Distributed learning with convex sum-of-non-convex objective

Mengfei Zhang, Jie Chen, Cédric Richard

DISTRIBUTED LEARNING WITH CONVEX SUM-OF-NON-CONVEX OBJECTIVE

Mengfei Zhang[*†], *Jie Chen*[*], *Cédric Richard*[†]

[*] Centre of Intelligent Acoustics and Immersive Communications
School of Marine Science and Technology, Northwestern Polytechnical University, China
[†] Université Côte d'Azur, CNRS, OCA, France
zhangmengfei@mail.nwpu.edu.cn, dr.jie.chen@ieee.org, cedric.richard@unice.fr

ABSTRACT

Recent research works have shown that some classical optimization methods originally designed for dealing with convex problems can demonstrate similar properties when applied to non-convex scenarios, where the problems are both locally and globally non-convex. This has led to the widespread development of distributed strategies for non-convex problems. When the sum of the local non-convex costs remains (strongly) convex and the individual local costs are smooth, it indicates a specific scenario that has notable applications and can benefit from existing solving methods. In this paper, drawing inspiration from the efficiency and stability of diffusion adaptation, we explore the minimization of a strongly convex sum of non-convex local costs. Specifically, we provide an analysis to demonstrate the convergence behavior of the network. Simulations are conducted to validate our theory.

***Index Terms*—** Stochastic optimization, non-convex cost, diffusion adaptation.

1. INTRODUCTION

Distributed adaptation over networks addresses global stochastic optimization problems by estimating and tracking the parameters of interconnected nodes from streaming data, without prior knowledge of its statistical properties. This is a collaborative strategy where each node shares its local estimates with all its neighbors to optimize a global cost. In comparison to alternative approaches [1–4], diffusion adaptation has gained popularity due to superior performance and a wider stability range [4–7].

Nonetheless, a fundamental assumption of these works is the convexity of the local cost at each node, which limits their application, particularly in handling complex optimization problems such as dictionary learning. Recent research has addressed non-convex optimization in a distributed manner through the use of stochastic local cost functions [8–10]. The studies conducted in [8, 10] demonstrate that the diffusion learning strategy continues to produce meaningful estimates in non-convex environments, as the iterates by individual agents tend to cluster in a narrow region close to the network centroid (an auxiliary variable during network evolution). In [9], the authors introduce a general stochastic unified decentralized algorithm (SUDA). It ensures convergence under both non-convex and the Polyak-Łojasiewicz condition settings. It is important to notice that all three aforementioned works emphasize the assumption that both individual and global costs are non-convex.

However, it is recognized that situations where the costs are locally non-convex but their sum is (strongly) convex, referred to as sum-of-non-convex objective, arise in solving problems such as regularized loss minimization [11], approximate principle component analysis (PCA) problems [12], distributed adaptive beamforming [13], and privacy-enhancing distributed optimization [14]. This observation has motivated researchers to develop distributed adaptation strategies based on the sum-of-non-convex objective. The work presented in [14] demonstrates that coupled consensus and projected gradient descent algorithms can converge to the optimum under the assumption of non-increasing step-sizes and gradient Lipschitzness, which, unfortunately, limits the adaptive filter's ability to effectively estimate and track over time.

In this paper, we investigate how the diffusion adaptation strategy can minimize a sum-of-non-convex objective using stochastic gradients in complex optimization problems. We analyze its convergence behavior in a mean-square sense, where the individual estimates will be close to certain neighbors of the global optimal point, forming a small cluster, for a small fixed step-size. Finally, we perform simulations to validate our theory and compare it with its counterpart of (strongly) convex local costs.

Notation. Lowercase letters x, boldface lowercase letters $\boldsymbol{x}$ and boldface capital letters $\boldsymbol{X}$ denote scalars, column vectors and matrices, respectively. The expected value of a random variable is denoted by $\mathbb{E}\{\cdot\}$. $\mathfrak{Re}\{\cdot\}$ and $\mathfrak{Im}\{\cdot\}$ denote the real part and imaginary part of the argument, respectively. Operator $\|\cdot\|$ denotes the ℓ_2-norm of a vector or matrix. Operator ∇ refers to the gradient of a function with respect to its variables. Hermitian transpose of a vector or matrix is denoted by $(\cdot)^{\mathrm{H}}$. Operator $\mathrm{col}\{\cdot\}$ provides a column vector by stacking its arguments on top of each other. Operator $\otimes$ refers to the Kronecker product. Vector $\mathbb{1}_N$ denotes the all-one vector of length N, and $\boldsymbol{I}_M$ is the identity matrix of size $M \times M$.

2. DIFFUSION ADAPTATION WITH SUM-OF-NON-CONVEX COST

2.1. Model assumptions

We consider a network of N interconnected nodes. Each node k has access to local measurements. The optimization problem is to find a Pareto solution $\boldsymbol{w}^o \in \mathbb{C}^M$ that minimizes the following real-valued

The work of M. Zhang was supported partly by the China Scholarship Council and partly by the Innovation Foundation for Doctor Dissertation of Northwestern Polytechnical University. The work of J. Chen is partially supported by NSFC grant 61671382, Shaanxi Key Industrial Innovation Chain Project 2022ZDLGY01-02, and Xi'an Technology Industrialization Plan XA2020-RGZNTJ-0076. The work of C. Richard was funded in part by the PIA program under its IDEX UCAJEDI project (ANR-15-IDEX-0001) and by ANR under grant ANR-19-CE48-0002.

global cost function defined over $\mathbb{C}^M$, i.e., $J^{\text{glob}}(\boldsymbol{w}) : \mathbb{C}^M \to \mathbb{R}$:

$$\boldsymbol{w}^o = \arg\min_{\boldsymbol{w}} \left(J^{\text{glob}}(\boldsymbol{w}) \triangleq \sum_{k=1}^{N} \alpha_k J_k(\boldsymbol{w}) \right), \tag{1}$$

in a collaborative and distributed manner. Factors α_k are positive and normalized to add up to one. Local cost $J_k(\boldsymbol{w})$ at each node k is defined as the expectation of an instantaneous risk of the form $J_k(\boldsymbol{w}) = \mathbb{E}_{\boldsymbol{x}}\{\mathcal{J}_k(\boldsymbol{w}; \boldsymbol{x}_k)\}$ with $\mathbb{E}_{\boldsymbol{x}}\{\cdot\}$ the expectation w.r.t the stochastic variable $\boldsymbol{x}_k$. We will keep each α_k fixed as a constant throughout this work, specifically $\alpha_k = \frac{1}{N}$. This specific case can be easily generalized. Our motivation for this choice stems from the need to tackle the challenges of non-convexity in certain real-world distributed signal processing problems. In [13, 15], the local costs of individual nodes are non-convex while their sum remains a strongly convex objective.

2.2. Diffusion adaptation for distributed estimation

Before presenting the diffusion strategy, let us introduce the definition of the gradient in the complex domain. Consider the case where J is a real-valued differentiable function of a complex variable $\boldsymbol{w} = \boldsymbol{u} + \mathrm{i}\boldsymbol{v} = \mathfrak{Re}\{\boldsymbol{w}\} + \mathrm{i}\mathfrak{Im}\{\boldsymbol{w}\}$, where $\boldsymbol{u}$ and $\boldsymbol{v}$ are real vectors. The gradient of J is defined as [16, Equation (1.53)]:

$$\nabla_{\boldsymbol{w}} J(\boldsymbol{w}) \triangleq \nabla_{\boldsymbol{u}} J(\boldsymbol{w}) + \mathrm{i}\nabla_{\boldsymbol{v}} J(\boldsymbol{w}). \tag{2}$$

Inspired by the distributed optimization algorithm proposed in [14], this paper aims to demonstrate how the existing diffusion strategy can be utilized to optimize the sum-of-non-convex costs (1) using stochastic gradient descent. The adapt-then-combine (ATC) diffusion strategy is formulated as follows:

$$\boldsymbol{\psi}_{k,i} = \boldsymbol{w}_{k,i-1} - \mu \hat{\nabla} J_k(\boldsymbol{w}_{k,i-1}) \tag{3}$$

$$\boldsymbol{w}_{k,i} = \sum_{\ell \in \mathcal{N}_k} a_{\ell k} \boldsymbol{\psi}_{\ell,i}. \tag{4}$$

Here, $\hat{\nabla} J_k(\boldsymbol{w}_{k,i-1})$ represents a stochastic approximation of the true local gradient $\nabla J_k(\boldsymbol{w}_{k,i-1})$. The coefficients $a_{\ell k}$ are non-negative and correspond to the (ℓ, k)-th entries of a left-stochastic matrix $\boldsymbol{A}$, which satisfies:

$$\boldsymbol{A}^\top \mathbb{1}_N = \mathbb{1}_N, \;\; a_{\ell k} = 0 \;\text{ if }\; \ell \notin \mathcal{N}_k. \tag{5}$$

Here, $\mathcal{N}_k$ represents the set of neighboring nodes of node k, including k, and μ is a uniform positive step-size used by all nodes. Let $\boldsymbol{p}$ denote the right-eigenvector, known as the Perron eigenvector, of matrix $\boldsymbol{A}$. After normalizing its entries p_k, which are all strictly positive, we have $\boldsymbol{A}\boldsymbol{p} = \boldsymbol{p}$ and $\mathbb{1}_N^\top \boldsymbol{p} = 1$. A special case is when $p_k = \frac{1}{N}$ for all k. Matrix $\boldsymbol{A}$ is then said to be a doubly stochastic matrix because it satisfies:

$$\boldsymbol{A}^\top \mathbb{1}_N = \boldsymbol{A}\mathbb{1}_N = \mathbb{1}_N. \tag{6}$$

We introduce the following definitions and assumptions to facilitate the derivation. It is important to note that these definitions and assumptions are similar to those used in previous works [17–21].

D.1. *The collection of all possible random events generated by the past iterations $\{\boldsymbol{w}_j\}$ up to time $j \le i$ over the network is defined as:*

$$\boldsymbol{\mathcal{F}}_i \triangleq \textit{filtration}\{\boldsymbol{w}_0, \boldsymbol{w}_1, \cdots, \boldsymbol{w}_i\}. \tag{7}$$

D.2. *The gradient noise vector $\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})$ at node k and time i is defined as the difference:*

$$\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1}) \triangleq \hat{\nabla} J_k(\boldsymbol{w}_{k,i-1}) - \nabla J_k(\boldsymbol{w}_{k,i-1}). \tag{8}$$

A.1. *Each $J_k(\boldsymbol{w})$ for $k \in \{1, \ldots, N\}$ may not be convex, while the weighted aggregate of the local costs $J^{glob}(\boldsymbol{w})$ is strongly convex. Specifically, there exists a constant $c > 0$ such that the following inequality holds for all $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{C}^M$:*

$$J^{glob}(\boldsymbol{w}_1) \ge J^{glob}(\boldsymbol{w}_2) + \mathfrak{Re}\{\nabla J^{glob}(\boldsymbol{w}_2)^{\mathrm{H}}(\boldsymbol{w}_1 - \boldsymbol{w}_2)\} + \frac{c}{2}\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|^2.$$

A.2. *Each $J_k(\boldsymbol{w})$ is δ-upper smooth and Δ-lower smooth, that is: $-\frac{\Delta}{2}\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|^2 \le J_k(\boldsymbol{w}_1) - [J_k(\boldsymbol{w}_2) + \mathfrak{Re}\{\nabla J_k(\boldsymbol{w}_2)^{\mathrm{H}}(\boldsymbol{w}_1 - \boldsymbol{w}_2)\}] \le \frac{\delta}{2}\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|^2$, for all $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{C}^M$. Constants δ and Δ satisfy the conditions $c \le \delta$ and $\Delta \ge 0$, respectively.*

A.3. *For all pairs of agents k and ℓ, the gradient disagreement is bounded, namely, for any $\boldsymbol{w} \in \mathbb{C}^M$: $\|\nabla J_k(\boldsymbol{w}) - \nabla J_\ell(\boldsymbol{w})\| \le \zeta$.*

A.4. *Matrix $\boldsymbol{A}$ is a doubly stochastic matrix. This property implies that $\|\boldsymbol{A} - \frac{1}{N}\mathbb{1}_N\mathbb{1}_N^\top\| = \lambda \in (0, 1)$, where λ is the second largest eigenvalue of $\boldsymbol{A}$.*

A.5. *The gradient noise vector $\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})$ satisfies:*

$$\mathbb{E}\{\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})|\boldsymbol{\mathcal{F}}_{i-1}\} = 0, \tag{9}$$

$$\mathbb{E}\{\|\boldsymbol{s}_{k,i}(\boldsymbol{w}_{k,i-1})\|^2|\boldsymbol{\mathcal{F}}_{i-1}\} \le \sigma^2. \tag{10}$$

3. CONVERGENCE ANALYSIS

In this section, we shall examine the convergence of the ATC algorithm for minimizing the sum-of-non-convex costs. Due to space limitations, we provide only a brief overview of the analysis. The detailed proofs are omitted but will be available as a supplementary material.

The convergence theorem can be proved in two steps. First, we establish a consensus lemma that shows the consensus distance $\mathbb{E}\{\|\boldsymbol{w}_{k,i} - \bar{\boldsymbol{w}}_i\|^2\} \le O(\mu^2)$ is bounded. This implies that $\boldsymbol{w}_{k,i}$ converges to a small cluster in the vicinity of the network average $\bar{\boldsymbol{w}}_i$ for small step-size μ. Second, we establish a theorem that demonstrates the geometric convergence of $J^{\text{glob}}(\bar{\boldsymbol{w}}_i)$ towards a neighbor of the optimal value. By combining these two steps, we conclude that, after a sufficient number of iterations or with the use of a suitable warm-up strategy, the estimates $\boldsymbol{w}_{k,i}$ will converge to neighbors of the global optimal point within a small cluster for a small step-size μ.

By collecting $\boldsymbol{w}_{k,i}$ and $\hat{\nabla} J_k(\boldsymbol{w}_{k,i})$ over the entire network, we obtain block vectors of dimension $MN \times 1$ as:

$$\boldsymbol{w}_i \triangleq \text{col}\{\boldsymbol{w}_{1,i}, \cdots, \boldsymbol{w}_{N,i}\}, \tag{11}$$

$$\hat{\boldsymbol{g}}_i \triangleq \text{col}\{\hat{\nabla} J_1(\boldsymbol{w}_{1,i}), \cdots, \hat{\nabla} J_N(\boldsymbol{w}_{N,i})\}. \tag{12}$$

Based on the definitions above, we can write the diffusion recursion (3)–(4) compactly as:

$$\boldsymbol{w}_i = \boldsymbol{\mathcal{A}}^\top(\boldsymbol{w}_{i-1} - \mu\hat{\boldsymbol{g}}_{i-1}), \tag{13}$$

with $\boldsymbol{\mathcal{A}} \triangleq \boldsymbol{A} \otimes \boldsymbol{I}_M$. Left-multiplying both sides of equation (13) by $\frac{1}{N}\mathbb{1}_N^\top \otimes \boldsymbol{I}_M$, we obtain:

$$\begin{aligned}
&\Big(\frac{1}{N}\mathbb{1}_N^\top \otimes \boldsymbol{I}_M\Big)\boldsymbol{w}_i \\
&= \Big(\frac{1}{N}\mathbb{1}_N^\top \otimes \boldsymbol{I}_M\Big)\boldsymbol{\mathcal{A}}^\top(\boldsymbol{w}_{i-1} - \mu\hat{\boldsymbol{g}}_{i-1}) \\
&= \Big(\frac{1}{N}\mathbb{1}_N^\top \otimes \boldsymbol{I}_M\Big)\boldsymbol{w}_{i-1} - \frac{\mu}{N}(\mathbb{1}_N^\top \otimes \boldsymbol{I}_M)\hat{\boldsymbol{g}}_{i-1},
\end{aligned} \tag{14}$$

where we have used the mixed-product property: $(\boldsymbol{M}\otimes\boldsymbol{N})(\boldsymbol{X}\otimes\boldsymbol{Y})=(\boldsymbol{MX})\otimes(\boldsymbol{NY})$.

We define $\bar{\boldsymbol{w}}_i$ as the average of the estimates over the network, given by $\bar{\boldsymbol{w}}_i \triangleq \frac{1}{N}\sum_{k=1}^{N}\boldsymbol{w}_{k,i}=\big(\frac{1}{N}\mathbb{1}_N^{\top}\otimes\boldsymbol{I}_M\big)\boldsymbol{w}_i$. This intermediate variable $\bar{\boldsymbol{w}}_i$ plays a role during the convergence of $\boldsymbol{w}_{k,i}$, serving as the network-wide average of the estimates at iteration i. Combining (14) with the definition of $\bar{\boldsymbol{w}}_i$, we can derive the recursion for $\bar{\boldsymbol{w}}_i$ as follows:

$$\begin{aligned}\bar{\boldsymbol{w}}_i &= \bar{\boldsymbol{w}}_{i-1}-\frac{\mu}{N}\sum_{k=1}^{N}\hat{\nabla}J_k(\boldsymbol{w}_{k,i})\\ &= \bar{\boldsymbol{w}}_{i-1}-\frac{\mu}{N}\sum_{k=1}^{N}\nabla J_k(\bar{\boldsymbol{w}}_{i-1})-\mu\boldsymbol{f}_{i-1}-\mu\boldsymbol{s}_i,\end{aligned}\tag{15}$$

where we define the following variables as perturbation terms:

$$\boldsymbol{f}_{i-1}\triangleq\frac{1}{N}\sum_{k=1}^{N}\big(\nabla J_k(\boldsymbol{w}_{k,i-1})-\nabla J_k(\bar{\boldsymbol{w}}_{i-1})\big),\tag{16}$$

$$\boldsymbol{s}_i\triangleq\frac{1}{N}\sum_{k=1}^{N}\big(\hat{\nabla}J_k(\boldsymbol{w}_{k,i-1})-\nabla J_k(\boldsymbol{w}_{k,i-1})\big).\tag{17}$$

Note that the subscript i for $\boldsymbol{s}_i$ is due to definition **D.2**, indicating that $\boldsymbol{s}_i$ depends on the data up to time instant i. Making N copies of $\bar{\boldsymbol{w}}_i$ and stacking them entry-wise, we obtain:

$$\bar{\boldsymbol{\mathfrak{w}}}_i\triangleq\frac{1}{N}(\mathbb{1}_N\mathbb{1}_N^{\top}\otimes\boldsymbol{I}_M)\boldsymbol{w}_i.\tag{18}$$

Lemma 1 (**Consensus Distance**). *Under assumptions **A.2** to **A.5**, the consensus residue is bounded by:*

$$\begin{aligned}&\mathbb{E}\{\|\boldsymbol{w}_i-\bar{\boldsymbol{\mathfrak{w}}}_i\|^2\}\leq\\ &\Big(\frac{1+\lambda^2}{2}\Big)^i\mathbb{E}\{\|\boldsymbol{w}_0-\bar{\boldsymbol{\mathfrak{w}}}_0\|^2\}+\frac{8\lambda^2(N\zeta^2+N^2\sigma^2)}{(1-\lambda^2)^2}\mu^2.\end{aligned}\tag{19}$$

The first term on the right-hand side of equation (19) becomes negligible as i becomes sufficiently large, given that $\lambda\in(0,1)$. Additionally, we can employ a warm-up strategy where all nodes' iterates at the initial time instant $i=0$ are set to be identical, ensuring that $\mathbb{E}\{\|\boldsymbol{w}_0-\bar{\boldsymbol{\mathfrak{w}}}_0\|^2\}=0$.

Corollary 1. *Under assumptions **A.2** to **A.5** and using the warm-up strategy, the iterates $\boldsymbol{w}_{k,i}$ will converge to a small cluster in the vicinity of the network average $\bar{\boldsymbol{w}}_i$ for a small step-size μ:*

$$\mathbb{E}\{\|\boldsymbol{w}_i-\bar{\boldsymbol{\mathfrak{w}}}_i\|^2\}\leq\frac{8\lambda^2(N\zeta^2+N^2\sigma^2)}{(1-\lambda^2)^2}\mu^2.\tag{20}$$

Assumption **A.2** implies that $\|\nabla J^{\text{glob}}(\boldsymbol{w}_1)-\nabla J^{\text{glob}}(\boldsymbol{w}_2)\|\leq\delta\|\boldsymbol{w}_1-\boldsymbol{w}_2\|$. This allows us to establish the following bound:

$$\begin{aligned}&J^{\text{glob}}(\bar{\boldsymbol{w}}_i)\\ &\quad\leq J^{\text{glob}}(\bar{\boldsymbol{w}}_{i-1})+\mathfrak{Re}\Big\{\nabla J^{\text{glob}}(\bar{\boldsymbol{w}}_{i-1})^{\text{H}}(\bar{\boldsymbol{w}}_i-\bar{\boldsymbol{w}}_{i-1})\Big\}\\ &\quad+\frac{\delta}{2}\|\bar{\boldsymbol{w}}_i-\bar{\boldsymbol{w}}_{i-1}\|^2\\ &\quad\overset{(15)}{\leq} J^{\text{glob}}(\bar{\boldsymbol{w}}_{i-1})-\mu\|\nabla J^{\text{glob}}(\bar{\boldsymbol{w}}_{i-1})\|^2\\ &\quad+\mathfrak{Re}\Big\{-\mu\nabla J^{\text{glob}}(\bar{\boldsymbol{w}}_{i-1})^{\text{H}}(\boldsymbol{f}_{i-1}+\boldsymbol{s}_i)\Big\}\\ &\quad+\mu^2\frac{\delta}{2}\|\nabla J^{\text{glob}}(\bar{\boldsymbol{w}}_{i-1})+\boldsymbol{f}_{i-1}+\boldsymbol{s}_i\|^2.\end{aligned}\tag{21}$$

Lemma 2 (**Perturbation Bounds**). *Under assumptions **A.2** to **A.5**, considering **Corollary 1** and using the warm-up strategy, the perturbation terms in* (15) *satisfy the following inequalities for small step-size μ:*

$$\mathbb{E}\{\|\boldsymbol{f}_{i-1}\|^2\}\leq\frac{8(\delta^2+3\delta\Delta+4\Delta^2)\lambda^2(\zeta^2+N\sigma^2)}{(1-\lambda^2)^2}\mu^2,\tag{22}$$

$$\mathbb{E}\{\|\boldsymbol{s}_i\|^2|\boldsymbol{\mathcal{F}}_{i-1}\}\leq\sigma^2.\tag{23}$$

The results of **Lemma 2** and the relation (21) establish the bound for the network average as follows:

Theorem 1 (**Network Average Bound**). *Under assumptions **A.1** to **A.5**, and based on **Lemma 2**, suppose that the diffusion adaptation algorithm minimizing the global cost* (1) *is run with the warm-up strategy and a fixed step-size μ satisfying:*

$$0\leq\mu<\frac{1}{4\delta}.\tag{24}$$

Then, the expected optimality gap satisfies the following inequality:

$$\begin{aligned}&\mathbb{E}\{J^{glob}(\bar{\boldsymbol{w}}_i)-J_o^{glob}\}-\frac{c_2\mu+c_1\mu^2+2c_1\delta\mu^3}{2c(1-2\mu\delta)}\\ &\leq\big(1-c\mu(1-2\mu\delta)\big)^i\Big[\mathbb{E}\{J^{glob}(\bar{\boldsymbol{w}}_0)-J_o^{glob}\}\\ &\quad-\frac{c_2\mu+c_1\mu^2+2c_1\delta\mu^3}{2c(1-2\mu\delta)}\Big],\end{aligned}\tag{25}$$

with $c_1\triangleq\frac{8(\delta^2+3\delta\Delta+4\Delta^2)\lambda^2(\zeta^2+N\sigma^2)}{(1-\lambda^2)^2}$, $c_2\triangleq\delta\sigma^2$.

Theorem 1 illustrates the interplay between the step-size μ, the upper-smooth parameter δ and the bound σ^2 on the gradient noise. When the gradient computation is noisy, one can still use a fixed step-size and ensure that the expected global cost values w.r.t. the network average $\bar{\boldsymbol{w}}_i$ will converge at a geometric rate to a value close to the optimal value. The upper bound on the expected optimality gap is given by:

$$\limsup_{i\to\infty}\mathbb{E}\{J^{\text{glob}}(\bar{\boldsymbol{w}}_i)-J_o^{\text{glob}}\}\leq\frac{c_2\mu+c_1\mu^2+2c_1\delta\mu^3}{c},\tag{26}$$

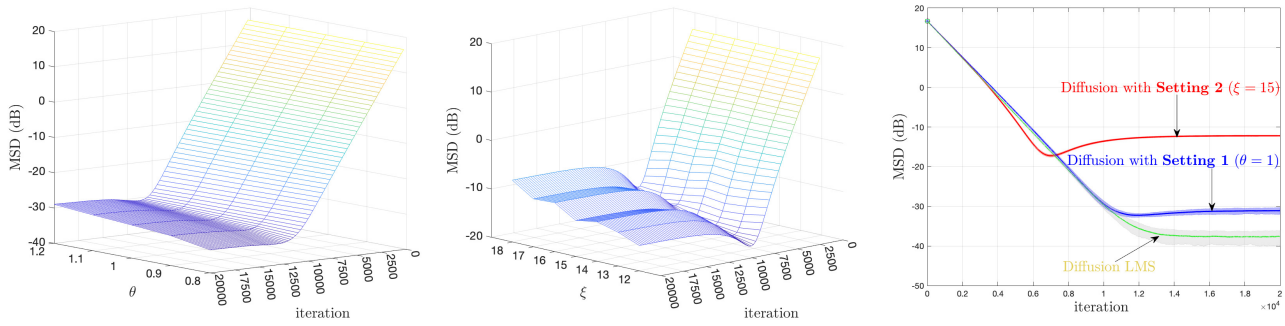
for sufficiently small step-size μ under condition (24). As shown in the definition of c_1, the lower-smooth parameter Δ, related to the non-convexity of local costs, clearly affects this upper bound. This will be illustrated in the experiments section.

4. SIMULATIONS

We present simulation results to demonstrate the effectiveness of the diffusion strategy in minimizing a strongly convex sum of non-convex local costs. In our simulations, we considered a network with 16 nodes as depicted in Figure 2(a). All nodes were initialized with all-zero parameter vectors, $\boldsymbol{w}_{k,0}=\boldsymbol{0}$. We conducted 100 Monte Carlo runs and averaged the results to obtain the reported curves.

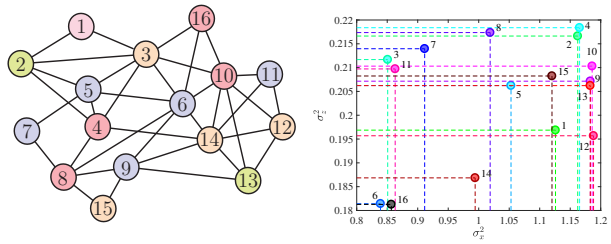
In the considered scenario, each agent k in the network observes streaming complex realizations $\{d_{k,i},\boldsymbol{x}_{k,i}\}$ from the linear model $d_{k,i}=\boldsymbol{x}_{k,i}^{\text{H}}\boldsymbol{w}^o+z_{k,i}$, where $z_{k,i}$ denotes complex zero-mean Gaussian noise with variance σ_z^2. The noise $z_{k,i}$ is independent of any other signals. In a distributed context, a common method for estimating $\boldsymbol{w}^o$ is the least-mean-square error (LMS) estimation, which leads to the following local cost functions:

$$J_k(\boldsymbol{w})=\mathbb{E}_{\boldsymbol{x}}(d_{k,i}-\boldsymbol{x}_{k,i}^{\text{H}}\boldsymbol{w}_{k,i})^2.\tag{27}$$



(a) MSD learning curves with varying θ. (b) MSD learning curves with varying ξ. (c) MSD learning curves.

Fig. 1: MSD learning curve properties and comparisons.



(a) Network topology. (b) Input and noise variances for each node.

Fig. 2: Network topology, node input and noise variances.

The regression inputs $\boldsymbol{x}_{k,i}$ are complex zero-mean random vectors of length $M = 50$ governed by a complex Gaussian distribution with covariance matrices $\boldsymbol{R}_{x,k} = \sigma^2_{\boldsymbol{x},k}\boldsymbol{I}_M$. Both σ^2_z and $\sigma^2_{\boldsymbol{x},k}$ for all nodes are provided in Figure 2(b). We set the unknown vector $\boldsymbol{w}^o$ to be a fixed set of variables sampled from $\mathcal{CN}(0,1)$. We set the combination matrix $\boldsymbol{A}$ as a doubly-stochastic matrix following the maximum-degree rule [22] with its entry $a_{\ell k}$ in (4) as:

$$a_{\ell k} = \begin{cases} \frac{1}{N}, & \text{if } k \neq \ell \quad \text{are neighbors} \\ 1 - \frac{|\mathcal{N}_k|-1}{N}, & k = \ell \\ 0, & \text{otherwise.} \end{cases} \tag{28}$$

Similar to the work [21], we designed two synthetic experiments based on the construction of the sum-of-non-convex costs. Specifically, given N diagonal matrices $\boldsymbol{D}_1, \ldots, \boldsymbol{D}_N$ that satisfy $\boldsymbol{D}_1 + \ldots + \boldsymbol{D}_N = \boldsymbol{0}$, we set each cost function $\mathcal{J}_k(\boldsymbol{w})$ as follows:

$$\mathcal{J}_k(\boldsymbol{w}) = (d_{k,i} - \boldsymbol{x}^{\mathrm{H}}_{k,i}\boldsymbol{w}_{k,i})^2 + \boldsymbol{w}^{\mathrm{H}}_{k,i}\boldsymbol{D}_k\boldsymbol{w}_{k,i}. \tag{29}$$

Under this construction, each $J_k = \mathbb{E}_{\boldsymbol{x}}\{\mathcal{J}_k\}$ is non-convex if $\boldsymbol{D}_k$ has negative entries on the diagonal smaller than $-\sigma^2_k$. We now present two different settings for designing $\boldsymbol{D}_k$.

Setting 1. Given $\theta \in [\min\{\sigma^2_{\boldsymbol{x},k}\}_{k=1}^N, \max\{\sigma^2_{\boldsymbol{x},k}\}_{k=1}^N)]$, for each $j \in \{1, \ldots, M\}$, we randomly select half of the indices k and we set $[\boldsymbol{D}_k]_{jj} = \theta$. We set $[\boldsymbol{D}_\ell]_{jj} = -\theta$ for the other half of the indices ℓ. In this way, for nodes k where $\theta \geq \sigma^2_{\boldsymbol{x},k}$, J_k is δ-upper smooth and Δ-lower smooth with $\delta = \sigma^2_{\boldsymbol{x},k} + \theta$ and $\Delta = \theta - \sigma^2_{\boldsymbol{x},k}$. Otherwise, J_k is strongly convex with $c = \sigma^2_{\boldsymbol{x},k} - \theta$.

Setting 2. Let $\xi = (N-1)\theta$, where θ is defined in **Setting 1**. For each $j \in \{1, \ldots, M\}$, we randomly select one matrix $\boldsymbol{D}_k$ and set its j-th diagonal entry, denoted as $[\boldsymbol{D}_k]_{jj}$, to $-\xi$. Simultaneously, we set the j-th diagonal entry of all other $N-1$ matrices to $\xi/(N-1)$. To ensure generality and prevent the diagonal elements of each matrix $\boldsymbol{D}_k$ from being uniformly positive or negative, we guarantee that each index k is selected at least once during this setting procedure. This approach ensures that each J_k is δ-upper smooth and Δ-lower smooth, where $\delta = \sigma^2_{\boldsymbol{x},k} + \theta$ and $\Delta = (N-1)\theta - \sigma^2_{\boldsymbol{x},k}$. Observe that **Setting 2** results in the same upper-smooth property as **Setting 1**, but it yields a distinct lower-smooth property with $\Delta \gg \delta$.

To analyze the effect of parameters on the performance of the diffusion algorithm, we varied the upper and lower smooth parameters δ and Δ. We conducted two scenarios to investigate the convergence behavior of the mean-square deviation (MSD) for different parameters θ and ξ. In scenario 1, we varied the parameter θ within the range of 0.8 to 1.2, with an increment of 0.1. In scenario 2, we explored the effect of different values of ξ by varying it with an increment of 1, from 12 to 18, which corresponds to approximately $0.8\times(N-1)$ to $1.2\times(N-1)$, with $N = 16$. For both scenarios, the diffusion adaptation step-size was set to $\mu = 5\cdot 10^{-4}$. Figure 1(a) illustrates the convergence behavior of the MSD with varying θ, while Figure 1(b) displays the convergence behavior with varying ξ. We observe that, as the parameters θ or ξ decrease, diffusion algorithm achieves superior steady-state performance.

Then we run the diffusion algorithm over sums of non-convex local costs with Settings 1 and 2, but also over the sum of strongly convex local costs by setting $\boldsymbol{D}_k = \boldsymbol{0}$ for all k. To validate the results obtained in (26), which indicate that parameter Δ significantly affects the upper bound on the expected optimality gap, we conducted the following experiment. We set $\theta = 1$ and $\xi = 15$ in Settings 1 and 2, ensuring that both have the same δ-upper smoothness characteristics. However, note that **Setting 2** was set with a large $\Delta \approx 14$-lower smoothness, while **Setting 1** was with a smaller value of $\Delta \approx 0.2$. The step-size μ in the diffusion LMS algorithm was set to $\mu = 5 \cdot 10^{-4}$. The MSD learning curves for the three settings are depicted in Figure 1(c). It can be observed that **Setting 1** achieves superior steady-state performance compared to **Setting 2** due to its relatively smaller upper bound on the expected optimality gap. Additionally, the performance of the diffusion LMS algorithm when run over the sum of strongly convex local costs is enhanced.

5. CONCLUSION

In this paper, we investigated the minimization of a strongly convex sum of non-convex local costs with the diffusion LMS. Through a comprehensive analysis, we have analyzed the convergence behavior of the network. Simulations were conducted to validate our theoretical findings and provide a comparison with a scenario involving (strongly) convex local costs.

6. REFERENCES

[1] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Trans. Automatic Control*, vol. 54, no. 1, pp. 48–61, 2009.

[2] C. G. Lopes and A. H. Sayed, "Incremental adaptive strategies over distributed networks," *IEEE Trans. Signal Process.*, vol. 55, no. 8, pp. 4064–4077, Aug. 2007.

[3] S. Kar and J. M. F. Moura, "Distributed consensus algorithms in sensor networks with imperfect communication: Link failures and channel noise," *IEEE Trans. Signal Process.*, vol. 57, no. 1, pp. 355–369, 2009.

[4] A. H. Sayed, "Diffusion adaptation over networks," in *Academic Press Library in Signal Processing*, vol. 3, pp. 323–453. Elsevier, 2014.

[5] C. G. Lopes and A. H. Sayed, "Diffusion least-mean squares over adaptive networks: Formulation and performance analysis," *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 3122–3136, July 2008.

[6] J. Chen, C. Richard, and A. H. Sayed, "Multitask diffusion adaptation over networks," *IEEE Trans. Signal Process.*, vol. 62, no. 16, pp. 4129–4144, Aug. 2014.

[7] J. Chen, C. Richard, and A. H. Sayed, "Diffusion LMS over multitask networks," *IEEE Trans. Signal Process.*, vol. 63, no. 11, pp. 2733–2748, Jun. 2015.

[8] S. Vlaski and A. H. Sayed, "Distributed learning in non-convex environments—part i: Agreement at a linear rate," *IEEE Trans. Signal Process.*, vol. 69, pp. 1242–1256, 2021.

[9] S. A. Alghunaim and K. Yuan, "A unified and refined convergence analysis for non-convex decentralized learning," *IEEE Trans. Signal Process.*, vol. 70, pp. 3264–3279, 2022.

[10] S. Vlaski and A. H. Sayed, "Distributed learning in non-convex environments—part ii: Polynomial escape from saddle-points," *IEEE Trans. Signal Process.*, vol. 69, pp. 1257–1270, 2021.

[11] Sh. Shalev-Shwartz, "Sdca without duality, regularization, and individual convexity," in *Proc. Int. Conf. Mach. Learn.* PMLR, 2016, pp. 747–754.

[12] D. Garber and E. Hazan, "Fast and simple pca via convex optimization," *arXiv preprint arXiv:1509.05647*, 2015.

[13] M. O'Connor, W. B. Kleijn, and T. Abhayapala, "Distributed sparse mvdr beamforming using the bi-alternating direction method of multipliers," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 106–110.

[14] S. Gade and N. H. Vaidya, "Distributed optimization of convex sum of non-convex functions," *arXiv preprint arXiv:1608.05401*, 2016.

[15] K. Hu, D. Jin, W. Zhang, and J. Chen, "Distributed optimization of quadratic costs with a group-sparsity regularization via pdmm," in *2018 Asia-Pacific Signal Inf. Process. Ass. Ann. Sum. and Conf. (APSIPA ASC)*, 2018, pp. 1825–1830.

[16] C. Y. Chi, W. C. Li, and C. H. Lin, *Convex optimization for signal processing and communications: from fundamentals to applications*, CRC press, 2017.

[17] A. H. Sayed, "Adaptation, learning, and optimization over networks," *Foundations and Trends in Machine Learning*, vol. 7, no. 4-5, pp. 311–801, 2014.

[18] B. Swenson, S. Kar, H. V. Poor, and J. M. F. Moura, "Annealing for distributed global optimization," in *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, 2019, pp. 3018–3025.

[19] S. Vlaski and A. H. Sayed, "Distributed learning in non-convex environments—part I: Agreement at a linear rate," *IEEE Trans. Signal Process.*, vol. 69, pp. 1242–1256, 2021.

[20] B. Ying, K. Yuan, Y. Chen, H. Hu, P. Pan, and W. Yin, "Exponential graph is provably efficient for decentralized deep training," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 13975–13987, 2021.

[21] Z. Allen-Zhu and Y. Yuan, "Improved svrg for non-strongly-convex or sum-of-non-convex objectives," in *Proc. Int. Conf. Mach. Learn.* PMLR, 2016, pp. 1080–1089.

[22] L. Xiao, S. Boyd, and S. Lall, "A scheme for robust distributed sensor fusion based on average consensus," in *Proc. IPSN, 2005*. Los Angeles, CA, April 2005, pp. 63–70.