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 overlapping 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.

Introduction

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 where effective algorithms have been developed for the distributed solution of constrained optimization problems of the form:

$$\underset{w}{\text{minimize}} \quad \sum_{k=1}^{N} J_k(w), \text{ subject to} \quad w \in \mathbb{W}_1 \cap \cdots \cap \mathbb{W}_N \tag{1}$$

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 distributed wireless localization [3], minimum-cost flow problems [1], and distributed power systems monitoring [4], 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.

Problem Formulation

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:

$$w_k \triangleq \text{col}\{w^\ell\}_{\ell \in \mathcal{I}_k} \in \mathbb{R}^{Q_k} \tag{2}$$
$$Q_k \triangleq \sum_{\ell \in \mathcal{I}_k} M_\ell. \tag{3}$$

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:

$$\begin{aligned} \underset{w}{\text{minimize}} \quad & J^{\text{glob}}(w) \triangleq \sum_{k=1}^{N} J_k(w_k) \\ \text{subject to} \quad & w \in \mathbb{W}_1 \cap \cdots \cap \mathbb{W}_N \end{aligned} \tag{4}$$

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

$$\mathbb{W}_k = \left\{ w : \begin{array}{l} h_{k,u}(w_k) = 0, \; u = 1, ..., U_k \\ g_{k,v}(w_k) \leq 0, \; v = 1, ..., V_k \end{array} \right. \tag{5}$$

where $\{h_{k,u}(\cdot), g_{k,v}(\cdot)\}$ are convex functions. Problem (4) is assumed to be feasible and therefore, a minimizer exists

$$w^o = \text{col}\{w^{1,o}, \cdots, w^{L,o}\} \triangleq \underset{w \in \mathbb{W}_1 \cap \cdots \cap \mathbb{W}_N}{\arg\min} J^{\text{glob}}(w) \tag{6}$$

It is clear that algorithms that solve (1) can be used to solve (4). 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 [2]) to lead to performance degradation. It is therefore necessary to solve problem (4) more directly and also more effectively.

Illustration Figure for the Problem Formulation

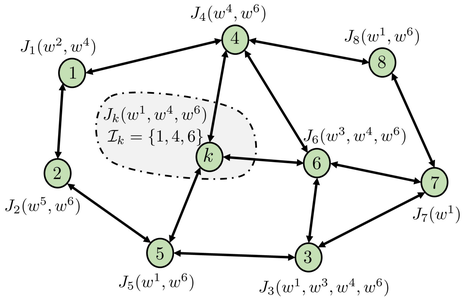

Figure: A connected network of agents where the local costs depend on different subsets of the global parameter vector $w = [w^1, w^2, w^3, w^4, w^5, w^6]$.

Penalized Formulation

We first relax problem (4) and replace it by the following penalized form parametrized by a scalar $\eta \geq 0$ ($\eta = 0$ for the unconstrained case):

$$\underset{w}{\text{minimize}} \quad J_\eta^{\text{glob}}(w) \triangleq \sum_{k=1}^{N} J_{k,\eta}(w_k) \tag{7}$$

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

$$J_{k,\eta}(w_k) \triangleq J_k(w_k) + \eta p_k(w_k) \tag{8}$$

with each penalty function in (8) given by

$$p_k(w_k) \triangleq \sum_{u=1}^{U_k} \delta^{\text{EP}}(h_{k,u}(w_k)) + \sum_{v=1}^{V_k} \delta^{\text{IP}}(g_{k,v}(w_k)) \tag{9}$$

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:

$$\delta^{\text{EP}}(x) = \begin{cases} 0, & x = 0 \\ > 0, & x \neq 0 \end{cases}, \quad \delta^{\text{IP}}(x) = \begin{cases} 0, & x \leq 0 \\ > 0, & \text{otherwise} \end{cases} \tag{10}$$

We denote the optimal solution of (7) by:

$$w^\star = \text{col}\{w^{1,\star}, \cdots, w^{L,\star}\} \triangleq \underset{w^1, \cdots, w^L}{\arg\min} J_\eta^{\text{glob}}(w) \tag{11}$$

Assumption 1. *(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:*

$$\nabla^2_{w_k} J_k(w_k) \leq \delta_k I_{Q_k} \tag{12}$$

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

$$\nabla^2_{w_{k_o}} J_{k_o}(w_{k_o}) > \nu_{k_o} I_{Q_{k_o}} \tag{13}$$

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

This assumption guarantees that the aggregate cost is strongly convex, and therefore a unique minimizer exists.

Assumption 2. *(Penalty functions): The penalty function $p_k(w_k)$ is twice-differentiable and its Hessian matrix is upper bounded:*

$$\nabla^2_{w_k} p_k(w_k) \leq \delta_{p,k} I_{Q_k} \tag{14}$$

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

Distributed Reformulation

In order to solve (7) 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 (3) using the local copies instead, namely, we now write

$$w_k \triangleq \text{col}\{w_k^\ell\}_{\ell \in \mathcal{I}_k} \in \mathbb{R}^{Q_k} \tag{15}$$

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

$$\mathcal{C}_\ell = \{k \mid \ell \in \mathcal{I}_k\} \tag{16}$$

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

$$w_k^\ell = w_s^\ell, \quad \forall \; k, s \in \mathcal{C}_\ell \tag{17}$$

Using relations (15) and (17), we can rewrite problem (7) as

$$\begin{aligned} \underset{w_1, ..., w_N}{\text{minimize}} \quad & J_\eta^{\text{glob}}(w_1, ..., w_N) \triangleq \sum_{k=1}^{N} J_{k,\eta}(w_k) \\ \text{subject to} \quad & w_k^\ell = w_s^\ell, \, \forall \, k, s \in \mathcal{C}_\ell, \forall \, \ell \end{aligned} \tag{18}$$

Coupled Diffusion Strategy

Algorithm 1 (Coupled diffusion strategy)

$$\boldsymbol{\zeta}_{k,i} = \boldsymbol{w}_{k,i-1} - \mu\eta \nabla_{w_k} p_k(\boldsymbol{w}_{k,i-1}) \tag{19a}$$
$$\boldsymbol{\psi}_{k,i} = \boldsymbol{\zeta}_{k,i} - \mu \nabla_{w_k} J_k(\boldsymbol{\zeta}_{k,i}) \tag{19b}$$
$$\boldsymbol{w}_{k,i}^\ell = \sum_{s \in \mathcal{N}_k \cap \mathcal{C}_\ell} a_{\ell,sk} \boldsymbol{\psi}_{s,i}^\ell, \; \forall \, \ell \in \mathcal{I}_k \tag{19c}$$

where $\{a_{\ell,sk}\}_{s,k \in \mathcal{C}_\ell}$ are combination weights that are chosen to satisfy:

$$\sum_{s \in \mathcal{C}_\ell} a_{\ell,sk} = 1, \quad \sum_{k \in \mathcal{C}_\ell} a_{\ell,sk} = 1 \tag{20}$$
$$a_{\ell,sk} \geq 0, \text{ and } a_{\ell,sk} = 0 \text{ if } s \notin \mathcal{N}_k \tag{21}$$

In steps (19a)–(19b), a traditional gradient-descent step is applied by each agent using the gradients of the corresponding risk and penalty functions. The last step (19c) 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.

Noise Model

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:

$$\widehat{\nabla_{w_k} J_k}(\boldsymbol{\zeta}_{k,i}) \triangleq \nabla_{w_k} J_k(\boldsymbol{\zeta}_{k,i}) - \boldsymbol{v}_{k,i}(\boldsymbol{\zeta}_{k,i}) \tag{22}$$

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

Assumption 3. *(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:*

$$\mathbb{E}\{\boldsymbol{v}_{k,i}(\boldsymbol{\zeta}_k) \mid \boldsymbol{\mathcal{F}}_i\} = 0 \tag{23}$$
$$\mathbb{E}\{\|\boldsymbol{v}_{k,i}(\boldsymbol{\zeta}_k)\|^2 \mid \boldsymbol{\mathcal{F}}_i\} \leq \bar{\alpha}_k \|\boldsymbol{\zeta}_k\|^2 + \bar{\sigma}_k^2 \tag{24}$$

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

Using (22), the coupled diffusion algorithm (19) becomes

$$\boldsymbol{\zeta}_{k,i} = \boldsymbol{w}_{k,i-1} - \mu\eta \nabla_{w_k} p_k(\boldsymbol{w}_{k,i-1}) \tag{25a}$$
$$\boldsymbol{\psi}_{k,i} = \boldsymbol{\zeta}_{k,i} - \mu \nabla_{w_k} J_k(\boldsymbol{\zeta}_{k,i}) + \mu \boldsymbol{v}_{k,i}(\boldsymbol{\zeta}_{k,i}) \tag{25b}$$
$$\boldsymbol{w}_{k,i}^\ell = \sum_{s \in \mathcal{N}_k \cap \mathcal{C}_\ell} a_{\ell,sk} \boldsymbol{\psi}_{s,i}^\ell, \; \forall \, \ell \in \mathcal{I}_k \tag{25c}$$

Network Model

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

$$A_\ell \triangleq [a_{\ell,sk}]_{s,k \in \mathcal{C}_\ell} \tag{26}$$

Assumption 4. *(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.* ■

Convergence Result

Theorem 1. *(Mean-square convergence): If w^o is a regular[a] point for the constraints, then, under Assumptions 4–3, the coupled diffusion algorithm (25) converges for sufficiently small step-sizes μ. Moreover, for every agent k, it holds that:*

$$\limsup_{i \to \infty} \mathbb{E}\|w^{\ell,\star} - \boldsymbol{w}_{k,i}^\ell\|^2 \leq O(\mu) + O(\mu^2 \eta^2), \; \forall \; \ell \in \mathcal{I}_k \tag{27}$$

■

[a] 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 $w_k^o = \text{col}\{w^{\ell,o}\}_{\ell \in \mathcal{I}_k}$).

Simulation Results

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:

$$\boldsymbol{d}_k(i) + \boldsymbol{u}_{k,i} w^{k,\bullet} + \boldsymbol{v}_k(i) \tag{28}$$

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 goal of the network is to solve the following problem:

$$\min_{w^1, ..., w^N} \sum_{k=1}^{N} \mathbb{E}|\boldsymbol{d}_k(i) - \boldsymbol{u}_{k,i} w^k|^2, \quad \text{s.t} \quad \sum_{s \in \mathcal{N}_k} B_{sk} w^s = b_k, \; \forall \; k \tag{29}$$

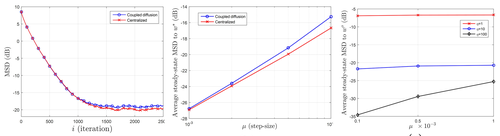

Figure: (a) MSD learning curve for $\mu = 0.005$ and $\eta = 10$. (b) MSD for different values of step size μ with $\eta = 10$. (c) Average steady-state error to w^o for different values of η and μ.

References

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

[2] S. A. Alghunaim, K. Yuan, and A. H. Sayed. Decentralized exact coupled optimization. In *Proc. Allerton Conference on Communication, Control, and Computing*, pages 338–345, Allerton, IL, October 2017.

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

[4] V. Kekatos and G. B. Giannakis. Distributed robust power system state estimation. *IEEE Trans. Power Syst.*, 28(2):1617–1626, May 2013.

Acknowledgements and Contact Information

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