{mP+P-1} \mathbf{x}_k(i)\epsilon_k(i) &= \sum_{i=mP}^{mP+P-1} \mathbf{x}_k(i)[d_k(i) - \mathbf{x}_k^H(i)\boldsymbol{\psi}_k(i)] \\ &= \sum_{i=mP}^{mP+P-1} \mathbf{x}_k(i)[\mathbf{x}_k^H(i)\mathbf{w}_o - \mathbf{x}_k^H(i)\boldsymbol{\psi}_k(i) + v_k(i)] \quad (15) \\ &= \sum_{i=mP}^{mP+P-1} \mathbf{x}_k(i)\mathbf{x}_k^H(i)\mathbf{y}_k(i) + \sum_{i=mP}^{mP+P-1} \mathbf{x}_k(i)v_k(i)\end{aligned}$$

can be used to find an estimate of $\mathbf{y}_k(i)$ at $i = mP + P - 1$ as

$$\hat{\mathbf{y}}_k(mP+P-1) = \left[\sum_{i=mP}^{mP+P-1} \mathbf{x}_k(i)\mathbf{x}_k^H(i)\right]^{-1} \sum_{i=mP}^{mP+P-1} \mathbf{x}_k(i)\epsilon_k(i). \quad (16)$$

It is assumed that $\mathbf{y}_k(i)$ does not change a lot within P time instants.

Motivated by (16) the algorithm proceeds as follows. At each time instant

$$\begin{aligned} i &= mP + j \\ j &\in \{0, 1, 2, \ldots, P-1\} \\ m &\in \{0\} \cup \mathbb{N} \end{aligned}$$

and each node, using $\mathbf{w}_k(mP-1)$ and $\boldsymbol{\psi}_k^B(mP+P-1)$ we calculate

$$\begin{aligned} e_k(mP+j) &= d_k(mP+j) - \mathbf{x}_k^H(mP+j)\mathbf{w}_k(mP-1) \\ \epsilon_k^B(mP+j) &= d_k(mP+j) - \mathbf{x}_k^H(mP+j)\boldsymbol{\psi}_k^B(mP+P-1). \end{aligned}$$

where the block updated $\boldsymbol{\psi}_k^B(mP+P-1)$ is formed as

$$\boldsymbol{\psi}_k^B(mP+P-1) = \mathbf{w}_k(mP-1) + \mu \sum_{i=mP}^{mP+P-1} \mathbf{x}_k(i)e_k(i). \quad (17)$$

Note that combination has not taken place yet.

At time instance $i = mP+P-1$, we calculate (16) using $\epsilon_k^B(i)$ instead of $\epsilon_k(i)$. Next form

$$\hat{\mathbf{Y}}(i) \equiv [\hat{\mathbf{y}}_1(i), \hat{\mathbf{y}}_2(i), \ldots \hat{\mathbf{y}}_N(i)]$$

and estimate $E\{\mathbf{Y}^H(i)\mathbf{Y}(i)\}$ with $\hat{\mathbf{Y}}^H(i)\hat{\mathbf{Y}}(i)$. At this point the minimization problem in (14) can be solved and the combination in (10) can be performed based on block updated $\boldsymbol{\psi}_k^B(mP+P-1)$. That is we are performing combination only after P iterations. During other time instances no combination is performed. This approach is motivated by the Block LMS algorithm [2]. The neighbors of node k at $i = mP+P-1$ provide node k with $\hat{\mathbf{y}}_l(i)$ and $\boldsymbol{\psi}_l^B(i)$ vectors, where $l \in \mathcal{N}_k$.

The block implementations can be extended to the CTA algorithm by replacing in the above expressions $\mathbf{w}$ with $\boldsymbol{\psi}$, $\boldsymbol{\psi}$ with $\mathbf{w}$, and $\hat{\mathbf{y}}$ with $\hat{\mathbf{z}}$. Now we form

$$\hat{\mathbf{Z}}(i) \equiv [\hat{\mathbf{z}}_1(i), \hat{\mathbf{z}}_2(i), \ldots \hat{\mathbf{z}}_N(i)]$$

and estimate $E\{\mathbf{Z}^H(i)\mathbf{Q}\mathbf{Z}(i)\}$ with $\hat{\mathbf{Z}}^H(i)\mathbf{Q}\hat{\mathbf{Z}}(i)$. Note that $\mathbf{Q}$ requires knowledge of the input signals covariance matrix, which can be estimated if that information is not available.

4.2. Constrained and Unconstrained Quadratic Programming Solution

In [1] it was assumed that $\mathbf{a}_k^T(i)\mathbf{1} = 1$ and that all combination coefficients were greater than or equal to zero. Solving (9) and (14) under these conditions has been named the quadratic programming optimal combination coefficients solution.

The second constraint on the combination coefficients can be removed [7]. Solving (9) and (14) allowing the coefficients to become negative has been named the unconstrained optimal combination coefficients solution. This may result in increased instability for a given step size μ in exchange for faster convergence rates.

A formal treatment of the proposed algorithms stability is not presented here. However, the solution with convex-combination coefficients will be stable so long as μ is chosen such that the individual nodes do remain stable. The solution allowing negative combination coefficients may require smaller values of μ to be chosen to ensure stability.

5. SIMULATION RESULTS

In this section the arithmetic mean of the MSD across all nodes in the network is plotted versus iteration for different diffusion strategies. The combination-coefficient methods employed consisted of the noncooperative network (NCN), relative variance [8], Hastings rule network [9], unconstrained optimal combination coefficients solution, and quadratic programming optimal combination coefficients solution. These methods were implemented within the CTA and ATC network diffusion strategies.

The following parameters were used in all simulations. The network examined consisted of three nodes where node 1 was connected to nodes 1 and 2, node 2 was connected to all nodes, and node 3 was connected to nodes 2 and 3. The value of μ was set to 0.005 across all nodes. The measurement noise had variance $\sigma_{v,k}^2 = 0.0001$ and was real-valued. The impulse response had length three and was given by $\mathbf{w}_o = [0.3787, 0.2395, -0.8940]^T$.

We considered white and colored real-valued input signals separately. The colored input signal was generated by a single pole system as follows:

$$x_k(i) = \gamma x_k(i-1) + \alpha_k(i) \quad (18)$$

where $x_k(0) = \alpha_k(0)$, $\alpha_k(i)$ is a white, real, Gaussian, stationary process with power $\sigma_{k,\alpha}^2 = 1$, and γ is a real pole. The value $\gamma = -0.9$ was used in this simulation, which implies $\sigma_x^2 = \sigma_\alpha^2/(1 - |\gamma|^2) = 5.263$.

The figures which will be presented are a valid form of comparison. In each node, the algorithms are the same and only the combination coefficients are varied. The effects on transient and steady-state convergence properties, of the methodology used to choose the combination coefficients, are shown.

In Figure 1 the arithmetic mean of the MSD across all nodes is plotted when using the CTA distributed estimation strategy for the ideal case. The ideal case refers to using the non-feasible implementation discussed in section 4.1.1 as well as substituting the known measurement noise variances into the relative variance and Hastings rule algorithms. The unconstrained optimal combination coefficients solution algorithm provides the fastest convergence as well as the lowest steady-state value, followed by the quadratic programming optimal combination coefficients solution algorithm. The relative variance and Hastings rule algorithms provide lower steady-state values than the NCN. The proposed algorithms exhibit faster tran-

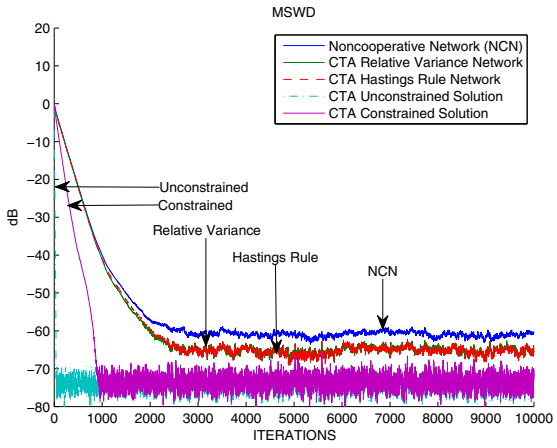

Fig. 1. CTA Distributed Estimation Strategy Learning Curve Comparison with Colored Input (Ideal Case)

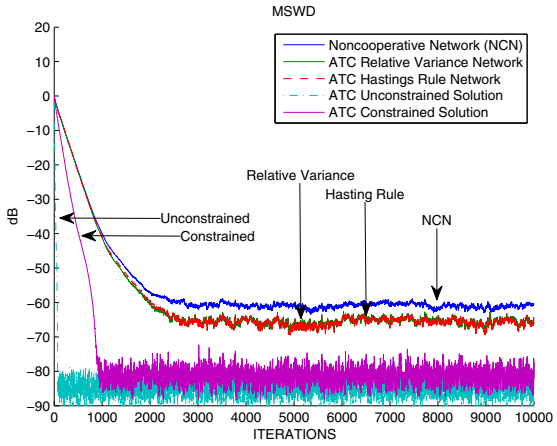

Fig. 2. ATC Distributed Estimation Strategy Learning Curve Comparison with Colored Input (Ideal Case)

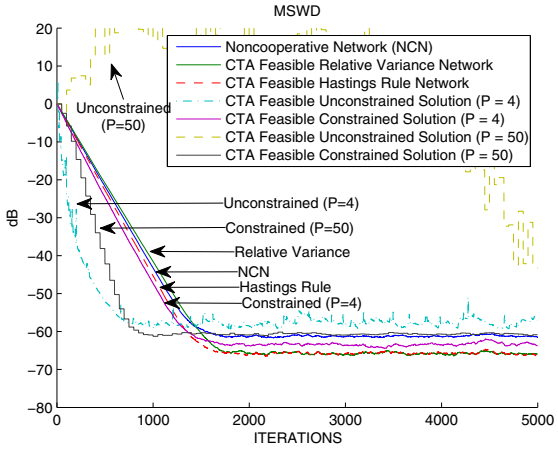

Fig. 3. CTA Distributed Estimation Strategy Learning Curve Comparison with White Input (Block Implementation)

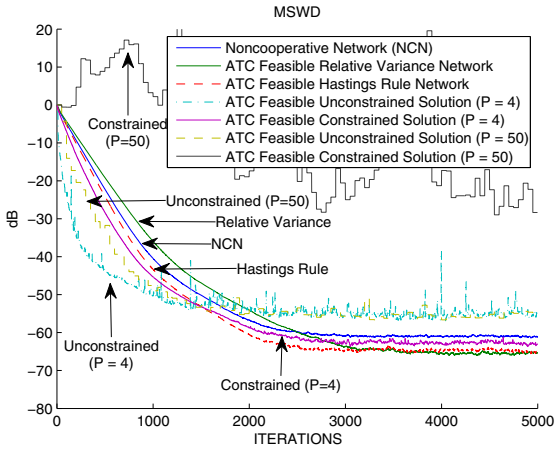

Fig. 4. ATC Distributed Estimation Strategy Learning Curve Comparison with Colored Input (Block Implementation)

sient convergence performance because they seek to minimize the mean square deviation at every time instance. The relative variance and Hastings rule algorithms assigns combination weights based on the estimated noise and therefore do not have superior transient regime performance relative to the proposed algorithms.

Similar performance trends for the ideal case can be seen with the ATC distributed estimation strategy in Figure 2. The ATC distribution strategy results in lower steady-state values than the CTA implementation.

In Figure 3 the learning curve for the CTA distributed estimation strategy using block implementation with a white input signal is depicted. A total of 100 Monte Carlo trials were used to generate this figure. The unconstrained optimal combination coefficients solution with block size of $P = 4$ has the best transient performance followed by the quadratic programming optimal combination coefficients solution with block size of $P = 50$. The unconstrained solution with $P = 50$ becomes unstable. Similar performance is observed for the ATC distributed estimation strategy, but is not presented in this paper.

Finally, in Figure 4 the learning curve for the ATC distributed estimation strategy using block implementation with the color input signal is shown. The unconstrained optimal combination coefficients with block size of $P = 4$ has the best transient regime performance. For block size $P = 50$ the algorithms convergence performance are degraded.

6. CONCLUSION

In this work methods for reducing the transient MSD across all nodes where presented within the context of the CTA and ATC distributed estimation strategies. The first method allowed the combination coefficients to have negative values while the second method presented restricted the combination coefficient values to the interval $[0, 1]$. In both cases the sum of the combination coefficients is one. Estimation techniques were presented in order to implement the proposed algorithms. They required adapting more frequently than combining to form estimates of required quantities such as $E\{\mathbf{Y}^H(i)\mathbf{Y}(i)\}$ and $E\{\mathbf{Z}^H(i)\mathbf{QZ}(i)\}$. Simulations demonstrate the proposed algorithms result in faster or equal transient regime convergence rate relative to existing algorithms.

7. REFERENCES

[1] A.H. Sayed, Sheng-Yuan Tu, Jianshu Chen, Xiaochuan Zhao, and Z.J. Towfic, "Diffusion strategies for adaptation and learning over networks: an examination of distributed strategies and network behavior," *Signal Processing Magazine, IEEE*, vol. 30, no. 3, pp. 155–171, 2013.

[2] S. Haykin, *Adaptive Filter Theory*, Prentice Hall, Upper Saddle River, New Jersey, fourth edition, 2002.

[3] D. Luenberger, *Linear and Nonlinear Programming*, Addison-Wesley, Reading, Massachusetts, second edition, 1984.

[4] R. Fletcher, *Practical Methods of Optimization*, vol. 2: Constrained optimization, John Wiley and Sons, New York, NY, 1981.

[5] G. Strang, *Introduction to Applied Mathematics*, Wellesley-Cambridge Press, Wellesley, MA, 1961.

[6] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2004.

[7] N. Takahashi, I. Yamada, and A.H. Sayed, "Diffusion least-mean squares with adaptive combiners: Formulation and performance analysis," *Signal Processing, IEEE Transactions on*, vol. 58, no. 9, pp. 4795–4810, Sept 2010.

[8] Xiaochuan Zhao, Sheng-Yuan Tu, and A.H. Sayed, "Diffusion adaptation over networks under imperfect information exchange and non-stationary data," *Signal Processing, IEEE Transactions on*, vol. 60, no. 7, pp. 3460–3475, 2012.

[9] Xiaochuan Zhao and A.H. Sayed, "Performance limits for distributed estimation over lms adaptive networks," *Signal Processing, IEEE Transactions on*, vol. 60, no. 10, pp. 5107–5124, 2012.