DISTRIBUTED COUPLED LEARNING OVER ADAPTIVE NETWORKS

Sulaiman A. Alghunaim⋆ and Ali H. Sayed†

⋆Department of Electrical and Computer Engineering, University of California, Los Angeles
†School of Engineering, Ecole Polytechnique Federale de Lausanne, Switzerland

ABSTRACT

This work develops an effective distributed algorithm for the solution of stochastic optimization problems that involve partial coupling among both local constraints and local cost functions. While the collection of networked agents is interested in discovering a global model, the individual agents are sensing data that is only dependent on parts of the model. Moreover, different agents may be dependent on different subsets of the model. In this way, cooperation is justified and also necessary to enable recovery of the global information. In view of the local constraints, we show how to relax the optimization problem to a penalized form, and how to enable cooperation among neighboring agents. We establish mean-square-error convergence of the resulting strategy for sufficiently small step-sizes and large penalty factors. We also illustrate performance by means of simulations.

Index Terms— Distributed learning, diffusion strategy, stochastic optimization, coupled optimization, multi-agent networks.

I. INTRODUCTION AND RELATED WORK

Consider a multi-agent optimization problem consisting of N networked agents, where each agent is associated with an individual cost function, $J_k(w)$. There have been extensive works in the literature (e.g., [1]–[12] and the references therein) where effective algorithms have been developed for the distributed solution of constrained optimization problems of the form:

$$\min_{w} \quad \sum_{k=1}^{N} J_k(w), \ \text{s.t.} \quad w \in \mathbb{W}_1 \cap \cdots \cap \mathbb{W}_N \tag{1}$$

where $\mathbb{W}_k$ denotes a convex constraint set at node k. In this formulation, each cost $J_k(w)$ is a function of the same parameter vector, $w \in \mathbb{R}^M$. However, in many applications such as in multitask learning [13]–[15], distributed wireless localization [16], minimum-cost flow problems [17], and distributed power systems monitoring [18], the individual costs $J_k(\cdot)$ may be functions of only a few entries of w; moreover, different agents may be functions of different subsets of these parameters. Motivated by these scenarios, we consider in this work a more general problem where we assume that there are L variables, denoted by $\{w^1, w^2, \ldots, w^L\}$ with each $w^\ell \in \mathbb{R}^{M_\ell}$. We also assume that the cost of each agent is a function of only a subset of these variables. Without loss of generality, we assume the $\{w^\ell\}$ are distinct and do not share entries. In reference [19], we examined a special case of this situation without constraints and under the assumption that the exact gradient vectors of the costs $J_k(.)$ are available to the designer. Under these conditions, it was possible to rely on exact diffusion techniques [20] to solve the optimization problem *exactly*. In this work, we generalize the results in several respects: (a) we add local constraints represented by the sets $\{\mathbb{W}_k\}$, which are only available locally; (b) we do not assume knowledge of exact gradients and replace them by stochastic approximations from streaming data; and (c) we reformulate the optimization problem by employing penalty-methods to arrive at a fully-distributed solution for coupled costs.

This work was supported in part by NSF grant CCF-1524250. Emails: salghunaim@ucla.edu and ali.sayed@epfl.ch

Thus, let $\mathcal{I}_k$ denote the set of variable indices that affect the cost of agent k and let w_k denote the collection of variables that affect this agent:

$$w_k \triangleq \text{col}\{w^\ell\}_{\ell \in \mathcal{I}_k} \in \mathbb{R}^{Q_k}, \quad Q_k \triangleq \sum_{\ell \in \mathcal{I}_k} M_\ell. \tag{2}$$

If we stack all variables into a larger $L \times 1$ block vector $w \triangleq \text{col}\{w^1, w^2, ..., w^L\} \in \mathbb{R}^M$, then we are reduced to determining the solution of the optimization problem:

$$\min_{w} \ J^{\text{glob}}(w) \triangleq \sum_{k=1}^{N} J_k(w_k), \ \text{s.t.} \quad w \in \mathbb{W}_1 \cap \cdots \cap \mathbb{W}_N \tag{3}$$

Since different agents may be influenced by common vectors $\{w^\ell\}$, cooperation becomes desirable and is often necessary to improve accuracy and to ensure that agents reach agreement about the unknown shared parameters. Figure 1 illustrates the formulation for a simple network.

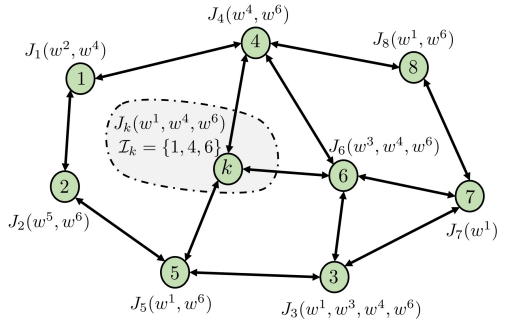


Fig. 1: A connected network of agents where the local costs depend on different subsets of the global parameter vector. For this example, we have $w = [w^1, w^2, w^3, w^4, w^5, w^6]$.

The constraint sets $\mathbb{W}_k$ are generally described by equality and inequality conditions of the form:

$$\mathbb{W}_k = \left\{ w : \begin{array}{l} h_{k,u}(w_k) = 0, \quad u = 1, ..., U_k \\ g_{k,v}(w_k) \le 0, \quad v = 1, ..., V_k \end{array} \right. \tag{4}$$

where $\{h_{k,u}(\cdot), g_{k,v}(\cdot)\}$ are convex functions. Problem (3) is assumed to be feasible and therefore, a minimizer exists

$$w^o = \text{col}\{w^{1,o}, \cdots, w^{L,o}\} \triangleq \underset{w \in \mathbb{W}_1 \cap \cdots \cap \mathbb{W}_N}{\arg\min} \ J^{\text{glob}}(w) \tag{5}$$

It is clear that algorithms that solve (1) can be used to solve (3). For example, this can be achieved by extending each local variable w_k into the longer global variable w. However, this solution method would require unnecessary communications and memory

allocation, and has been observed in simulations (see [19], [21]) to lead to performance degradation. It is therefore necessary to solve problem (3) more directly and also more effectively. As noted in [19], problems of the type (3), with partial coupling among the local costs, have been less studied in the literature than problems similar to (1). Some useful special cases, including variations without constraints or variations that assume access to exact gradient calculations, appear in [18], [21]–[23] using ADMM methods or similar primal-dual methods. In [24] a special case of (3) is solved under a stochastic environment where each agent cost is a quadratic function of w^k and coupling occurs with neighboring nodes through linear constraints. The works [4] and [5] consider stochastic settings for problem (1), and thus, cannot solve (3) directly.

In this work, we will solve problem (3) directly under a *stochastic* environment where agents do not necessarily know the exact gradient information but are subject to noisy perturbations as is the case in learning with streaming data. We will employ *constant* step-size learning in order to endow the resulting recursions with adaptation abilities to drifts in the models. It was shown in [25], [26] that under such scenarios, diffusion strategies [1] have superior performance than consensus strategies and primal-dual methods. Additionally, it is explained in [20] that diffusion strategies can be motivated by optimizing penalized costs. For these reasons, we shall employ in this work penalized diffusion methods to solve problem (3).

Notation. We use boldface letters to denote random quantities and regular font to denote their realizations or deterministic variables. For any set $\mathcal{X} = \{n_1, n_2, \cdots, n_x\}$, where n_s is an integer. We let $y = \mathrm{col}\{b_i\}_{i\in\mathcal{X}}$ denote a column vector with r-th entry $y(r) = b_{n_r}$ and $U = [c_{ij}]_{i,j\in\mathcal{X}}$ denote a matrix with (r,q) entry $U(r,q) = c_{n_r n_q}$.

II. PENALIZED FORMULATION

We start our derivation by adapting the technique from [2] to fit problem (3). We first relax problem (3) and replace it by the following penalized form parametrized by a scalar $\eta \geq 0$ ($\eta = 0$ for the unconstrained case):

$$\underset{w}{\text{minimize}} \quad J_\eta^{\text{glob}}(w) \triangleq \sum_{k=1}^{N} J_{k,\eta}(w_k) \tag{6}$$

where the individual costs on the right-hand side incorporate a penalty term, and are defined as follows:

$$J_{k,\eta}(w_k) \triangleq J_k(w_k) + \eta\, p_k(w_k) \tag{7}$$

with each penalty function in (7) given by

$$p_k(w_k) \triangleq \sum_{u=1}^{U_k} \delta^{\text{EP}}\big(h_{k,u}(w_k)\big) + \sum_{v=1}^{V_k} \delta^{\text{IP}}\big(g_{k,v}(w_k)\big) \tag{8}$$

Here, the terms $\delta^{\text{EP}}(x)$ and $\delta^{\text{IP}}(x)$ denote differentiable convex functions that penalize the violation of the constraints, namely, they satisfy the requirements:

$$\delta^{\text{EP}}(x) = \begin{cases} 0, & x = 0 \\ > 0, & x \neq 0 \end{cases}, \quad \delta^{\text{IP}}(x) = \begin{cases} 0, & x \leq 0 \\ > 0, & \text{otherwise} \end{cases} \tag{9}$$

Reference [2] provides useful examples of such functions. We denote the optimal solution of (6) by:

$$w^\star = \mathrm{col}\{w^{1,\star}, \cdots, w^{L,\star}\} \triangleq \underset{w^1,\cdots,w^L}{\arg\min}\, J_\eta^{\text{glob}}(w) \tag{10}$$

II-A. Reformulation for Distributed Implementation

In order to solve (6) in a distributed manner, we first need to adjust the notation to account for one additional degree of freedom. Since the costs of two arbitrary agents k and s, may depend on the same sub-vector, w^ℓ, and these two agents will be learning w^ℓ over time, each one of them will have its own local estimate for w^ℓ. Thus, we refer to w^ℓ at agent k by w_k^ℓ and to the same w^ℓ at agent s by w_s^ℓ. With this in mind, we redefine w_k; defined earlier in (2) using the local copies instead, namely, we now write

$$w_k \;\triangleq\; \mathrm{col}\{w_k^\ell\}_{\ell\in\mathcal{I}_k} \in \mathbb{R}^{Q_k} \tag{11}$$

We further let $\mathcal{C}_\ell$ denote the cluster of nodes that contains the variable w^ℓ in their costs:

$$\mathcal{C}_\ell = \{k \mid \ell \in \mathcal{I}_k\} \tag{12}$$

To require all local copies $\{w_k^\ell\}_{k\in\mathcal{C}_\ell}$ to coincide with each other, we introduce the constraint

$$w_k^\ell = w_s^\ell, \quad \forall\; k, s \in \mathcal{C}_\ell \tag{13}$$

Using relations (11) and (13), we can rewrite problem (6) as

$$\begin{aligned} \underset{w_1,\ldots,w_N}{\text{minimize}} \quad & J_\eta^{\text{glob}}(w_1, ..., w_N) \triangleq \sum_{k=1}^{N} J_{k,\eta}(w_k) \\ \text{subject to} \quad & w_k^\ell = w_s^\ell, \;\; \forall\; k, s \in \mathcal{C}_\ell, \; \forall\; \ell \end{aligned} \tag{14}$$

III. COUPLED DIFFUSION STRATEGY

To solve (14), we associate with each cluster $\mathcal{C}_\ell$ a set of coefficients $\{a_{\ell,sk}\}_{s,k\in\mathcal{C}_\ell}$ that are chosen to satisfy:

$$\sum_{s\in\mathcal{C}_\ell} a_{\ell,sk} = 1, \quad \sum_{k\in\mathcal{C}_\ell} a_{\ell,sk} = 1 \tag{15}$$

$$a_{\ell,sk} \geq 0, \text{ and } a_{\ell,sk} = 0 \text{ if } s \notin \mathcal{N}_k \tag{16}$$

Let N_ℓ denote the cardinality of cluster $\mathcal{C}_\ell$ and introduce the $N_\ell \times N_\ell$ matrices:

$$A_\ell \;\triangleq\; [a_{\ell,sk}]_{s,k\in\mathcal{C}_\ell} \tag{17}$$

Assumption 1. *(**Each cluster is strongly-connected**): The combinations matrices $\{A_\ell\}$ are assumed to be primitive, i.e., we assume that there exists a large enough j_0 such that the elements of $A_\ell^{j_0}$ have strictly positive entries. This implies that for any two arbitrary agents in cluster $\mathcal{C}_\ell$, there exists at least one path with nonzero weights $\{a_{\ell,sk}\}_{s,k\in\mathcal{C}_\ell}$ linking one agent to the other. Moreover, at least one self weight $\{a_{\ell,kk}\}_{k\in\mathcal{C}_\ell}$ is nonzero. We further assume the matrices $\{A_\ell\}$ to be symmetric and doubly stochastic.* ■

Assumption 1 is satisfied for most networks of interest. For example, applications in distributed power system monitering, distributed control, and maximum-flow problems satisfy this assumption — see [18], [21], [24]. Additionally, multitask applications satisfy this assumption [13]–[15]. In fact, this work is not limited to these scenarios, and can handle more general situations. Moreover, since most networks of interest are connected, then if some cluster $\mathcal{C}_\ell$ happens to be unconnected, we can embed it into a larger *connected* cluster $\mathcal{C}'_\ell$ such that $\mathcal{C}_\ell \subset \mathcal{C}'_\ell$ – see [19].

We state the following auxiliary result proven in [20].

Lemma 1. *For any $Q \times Q$ primitive, symmetric and doubly stochastic matrix A, it holds that $I_Q - A^{\mathsf{T}}$ is symmetric and positive semi-definite. Moreover, if we introduce the eigen-decomposition $\frac{1}{2}(I_Q - A^{\mathsf{T}}) = U\Sigma U^{\mathsf{T}}$, the symmetric square-root matrix: $V \triangleq U\Sigma^{1/2}U^{\mathsf{T}} \in \mathbb{R}^{Q\times Q}$ and let:*

$$\mathcal{A} = A \otimes I_M, \quad \mathcal{V} = V \otimes I_M \tag{18}$$

Then, for any block vector $x = \mathrm{col}\{x^1, ..., x^Q\}$ in the nullspace of $I - \mathcal{A}^{\mathsf{T}}$ with entries $x^k \in \mathbb{R}^M$ it holds that:

$$\mathcal{V}x = 0 \iff (I - \mathcal{A}^{\mathsf{T}})x = 0 \iff x^1 = x^2 = ... = x^Q \tag{19}$$

■

III-A. Coupled Diffusion Development

Lemma 1 allows us to rewrite (14) in an equivalent form that is amenable to distributed implementations. First, we introduce

$$\mathcal{w}^\ell \triangleq \mathrm{col}\{w_k^\ell\}_{k\in\mathcal{C}_\ell} \in \mathbb{R}^{N_\ell M_\ell} \tag{20}$$

as the collection of all local copies of w^ℓ across the agents in cluster $\mathcal{C}_\ell$. Next we use Lemma 1 to rewrite the constraints of problem (14) in an equivalent manner. Recall that each cluster $\mathcal{C}_\ell$ is associated with a symmetric doubly stochastic combination matrix A_ℓ. We appeal to Lemma 1 to decompose $\frac{1}{2}(I_{N_\ell} - A_\ell^{\mathsf{T}}) = U_\ell \Sigma_\ell U_\ell^{\mathsf{T}}$. If we let

$$V_\ell \triangleq U_\ell \Sigma_\ell^{1/2} U_\ell^{\mathsf{T}}, \quad \mathcal{V}_\ell \triangleq V_\ell \otimes I_{M_\ell}, \tag{21}$$

then using Lemma 1 and the definition of $\mathcal{w}^\ell$ in (20) we get

$$w_k^\ell = w_s^\ell, \ \forall \ k, s \in \mathcal{C}_\ell \Longleftrightarrow \mathcal{V}_\ell \mathcal{w}^\ell = 0, \quad \forall \ \ell. \tag{22}$$

Using relation (22), we can rewrite problem (14) equivalently as

$$\begin{aligned} \underset{\mathcal{w}^1,\ldots,\mathcal{w}^L}{\text{minimize}} \quad & \underbrace{\mathcal{J}(\mathcal{w}^1, \mathcal{w}^2, \cdots, \mathcal{w}^\ell)}_{\triangleq \sum_{k=1}^N J_k(w_k)} + \eta \underbrace{\mathcal{P}(\mathcal{w}^1, \mathcal{w}^2, \cdots, \mathcal{w}^\ell)}_{\triangleq \sum_{k=1}^N p_k(w_k)} \\ \text{subject to} \quad & \mathcal{V}_\ell \mathcal{w}^\ell = 0, \ \forall \ \ell \end{aligned} \tag{23}$$

To rewrite problem (23) more compactly, we introduce

$$\mathcal{V} \triangleq \text{diag}\{\mathcal{V}_\ell\}_{\ell=1}^L, \quad \mathcal{w} \triangleq \text{col}\{\mathcal{w}^\ell\}_{\ell=1}^L \in \mathbb{R}^S, \tag{24}$$

$$\mathcal{J}(\mathcal{w}) \triangleq \mathcal{J}(\mathcal{w}^1, \cdots, \mathcal{w}^L), \ \mathcal{P}(\mathcal{w}) \triangleq \mathcal{P}(\mathcal{w}^1, \cdots, \mathcal{w}^L) \tag{25}$$

where $S \triangleq \sum_{\ell=1}^L N_\ell M_\ell$. Then, problem (23) becomes:

$$\underset{\mathcal{w}}{\text{minimize}} \quad \mathcal{J}(\mathcal{w}) + \eta \mathcal{P}(\mathcal{w}), \ \text{s.t.} \quad \mathcal{V}\mathcal{w} = 0 \tag{26}$$

Instead of solving the constrained problem (26), we further relax it and solve the penalized version:

$$\underset{\mathcal{w}}{\text{minimize}} \quad \mathcal{J}(\mathcal{w}) + \eta \mathcal{P}(\mathcal{w}) + \frac{1}{\mu}\|\mathcal{V}\mathcal{w}\|^2, \tag{27}$$

with $\mu > 0$. The smaller the value of μ is, the closer the solutions of problem (26) and (27) become to each other. We now note that

$$\mathcal{V}^2 = \text{diag}\{\mathcal{V}_\ell^2\}_{\ell=1}^L = \frac{1}{2}(I_S - \mathcal{A}^{\mathsf{T}}), \ \mathcal{A} \triangleq \text{diag}\{\mathcal{A}_\ell\}_{\ell=1}^L \tag{28}$$

Applying an incremental gradient descent steps w.r.t. $\mathcal{w}$ using step-size μ to problem (27), we get:

$$\begin{cases} \zeta_i = \mathcal{w}_{i-1} - \mu\eta \nabla_{\mathcal{w}} \mathcal{P}(\mathcal{w}_{i-1}) \\ \psi_i = \zeta_i - \mu \nabla_{\mathcal{w}} \mathcal{J}(\zeta_i) \\ \mathcal{w}_i = \psi_i - \mu \left(\frac{2}{\mu}\mathcal{V}^2\right) \psi_i = \mathcal{A}^{\mathsf{T}} \psi_i \end{cases} \tag{29}$$

Using the definition of $\mathcal{J}(\mathcal{w})$ from (25), we have:

$$\nabla_{\mathcal{w}} \mathcal{J}(\mathcal{w}) = \text{col}\left\{\nabla_{\mathcal{w}^\ell} \mathcal{J}(\mathcal{w})\right\}_{\ell=1}^L \tag{30}$$

where

$$\nabla_{\mathcal{w}^\ell} \mathcal{J}(\mathcal{w}) = \text{col}\{\nabla_{w_k^\ell} J_k(w_k)\}_{k \in \mathcal{C}_\ell} \tag{31}$$

and likewise for $\nabla_{\mathcal{w}} \mathcal{P}(\mathcal{w})$. Therefore, using the definition of $\mathcal{A}$ in (28), recursion (29) can be written in a distributed form as listed in (32a)–(32c). In steps (32a)–(32b), a traditional gradient-descent step is applied by each agent using the gradients of the corresponding risk and penalty functions. The last step (32c) is a combination step, where for every $\ell \in \mathcal{I}_k$, each agent k combines its estimate for $\psi_{k,i}^\ell$ with the neighbors that belong to $\mathcal{C}_\ell$ using weights $\{a_{\ell,sk}\}_{s,k \in \mathcal{C}_\ell}$. It is assumed that $\psi_{k,i}$ and $\zeta_{k,i}$ have the same structure as $w_{k,i}$, i.e., $\psi_{k,i} = \text{col}\{\psi_{k,i}^\ell\}_{\ell \in \mathcal{I}_k}$ and $\zeta_{k,i} = \text{col}\{\zeta_{k,i}^\ell\}_{\ell \in \mathcal{I}_k}$. This latter step requires agent k to know the set $\mathcal{N}_k \cap \mathcal{C}_\ell$ for every $\ell \in \mathcal{I}_k$, i.e., to know the collection of neighboring agents that share the vector w^ℓ for every $\ell \in \mathcal{I}_k$ as part of their cost. In most networked problems of interest, this scenario is satisfied. For instance, in many applications $\mathcal{I}_k = \mathcal{N}_k$ and hence $\mathcal{C}_\ell$ will generally be defined by a subset of the neighboring agents (i.e., $L = N$ and $\mathcal{C}_k = \mathcal{N}_k$) [17], [18]. Therefore the set $\mathcal{N}_k \cap \mathcal{C}_\ell$ for every $\ell \in \mathcal{I}_k$ can be easily known by agent k. See simulation section for an example.

IV. CONVERGENCE ANALYSIS

To facilitate the convergence analysis, we introduce the following assumptions, which are common in the study of distributed learning methods. These assumptions are also automatically satisfied by many important cases of interest – see, e.g., [1], [2].

Algorithm 1 (Coupled diffusion strategy)

$$\zeta_{k,i} = w_{k,i-1} - \mu\eta \nabla_{w_k} p_k(w_{k,i-1}) \tag{32a}$$

$$\psi_{k,i} = \zeta_{k,i} - \mu \nabla_{w_k} J_k(\zeta_{k,i}) \tag{32b}$$

$$w_{k,i}^\ell = \sum_{s \in \mathcal{N}_k \cap \mathcal{C}_\ell} a_{\ell,sk} \psi_{s,i}^\ell, \ \ \forall \ \ell \in \mathcal{I}_k \tag{32c}$$

Assumption 2. *(**Individual costs**): It is assumed that the individual cost functions, $J_k(w_k)$, are each twice-differentiable, convex, and have Hessian matrices that are bounded from above:*

$$\nabla_{w_k}^2 J_k(w_k) \leq \delta_k I_{Q_k} \tag{33}$$

Moreover, for every cluster $\mathcal{C}_\ell$ there exists at least one agent k_o such that:

$$\nabla_{w_{k_o}}^2 J_{k_o}(w_{k_o}) > \nu_{k_o} I_{Q_{k_o}} \tag{34}$$

where the scalars $\{\delta_k\}$ and $\{\nu_{k_o}\}$ are strictly positive. ■

Note that assumption (34) is similar to requiring at least one of the costs $J_k(.)$ to be strongly convex within each cluster – see [2], [8]. This guarantees that the aggregate cost is strongly convex, and therefore a unique minimizer exists.

Assumption 3. *(**Penalty functions**): The penalty function $p_k(w_k)$ is twice-differentiable and its Hessian matrix is upper bounded:*

$$\nabla_{w_k}^2 p_k(w_k) \leq \delta_{p,k} I_{Q_k} \tag{35}$$

for some strictly positive scalars $\{\delta_{p,k}\}$. ■

In many applications in practice, the true gradient vectors are not available. Therefore, we model the approximate gradient vector for each agent at time i by:

$$\widehat{\nabla_{w_k} J_k}(\boldsymbol{\zeta}_{k,i}) \triangleq \nabla_{w_k} J_k(\boldsymbol{\zeta}_{k,i}) - \boldsymbol{v}_{k,i}(\boldsymbol{\zeta}_{k,i}) \tag{36}$$

where $\boldsymbol{v}_{k,i}(\boldsymbol{\zeta}_{k,i})$ is a random gradient noise term that is required to satisfy certain conditions.

Assumption 4. *(**Gradient noise model**): Conditioned on the past history of iterates $\boldsymbol{\mathcal{F}}_i \triangleq \{\boldsymbol{w}_{k,j-1} : k = 1, ..., N \text{ and } j \leq i\}$, the gradient noise $\boldsymbol{v}_{k,i}(\boldsymbol{\zeta}_k)$ is assumed to satisfy:*

$$\mathbb{E}\left\{\boldsymbol{v}_{k,i}(\boldsymbol{\zeta}_k) \mid \boldsymbol{\mathcal{F}}_i\right\} = 0 \tag{37}$$

$$\mathbb{E}\left\{\|\boldsymbol{v}_{k,i}(\boldsymbol{\zeta}_k)\|^2 \mid \boldsymbol{\mathcal{F}}_i\right\} \leq \bar{\alpha}_k \|\boldsymbol{\zeta}_k\|^2 + \bar{\sigma}_k^2 \tag{38}$$

for some nonnegative constants $\bar{\alpha}_k$ and $\bar{\sigma}_k^2$. ■

Using (36), the coupled diffusion algorithm (32) becomes

$$\boldsymbol{\zeta}_{k,i} = \boldsymbol{w}_{k,i-1} - \mu\eta \nabla_{w_k} p_k(\boldsymbol{w}_{k,i-1}) \tag{39a}$$

$$\boldsymbol{\psi}_{k,i} = \boldsymbol{\zeta}_{k,i} - \mu \nabla_{w_k} J_k(\boldsymbol{\zeta}_{k,i}) + \mu \boldsymbol{v}_{k,i}(\boldsymbol{\zeta}_{k,i}) \tag{39b}$$

$$\boldsymbol{w}_{k,i}^\ell = \sum_{s \in \mathcal{N}_k \cap \mathcal{C}_\ell} a_{\ell,sk} \boldsymbol{\psi}_{s,i}^\ell, \ \ \forall \ \ell \in \mathcal{I}_k \tag{39c}$$

We are now using boldface letters to highlight the fact that the variables are stochastic in nature due to the randomness in the gradient noise component. We will measure the performance of the distributed strategy by examining the mean-square-error between the random iterates $\boldsymbol{w}_{k,i}^\ell$ and the corresponding optimal component from (5), denoted by $w^{\ell,o}$. For this purpose, we note first that in terms of the optimal solution $w^{\ell,\star}$ for the penalized problem (10), we have:

$$\begin{aligned} &\limsup_{i \to \infty} \ \mathbb{E}\,\|w^{\ell,o} - \boldsymbol{w}_{k,i}^\ell\|^2 \\ &\leq 2 \underbrace{\|w^{\ell,o} - w^{\ell,\star}\|^2}_{\text{Approximation Error}} + 2 \limsup_{i \to \infty} \mathbb{E}\,\|w^{\ell,\star} - \boldsymbol{w}_{k,i}^\ell\|^2 \end{aligned} \tag{40}$$

It was shown in [2] that

$$\lim_{\eta\to\infty} \|w^o - w^\star\| = 0 \tag{41}$$

Theorem 1. *(**Mean-square convergence**): If w^o is a regular[1] point for the constraints, then, under Assumptions 1–4, the coupled diffusion algorithm* (39) *converges for sufficiently small step-sizes μ. Moreover, for every agent k, it holds that:*

$$\limsup_{i\to\infty} \mathbb{E}\,\|w^{\ell,\star} - \boldsymbol{w}_{k,i}^{\ell}\|^2 \le O(\mu) + O(\mu^2\eta^2),\ \forall\ \ell \in \mathcal{I}_k \tag{42}$$

Proof: See [27]. ■

Theorem 1 means that the expected squared distance between $\boldsymbol{w}_{k,i}^{\ell}$ and $w^{\ell,\star}$ is on the order μ or $(\mu\eta)^2$, whichever is larger. Thus we can get arbitrarily close to the optimal penalized solution $w^\star$ by choosing μ arbitrarily small. Moreover, from (41), we conclude that as $\eta\to\infty$ and $\mu\to 0$, the iterates $\boldsymbol{w}_{k,i}^{\ell}$ approach the optimal solutions of the unconstrained problem $w^{\ell,o}$ asymptotically.

V. SIMULATIONS

In this section we illustrate our results for mean square error (MSE) networks [3]. Consider a network of N agents where each agent k is observing streaming data $\{\boldsymbol{d}_k(i), \boldsymbol{u}_{k,i}\}$ that satisfy the regression model:

$$\boldsymbol{d}_k(i) + \boldsymbol{u}_{k,i} w^{k,\bullet} + \boldsymbol{v}_k(i) \tag{43}$$

where $\boldsymbol{u}_{k,i} \in \mathbb{R}^{1\times M_k}$ with covariance $R_{u,k} = \mathbb{E}\,\boldsymbol{u}_{k,i}^{\mathsf{T}}\boldsymbol{u}_{k,i}$, $w^{k,\bullet} \in \mathbb{R}^{M_k}$ is unknown, and $\boldsymbol{v}_k(i)$ is a noise process independent of $\boldsymbol{u}_{k,i}$ with variance $\sigma_{v,k}^2$. The individual mean-square-error costs are defined by:

$$J_k(w^k) = \mathbb{E}\,|\boldsymbol{d}_k(i) - \boldsymbol{u}_{k,i} w^k|^2 \tag{44}$$

The goal of the network is to solve the following problem:

$$\min_{w^1,\ldots,w^N} \sum_{k=1}^{N} J_k(w^k), \quad \text{s.t} \quad \sum_{s\in\mathcal{N}_k} B_{sk} w^s = b_k, \quad \forall\ k \tag{45}$$

where the matrices $B_{sk} \in \mathbb{R}^{P_k\times M_s}$ and the vector $b_k \in \mathbb{R}^{P_k}$ are known by agent k only. The linear constraints in (45) couples the parameters $\{w^1,\cdots,w^N\}$ across the network. For example, in beamforming applications, the constraints can be used to specify the desired response of the beamformers to certain directions [28] or it can be used in multitask applications where the task of each agent is coupled with its neighbors – see [24].

[1] w^o is a regular point if the gradients of the equality constraints and the active inequality constraints $\{\nabla_w h_{k,u}(w^o), \nabla_w g_{k,v'}(w^o)\}$ are linearly independent (where an active constraint means that $g_{k,v'}(w_k^o) = 0$ for some v', where $w_k^o = \mathrm{col}\{w^{\ell,o}\}_{\ell\in\mathcal{I}_k}$).

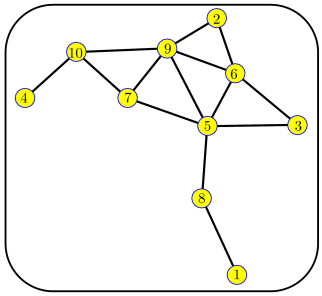


Fig. 3: Network topology used in simulation.

In our simulation, we considered the network with $N = 10$ agents shown in Fig. 3. Each parameter $w^{k,\bullet}$ is a 2×1 vector chosen from the standard Gaussian distribution. The inputs $\boldsymbol{u}_{k,i}$ are zero mean random vectors with covariance $R_{u,k} = \sigma_{u,k}^2 I_2$, where $\sigma_{u,k}^2$ was chosen uniformly at random between 1 and 3. The noise $\boldsymbol{v}_k(i)$ is a zero-mean i.i.d Gaussian random variable with variance $\sigma_{v,k}^2$ chosen uniformly at random between -25 dB and -35 dB. The elements of B_{sk} and b_k were generated uniformly at random between $(-2, 1)$ and $(0, 1)$ respectively. The constraints were penalized by using the quadratic penalty function $\delta^{\mathrm{EP}}(x) = x^2$. Note that existing algorithms in the literature require solving a deterministic inner minimization step in each iteration (see [18], [21]–[23]) and we are dealing with streaming data $\{\boldsymbol{d}_k(i), \boldsymbol{u}_{k,i}\}$ whose statistics are not known. Thus, these algorithms do not fit this scenario and cannot be directly applied to solve (45). Therefore, we will compare our algorithm with a penalty-based centralized recursion that assumes knowledge of all costs and constraints:

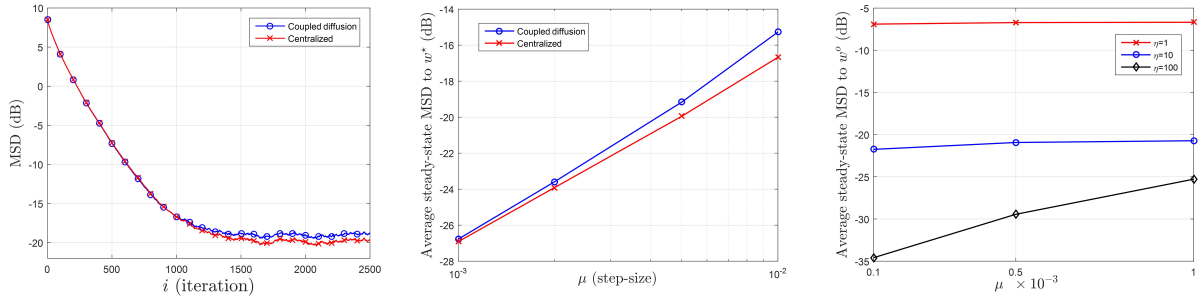
$$\psi_i = w_{i-1} - \mu\eta R\ \nabla_w p^{\mathrm{glob}}(w_{i-1}) \tag{46a}$$

$$w_i = \psi_i - \mu R\ \nabla_w J^{\mathrm{glob}}(\psi_i) \tag{46b}$$

where $p^{\mathrm{glob}}(w) = \sum_{k=1}^{N} p_k(w_k)$ and $R = \mathrm{diag}\{\frac{1}{N_\ell} I_{M_\ell}\}_{\ell=1}^{L}$ used to make the convergence rate similar for fair comparisons. Figure 2a shows the instantaneous network MSD

$$\mathrm{MSD} = \sum_{\ell=1}^{L} \frac{1}{N_\ell} \sum_{k\in\mathcal{C}_\ell} \mathbb{E}\,\|w^{\ell,\star} - \boldsymbol{w}_{k,i}^{\ell}\|^2 \tag{47}$$

to the penalized optimal values for $\mu = 0.005$ and $\eta = 10$ for both the coupled diffusion and the centralized recursions. Figure 2b plots the steady-state MSD for different values of step-sizes. We see that the smaller the value of μ is, the closer the solution is to $w^\star$ and the closer the coupled diffusion strategy become to the centralized one. Finally, Figure 2c shows the steady-state MSD of the coupled diffusion strategy to the constrained optimal w^o for different values of μ and η. We see that the larger η and the smaller μ are the closer we get to w^o. All results were averaged over 100 independent runs.



(a) **(b)** **(c)**

Fig. 2: (a) MSD learning curve for $\mu = 0.005$ and $\eta = 10$. (b) MSD for different values of step size μ with $\eta = 10$. (c) Average steady-state error to w^o for different values of η and μ.

VI. REFERENCES

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

[2] Z. J. Towfic and A. H. Sayed, "Adaptive penalty-based distributed stochastic convex optimization," *IEEE Trans. Signal Process*, vol. 62, no. 15, pp. 3924–3938, August 2014.

[3] A. H. Sayed, "Adaptive networks," *Proceedings of the IEEE*, vol. 102, no. 4, pp. 460–497, Apr. 2014.

[4] S. S. Ram, A. Nedic, and V. V. Veeravalli, "Distributed stochastic subgradient projection algorithms for convex optimization," *J. Optim. Theory Appl.*, vol. 147, no. 3, pp. 516–545, 2010.

[5] K. Srivastava and A. Nedic, "Distributed asynchronous constrained stochastic optimization," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 4, pp. 772–790, 2011.

[6] P. Braca, S. Marano, and V. Matta, "Enforcing consensus while monitering the environment in wireless sensor networks," *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 3375–3380, July 2008.

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

[8] J. Chen and A. H. Sayed, "Distributed Pareto optimization via diffusion strategies," *IEEE J. Sel. Topics Signal Process.*, vol. 7, no. 2, pp. 205–220, April 2013.

[9] J. F. Mota, J. M. Xavier, P. M. Aguiar, and M. Puschel, "D-ADMM: A communication-efficient distributed algorithm for sperable optimizatoin," *IEEE Trans. Signal Process.*, vol. 61, no. 10, pp. 2718–2723, 2013.

[10] W. Shi, Q. Ling, K. Yuan, G. Wu, and W. Yin, "On the convergence of the ADMM in decentralized consensus optimization," *IEEE Trans. Signal Process.*, vol. 62, no. 7, pp. 1750–1761, 2014.

[11] A. Mokhtari, W. Shi, Q. Ling, and A. Ribeiro, "DQM: Decentralized quadratically approximated alternating direction method of multipliers," *IEEE Trans. Signal Process.*, vol. 64, no. 19, pp. 5158–5173, 2016.

[12] K. Yuan, B. Ying, X. Zhao, and A. H. Sayed, "Exact diffusion strategy for optimization by networked agents," in *Proc. EUSIPCO*, Kos, Greece, Aug.–Sep. 2017, pp. 141–145.

[13] J. Chen, L. Tang, J. Liu, and J. Ye, "A convex formulation for learning shared structures from multiple tasks," in *Proc. ICML*, Montreal, QC, Canada, June, 2009, pp. 137–144.

[14] O. Chapelle, P. Shivaswmy, K. Q. Vadrevu, S. Weinberger, Y. Zhang, and B. Tseng, "Multi-task learning for boosting with applications to web search ranking," in *Proc. ACM SIGKDD*, Washington, DC, USA, Jul., 2010, pp. 1189–1198.

[15] J. Zhou, L. Yuan, J. Liu, and J. Ye, "A multi-task learning formulation for predicting disease progression," in *Proc. ACM SIGKDD*, San Diego, CA, USA, Aug., 2011, pp. 814–822.

[16] F. Cattivelli and A. H. Sayed, "Distributed nonlinear Kalman filtering with applications to wireless localization," in *Proc. IEEE ICASSP*, Dallas, TX, Mar., 2010, pp. 3522–3525.

[17] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin, *Network Flows: Theory, Algorithms, and Applications*, Prentice Hall, NJ, 1993.

[18] V. Kekatos and G. B. Giannakis, "Distributed robust power system state estimation," *IEEE Trans. Power Syst.*, vol. 28, no. 2, pp. 1617–1626, May 2013.

[19] S. A. Alghunaim and A. H. Sayed, "Decentralized exact coupled optimization," in *Proc. Allerton Conference on Communication, Control, and Computing*, Allerton, IL, October 2017, pp. 338–345.

[20] K. Yuan, B. Ying, X. Zhao, and A. H. Sayed, "Exact diffusion for distributed optimization and learning-part I: Algorithm development," submitted for publication, *also available on arXiv:1702.05122*, Feb. 2017.

[21] J. Mota, J. Xavier, P. Aguiar, and M. Puschel, "Distributed optimization with local domains: Application in MPC and network flows," *IEEE Trans. Autom. Contr.*, vol. 60, no. 7, pp. 2004–2009, July 2015.

[22] T. H. Summers and J. Lygeros, "Distributed model predictive consensus via the alternating directoin method of multipliers," in *Proc. Allerton Confrence on Communication, Control, and Computing*, Monticello, IL, USA, 2012, pp. 79–84.

[23] Y. Pu, M. N. Zeilinger, and C. N. Jones, "Inexact fast alternating minimization algorithm for distributed model predictive control," in *Proc. IEEE Conference on Decision and Control (CDC)*, 2014, pp. 5915–5921.

[24] R. Nassif, C. Richard, A. Ferrari, and A. H. Sayed, "Diffusion LMS for multitask problems with local linear equality constraints," *IEEE Trans. Signal Process*, vol. 65, no. 19, pp. 4979 – 4993, 2017.

[25] S. Y. Tu and A. H. Sayed, "Diffusion strategies outperform consensus strategies for distributed estimation over adaptive networks," *IEEE Trans. Signal Process.*, vol. 60, no. 12, pp. 6217–6234, 2012.

[26] Z. J. Towfic and A. H. Sayed, "Stability and performance limits of adaptive primal-dual networks," *IEEE Trans. Signal Process.*, vol. 63, no. 11, pp. 2888–2903, 2015.

[27] S. A. Alghunaim and A. H. Sayed, "Distributed coupled multi-agent stochastic optimization," submitted for publication, *also available on arXiv:1712.08817*, Dec. 2017.

[28] X. Zhao, J. Chen, and A. H. Sayed, "Beam coordination via diffusion adaptation over array networks," in *Proc. IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Cesme, Turkey, June, 2012, pp. 105–109.