DIFFUSION LMS OVER MULTITASK NETWORKS WITH NOISY LINKS

Roula Nassif(1), Cédric Richard (1), Jie Chen (2), André Ferrari (1), Ali H. Sayed (3)

(1) Université de Nice Sophia-Antipolis, France
(2) Northwestern Politechnical University, Xi’an, China
(3) University of California, Los Angeles, USA

ABSTRACT

Diffusion LMS is an efficient strategy for solving distributed optimization problems with cooperating agents. In some applications, the optimum parameter vectors may not be the same for all agents. Moreover, agents usually exchange information through noisy communication links. In this work, we analyze the theoretical performance of the single-task diffusion LMS when it is run, intentionally or unintentionally, in a multitask environment in the presence of noisy links. To reduce the impact of these nuisance factors, we introduce an improved strategy that allows the agents to promote or reduce exchanges of information with their neighbors.

1. INTRODUCTION

Distributed optimization allows to address inference problems in a decentralized manner over networks, where nodes are allowed to exchange information with their neighbors to improve their local estimates. In single-task networks, all nodes are interested in estimating the same parameter vector. Among the existing cooperation rules for single-task problems, we are interested in diffusion strategies [1–4] since they are scalable, robust, and enable continuous learning.

In multitask networks, nodes are grouped into clusters and each cluster is interested in estimating its own parameter vector, that is, each cluster has its own task. Recent studies on diffusion strategies over multitask networks have focused on two scenarios. In a first scenario, it is assumed that nodes know which cluster they belong to, and multitask diffusion strategies were derived to exploit intra-cluster and inter-cluster information exchanges in a meaningful way [5–9]. In a second scenario, nodes do not know the cluster they belong to. Several research efforts have focused on analyzing the performance of diffusion strategies when they are run, intentionally or unintentionally, in a multitask environment. It is shown in [10], for example, that the diffusion iterates converge to a Pareto optimal solution when the optimization problem consists of a sum of individual costs with possibly different minimizers. It is further shown in [11] that, when the tasks are sufficiently similar to each other, the single-task diffusion LMS can still perform better than noncooperative strategies. To avoid poor results resulting from cooperation between neighbors with sufficiently different objectives, extended diffusion strategies with a clustering step are proposed in [11–14] to enable agents to identify which neighbors belong to the same cluster and which neighbors should be ignored.

Usually, the exchange of raw data and local estimates between nodes may be corrupted by noisy communication links. Useful results dealing with the consequences of noisy communications on diffusion LMS behavior are presented in [15–18] for single-task environments. In this paper, we are interested in studying the degradation in the mean and mean-square error performance that would result from running the same diffusion LMS algorithm over a multitask network in the presence of noisy communication links. The analytical results reveal the influence of each nuisance factor on the dynamics of the network, on the biases in the weight estimates, and on the mean-square error performance. Since the mean-square error depends on the combination coefficients, we also show how these coefficients can be adjusted efficiently during the learning process in order to enable agents to cooperate only with neighbors sharing the same objective, and to simultaneously reduce the effect of exchanging information through noisy links.

Notation. Normal font letters denote scalars. Boldface lowercase letters denote matrices. The operator \( \otimes \) refers to the Kronecker product and \( \col{\cdot} \) stacks the column vectors entries on top of each other. The set \( \mathcal{N}_k \) denotes the neighbors of node \( k \), \( \mathcal{C}(k) \) denotes the cluster to which node \( k \) belongs, and \( C_i \) is the \( i \)-th cluster.

2. DIFFUSION LMS IN THE PRESENCE OF NOISY LINKS

Consider a connected network of \( N \) nodes. At each time instant \( i \), node \( k \) collects a zero-mean scalar measurement \( d_k(i) \) and a zero-mean \( L \times 1 \) regression vector \( x_k(i) \) with a positive covariance matrix denoted by \( R_{x_k} = \mathbb{E} x_k(i) x_k^\top(i) \). These data are assumed to be related to an \( L \times 1 \) unknown vector \( w_k^i \) via the linear model:

\[
d_k(i) = x_k^\top(i) w_k^i + z_k(i),
\]

where \( z_k(i) \) is a zero-mean measurement noise of variance \( \sigma^2 z_k \). The noise process is assumed to be temporally white and spatially independent. The problem is to estimate \( w_k^i \) at each node \( k \). To solve this problem, node \( k \) can minimize the mean-square error \( J_k(w) \):

\[
J_k(w) = \mathbb{E} |d_k(i) - x_k^\top(i) w|^2,
\]

using a stochastic gradient algorithm of the LMS type. In this case, the performance at node \( k \) depends on the variance \( \sigma^2 z_k \) [3].

In a single-task environment, all nodes are interested in estimating the same parameter vector \( w^o \), i.e., \( w_k^i = w^o \) \( \forall k \). In this case, it was shown that the use of a diffusion LMS strategy for minimizing, in a fully-distributed manner, the following aggregate cost [1–3]:

\[
J^{arb}(w) = \sum_{k=1}^{N} \mathbb{E} |d_k(i) - x_k^\top(i) w|^2,
\]

improves the estimation accuracy. In this work, we consider the adapt-then-combine (ATC) form of diffusion LMS [1,2]:

\[
\psi_k(i+1) = w_k(i) + \mu_k \sum_{\ell \in \mathcal{N}_k} c_{k,\ell} x_{\ell}(i) [d_{\ell}(i) - x_{\ell}^\top(i) w_k(i)]
\]

\[
w_k(i+1) = \sum_{\ell \in \mathcal{N}_k} a_{k,\ell} \psi_{\ell}(i+1)
\]
where $\mu_k$ is a small positive step-size parameter at node $k$ and $w_k(i)$ is the estimate of $w^o$ at node $k$ and iteration $i$. The non-negative coefficients $c_{\ell k}$ and $\sigma_{\ell k}$, which are used to scale the data $\{x_{\ell k}(i), d_{\ell k}(i)\}$ and the intermediate estimates $\psi_{\ell k}(i+1)$ transmitted from node $\ell$ to node $k$, are zero if node $\ell$ is not connected to node $k$, that is, $\ell \notin N_k$. These coefficients are the $(\ell, k)$-th entries of a right-stochastic matrix $C$ and a left-stochastic matrix $A$, respectively.

Each step of the ATC algorithm (4)-(5) involves the transmission of information from node $\ell \in N_k$ to node $k$. In the presence of noisy communication links, the ATC diffusion LMS algorithm becomes:

$$
\psi_{k}(i+1) = w_k(i) + \mu_k \sum_{\ell \in N_k} c_{\ell k} x_{\ell k}(i) [d_{\ell k}(i) - x_{\ell k}^T(i) w_k(i)], \quad (6)
$$

$$
w_k(i+1) = \sum_{\ell \in N_k} a_{\ell k} \psi_{\ell k}(i+1). \quad (7)
$$

where $x_{\ell k}(i)$, $d_{\ell k}(i)$, and $\psi_{\ell k}(i+1)$ are the noisy data received by node $k$ from its neighboring node $\ell$. For modeling noisy communication links, we adopt the model proposed in [3, 18]

$$
d_{\ell k}(i) = d_{\ell}(i) + z_{\ell,\ell k}(i), \quad (8)
$$

$$
x_{\ell k}(i) = x_{\ell}(i) + z_{x,\ell k}(i), \quad (9)
$$

$$
\psi_{\ell k}(i) = \psi_{\ell}(i) + z_{\psi,\ell k}(i), \quad (10)
$$

where $z_{\ell,\ell k}(i)$ is a scalar noise signal, $z_{x,\ell k}(i)$ and $z_{\psi,\ell k}(i)$ are noise vectors of dimension $L \times 1$. Note that this model is more general than in [15, 16] where diffusion LMS is considered without exchange of gradient information, that is, $C = I_N$.

In a multitask environment, the local costs $J_k(w)$ are not all minimized at the same location. It is shown in [10] that, in this case, when $C = I_N$, the ATC algorithm (4)-(5) leads to a Pareto optimum solution for (3). In [11], the authors study the behavior of the ATC algorithm (4)-(5) when it is run over a multitask environment, and analyze the critical role of the distance between tasks, $w_k$. In this work, we extend [11] to the case of noisy communication links. Before proceeding, we introduce the following assumptions:

**Assumption 1.** The regressors $x_{\ell}(i)$ arise from a zero-mean random process that is temporally white and spatially independent.

**Assumption 2.** The noise vectors $z_{\ell,\ell k}(i)$, $z_{x,\ell k}(i)$, and $z_{\psi,\ell k}(i)$ are temporally white, spatially independent zero-mean random variables. We denote by $\sigma^2_{\ell,\ell k}$, $R_{\ell,\ell k}$ and $R_{\psi,\ell}$ their variance and covariance matrices, respectively.

**Assumption 3.** $\{z_{d,\ell}(i)\}$, $\{x_{\ell,\ell k}(i)\}$, $\{z_{x,\ell k}(i)\}$, $\{x_{\ell}(i)\}$, and $\{z_{\psi,\ell}(i)\}$ are mutually independent for all $\{k, \ell, m, n, p, q, s, t\}$ and $\{i_1, i_2, i_3, i_4\}$.

**Assumption 4.** The step-sizes $\mu_k$ are sufficiently small so that terms depending on higher order powers of the step-sizes can be ignored.

### 3. PERFORMANCE ANALYSIS

Using model (1), the noisy data $\{d_{\ell k}(i), x_{\ell k}(i)\}$ in (8)-(9) at node $k$ can be related to the unknown vector $w^o_\ell$ at node $\ell$ via the relation:

$$
d_{\ell k}(i) = x_{\ell k}^T(i) w^o_\ell + z_{\ell k}(i), \quad (11)
$$

where we introduce the scalar zero-mean noise signal:

$$
z_{\ell k}(i) = z_{\ell}(i) + z_{\ell,\ell k}(i) - z_{x,\ell k}(i)^T w^o_\ell, \quad (12)
$$

whose variance is:

$$
\sigma^2_{\ell k} = \sigma^2_{\ell,\ell k} + \sigma^2_{z,\ell k} + (w^o_\ell)^T R_{x,\ell k} w^o_\ell. \quad (13)
$$

Let $\tilde{w}_k(i) = w^o_k - w_{\ell k}(i)$ be the error vector at node $k$ and time instant $i$. Using (11), the estimation error that appears in the adaptation step (6) can be written as:

$$
d_{\ell k}(i) - x_{\ell k}^T(i) w_k(i) = x_{\ell k}^T(i) \tilde{w}_k(i) + x_{\ell k}^T(i) w_{\ell k}(i) + z_{\ell k}(i), \quad (14)
$$

where $w^o_k \triangleq w^o_k - w_{\ell k}(i)$. Let $\tilde{w}(i)$ and $w^o$ denote the network block error vector and the network block optimum vector, namely:

$$
\tilde{w}(i) \triangleq \text{col}\{\tilde{w}_k(i)\}_{k=1}^N, \quad w^o \triangleq \text{col}\{w^o_k\}_{k=1}^N. \quad (15)
$$

Using relation (14), the network error vector recursion for the diffusion strategy (6)-(7) can be written as:

$$
\tilde{w}(i+1) = B(i) \tilde{w}(i) - g(i) - r(i) - z_{\psi}(i+1), \quad (16)
$$

where

$$
B(i) = A^T(I_{LN} - M R(i)), \quad (17)
$$

$$
M = \text{diag}\{\mu_k I_{L_\ell}\}_{k=1}^N, \quad (18)
$$

$$
R(i) = \text{diag}\{\sum_{\ell \in N_k} c_{\ell k} x_{\ell k}(i) x_{\ell k}^T(i)\}_{k=1}^N, \quad (19)
$$

$$
g(i) = A^T M s(i), \quad (20)
$$

$$
r(i) = A^T M h(i) - (I_{LN} - A^T) w^o, \quad (21)
$$

$$
h(i) = \text{col}\{\sum_{\ell \in N_k} c_{\ell k} x_{\ell k}(i) x_{\ell k}^T(i) w^o_\ell\}_{k=1}^N, \quad (22)
$$

$$
s(i) = \text{col}\{\sum_{\ell \in N_k} c_{\ell k} x_{\ell k}(i) z_{\ell k}(i)\}_{k=1}^N, \quad (23)
$$

$$
z_{\psi}(i+1) = \text{col}\{\sum_{\ell \in N_k} a_{\ell k} z_{\psi,\ell k}(i+1)\}_{k=1}^N. \quad (24)
$$

Based on recursion (16), we examine the performance of the ATC algorithm (6)-(7) in the mean and mean-square-error sense. Due to space limitations, we only list the main results of the analysis and omit the proofs.

### 3.1. Mean behavior analysis

Taking the expectation of both sides of (16), we get:

$$
E \tilde{w}(i+1) = B E \tilde{w}(i) - g - r, \quad (25)
$$

where

$$
B = A^T (I_{LN} - M R), \quad (26)
$$

$$
g = A^T M s, \quad (27)
$$

$$
r = A^T M h - (I_{LN} - A^T) w^o. \quad (28)
$$

$$
R = \text{diag}\{\sum_{\ell \in N_k} c_{\ell k} (R_{s,\ell} + R_{x,x,\ell k})\}_{k=1}^N, \quad (29)
$$

$$
h = \text{col}\{\sum_{\ell \in N_k} c_{\ell k} (R_{s,\ell} + R_{x,x,\ell k}) w^o_\ell\}_{k=1}^N, \quad (30)
$$

$$
s = - R_{x,x} w^o. \quad (31)
$$

and $R_{x,x}$ is the $N \times N$ block matrix whose $(i,j)$-th block is $c_{\ell k} R_{x,x,\ell k}$. It can be verified that for any initial conditions, the diffusion LMS algorithm (6)-(7) converges in the mean if the step-sizes $\mu_k$ satisfy:

$$
0 < \mu_k < \frac{1}{\lambda_{\text{max}}} \left\{\sum_{\ell \in N_k} c_{\ell k} (R_{s,\ell} + R_{x,x,\ell k})\right\} \quad (32)
$$

for $k = 1, \ldots, N$, where $\lambda_{\text{max}} \{\cdot\}$ is the maximum eigenvalue of its matrix argument. The asymptotic mean bias is given by:

$$
b = \lim_{i \to \infty} E \tilde{w}(i) = -(I_{LN} - B)^{-1} (g + r). \quad (33)
Note that $g$ in (27) is zero if the regressors are not corrupted by noise during their transmission. The vector $r$ in (28) is zero if there is no cooperation between neighbors with different objectives. Finally, we observe from (32) that the noise corrupting the communication of regressors affects the stability condition.

3.2. Mean-square-error behavior analysis

We now study the behavior of the variance $E\|\tilde{w}(i+1)\|_2^2$, where $\Sigma$ is a semi-positive definite matrix that we are free to choose. Let $\sigma$ denote the vectorized version of $\Sigma$, i.e., $\sigma^T \equiv \text{vec}(\Sigma)$. We obtain from (16) the following equation:

$$E\|\tilde{w}(i+1)\|_2^2 = E\|\tilde{w}(i)\|_2^2 + [\text{vec}(T^T) - 2(BE\tilde{w}(i)) \otimes (g + r)]^T \sigma,$$

(34)

where we use the notation $\|x\|_2^2$ and $\|x\|^2_2$ interchangeably to denote the same quantity $x^T \Sigma x$. The terms in (34) are given by:

$$\mathcal{F} = E\mathbf{B}^T(i) \otimes \mathbf{B}^T(i) \approx \mathbf{B}^T \otimes \mathbf{B}^T,$$

$$C = C \otimes I_L,$$

$$T = G + R_x + R_e + 2G_e,$$

$$G = A^T \mathcal{M} \mathcal{C} \mathcal{M} \mathcal{A},$$

$$S = \text{diag} \{\sigma_{\mathcal{A}}^2, R_{\mathcal{A}} \mathcal{C} \mathcal{M} \mathcal{A}\},$$

$$R_e = \mathbf{r}^T \mathbf{A} \mathbf{C} \mathcal{M} \mathcal{A},$$

$$G_e = \mathbf{g}^T - \mathbf{A} \mathbf{C} \mathcal{M} \mathcal{A},$$

$$\mathbf{R}_+ = \text{diag} \left\{ \sum_{k} \alpha^2_{ik} \mathbf{R}_{\mathcal{A} \mathcal{C} \mathcal{M} \mathcal{A}} \right\}^N_{k=1},$$

and

$$H_k = \sum_{\ell \in N_k} c^2_{ik} \left( R_{\mathcal{A} \mathcal{C} \mathcal{M} \mathcal{A}} (u_{\mathcal{A}}^2_{ik})^T R_{\mathcal{A} \mathcal{C} \mathcal{M} \mathcal{A}} (u_{\mathcal{A}}^2_{ik}) + (u_{\mathcal{A}}^2_{ik})^T R_{\mathcal{A} \mathcal{C} \mathcal{M} \mathcal{A}} (u_{\mathcal{A}}^2_{ik}) R_{\mathcal{A} \mathcal{C} \mathcal{M} \mathcal{A}} + (u_{\mathcal{A}}^2_{ik})^T R_{\mathcal{A} \mathcal{C} \mathcal{M} \mathcal{A}} (u_{\mathcal{A}}^2_{ik}) R_{\mathcal{A} \mathcal{C} \mathcal{M} \mathcal{A}} \right).$$

The approximations in (34)-(35) follow from Assumption 4. Furthermore, the evaluation of some terms of $T$ in (37) requires the calculation of 4-th order moments that are approximated by products of 2-nd order moments. Under these approximations, and for sufficiently small step-sizes, it can be verified that for any initial conditions, the diffusion LMS algorithm (6)-(7) is mean-square stable if the error recursion (16) is mean stable and the matrix $\mathcal{F}$ is stable.

From equation (34), we obtain a recursion that enables us to evaluate the variance over time [5]:

$$E\|\tilde{w}(i+1)\|_2^2 = E\|\tilde{w}(i)\|_2^2 - E\|\tilde{w}(0)\|_2^2 + [\text{vec}(T^T) - 2(BE\tilde{w}(i) \otimes (g + r))]^T \sigma,$$

(34)

where

$$\Gamma(i) = \Gamma(i-1) + [B \text{vec}(\tilde{w}(i-1)) \otimes (g + r)](I - \Sigma_{\mathcal{A} \mathcal{C} \mathcal{M} \mathcal{A}}),$$

and $\Gamma(0) = 0_{1 \times (LN)^2}$. Once convergence is achieved, we obtain in steady state:

$$\lim_{i \to \infty} E\{\|\tilde{w}(i)\|_2^2\} = [\text{vec}(T^T) - 2(Bb) \otimes (g + r)]^T \sigma.$$
The objectives were uniformly distributed on a circle of radius $r = 0.2$. The performance decreases when the level of noise over links increases.

In the following, we use $A(0) = A_0$ and $C(0) = C_0$. The coefficients $c_{\ell k}(i)$ were set such that $C(i + 1) = A^T(i) C$. Three different protocols for adjusting the combination coefficients $a_{\ell k}$ were considered: the rule (46)-(49) with $v_k = 0.05$ and $\epsilon = 0.01$, the rule in [11] with $\epsilon = 0.01$, and the rule in [12, 18] with $v_k = 0.05$. We ran algorithm (6)-(7) for $r = 0.02$ and $\sigma_k^2 = 10^{-2}$ with the adaptive combination rules mentioned earlier. Figure 3 illustrates the network MSD behavior for these algorithms. It appears that all these rules allow us to reduce the negative effects of noise over communication links. Our rule (46)-(49) achieves the best performance.

To test the clustering ability of the ATC algorithm (6)-(7) with adaptive combiners in the presence of noisy links, we increased the distance between tasks by setting $r = 1$. In Fig. 4, we compare the network MSD of the algorithm under perfect (left plot) and imperfect (right plot) information exchange, by setting $\sigma_k^2 = 0$ and $\sigma_k^2 = 10^{-4}$, respectively. In each case, we considered fixed combiners $\{a_{\ell k}, c_{\ell k}\}$ and adaptive combiners using the 3 different protocols mentioned earlier. As shown by the experiments, the use of adaptive combiners is necessary when the tasks are not close enough. Furthermore, our rule (46)-(49) provides the best performance especially in the presence of noisy information exchange. To better analyze this behavior, we report in Fig. 5 the probabilities of erroneous clustering decisions of types I and II. Consider the link $L_{\ell k}$ connecting $k$ to its neighbor $\ell$. The probability of type I for node $k$ is the probability that $L_{\ell k}$ is erroneously dropped while $w_k^o = w_\ell^o$. The probability of type II is the probability that $L_{\ell k}$ is erroneously connected while $w_k^o \neq w_\ell^o$. We considered that the link is dropped off if $a_{\ell k}(i) < 0.05$. The experiments show that the rule in [12, 18] suffers in the presence of imperfect information exchange. The rule in [11] tends to drop off links between agents of the same clusters, notably in the presence of noisy links. Our rule (46)-(49) is able to perform a perfect clustering in the presence and absence of noisy links since both types of probabilities are decaying to zero.

6. CONCLUSION

This work analyzed the performance of the diffusion LMS when it is run in a multitask environment in the presence of noisy links. An online strategy for adapting the combination coefficients was proposed to reduce the impact of these nuisance factors.

Fig. 4. Network MSD for different combination rules (distant tasks) with perfect (left) and imperfect (right) information exchange.

Fig. 5. Erroneous clustering decisions of type I (left) and II (right) with perfect (solid) and imperfect (dashed) information exchange.
7. REFERENCES


