Journal of Machine Learning Research 20 (2019) 1-62 Submitted 9/17; Revised 8/19; Published 9/19

Decentralized Dictionary Learning Over Time-Varying Digraphs

Amir Daneshmand adaneshm@purdue.edu
Ying Sun sun578@purdue.edu
Gesualdo Scutari gscutari@purdue.edu
School of Industrial Engineering
Purdue University
West-Lafayette, IN, USA

Francisco Facchinei facchinei@diag.uniroma1.it
Department of Computer, Control, and Management Engineering
University of Rome "La Sapienza"
Rome, Italy

Brian M. Sadler brian.m.sadler6.civ@mail.mil
U.S. Army Research Laboratory
Adelphi, MD, USA

Editor: Sathiya Keerthi

Abstract

This paper studies Dictionary Learning problems wherein the learning task is distributed over a multi-agent network, modeled as a time-varying directed graph. This formulation is relevant, for instance, in Big Data scenarios where massive amounts of data are collected/stored in different locations (e.g., sensors, clouds) and aggregating and/or processing all data in a fusion center might be inefficient or unfeasible, due to resource limitations, communication overheads or privacy issues. We develop a unified decentralized algorithmic framework for this class of *nonconvex* problems, which is proved to converge to stationary solutions at a sublinear rate. The new method hinges on Successive Convex Approximation techniques, coupled with a decentralized tracking mechanism aiming at locally estimating the gradient of the smooth part of the sum-utility. To the best of our knowledge, this is the first provably convergent decentralized algorithm for Dictionary Learning and, more generally, bi-convex problems over (time-varying) (di)graphs.

Keywords: Decentralized algorithms, dictionary learning, directed graph, non-convex optimization, time-varying network

1. Introduction and Motivation

This paper introduces, analyzes, and tests numerically the first *provably convergent distributed method* for a fairly general class of Dictionary Learning (DL) problems. More specifically, we study the problem of finding a matrix $\mathbf{D} \in \mathbb{R}^{M\times K}$ (a.k.a. the dictionary), by which the data matrix $\mathbf{S} \in \mathbb{R}^{M\times N}$ can be represented through a matrix $\mathbf{X} \in \mathbb{R}^{K\times N}$, with a favorable structure on $\mathbf{D}$ and $\mathbf{X}$ (e.g., sparsity). We target scenarios where computational resources and data are not centrally available, but distributed over a group of I agents, which can communicate through a (possibly) *time-varying, directed* network; see Fig. 1.

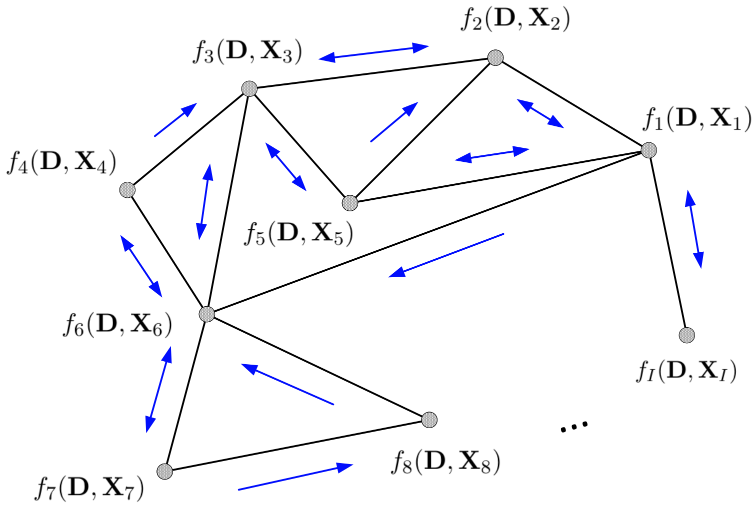


Figure 1: Directed network topology

Each agent $i \in \{1, 2, \ldots, I\}$ owns one block $\mathbf{S}_i \in \mathbb{R}^{M\times n_i}$ of the data $\mathbf{S} \triangleq [\mathbf{S}_1, \ldots, \mathbf{S}_I]$, with $\sum_{i=1}^{I} n_i = N$. Partitioning the representation matrix $\mathbf{X} \triangleq [\mathbf{X}_1, \ldots, \mathbf{X}_I]$ according to $\mathbf{S}$, with $\mathbf{X}_i \in \mathbb{R}^{K\times n_i}$, the class of *distributed* DL problems we aim at studying reads

$$\begin{aligned} \min_{\mathbf{D}, \{\mathbf{X}_i\}_{i=1}^{I}} \quad & U(\mathbf{D}, \mathbf{U}) \triangleq \underbrace{\sum_{i=1}^{I} f_i(\mathbf{D}, \mathbf{X}_i)}_{\triangleq F(\mathbf{D}, \mathbf{X})} + \sum_{i=1}^{I} g_i(\mathbf{X}_i) + G(\mathbf{D}) \\ \text{s.t.} \quad & \mathbf{D} \in \mathcal{D}, \ \mathbf{X}_i \in \mathcal{X}_i, \quad i = 1, \ldots, I, \end{aligned} \tag{P}$$

where $f_i : \mathcal{D} \times \mathcal{X}_i \to \mathbb{R}$ is the fidelity function of agent i, which measures the mismatch between the data $\mathbf{S}_i$ and the (local) model; this function is assumed to be smooth and *biconvex* (i.e., convex in $\mathbf{D}$ for fixed $\mathbf{X}_i$, and vice versa); $G : \mathcal{D} \to \mathbb{R}$ and $g_i : \mathcal{X}_i \to \mathbb{R}$ are (possibly non-smooth) convex functions, which are generally used to impose extra structure on the solution (e.g., low-rank or sparsity); and $\mathcal{D} \subseteq \mathbb{R}^{M\times K}$ and $\mathcal{X}_i \subseteq \mathbb{R}^{K\times n_i}$ are some closed convex sets. To avoid scaling ambiguity in the model, $\mathcal{D}$ is assumed to be bounded, without loss of generality. Since all f_i's share the common variable $\mathbf{D}$, we call it a *shared* variable and, by the same token, $\mathbf{X}_i$'s are termed *private* variables. Note that, in this distributed setting, agent i knows only its own functions f_i (and g_i) but not $\sum_{j\neq i} f_j$. Hence, agents aim to cooperatively solve Problem P leveraging local communications with their neighbors.

Problem P encompasses several DL-based formulations of practical interest, corresponding to different choices of the fidelity functions, regularizers, and feasible sets; examples include the *elastic net* (Zou and Hastie, 2005) sparse DL, sparse PCA (Shen and Huang, 2008), non-negative matrix factorization and low-rank approximation (Hastie et al., 2015), supervised DL (Mairal et al., 2008), sparse singular value decomposition (Lee et al., 2010), non-negative sparse coding (Hoyer, 2004), principal component pursuit (Candès et al., 2011), robust non-negative sparse matrix factorization, and discriminative label consistent learning

(Jiang et al., 2011). More details on explicit customizations of the general model P can be found in Sec. 2.

Our distributed setting is motivated by several data-intensive applications in several fields, including signal processing and machine learning, and network systems (such as clouds, cluster computers, networks of sensor vehicles, or autonomous robots) wherein the sheer volume and spatial/temporal disparity of scattered data, energy constraints, and/or privacy issues, render centralized processing and storage infeasible or inefficient. Also, time-varying communications arise, for instance, in mobile wireless networks (e.g., ad-hoc networks), wherein nodes are mobile and/or communicate through fading channels. Moreover, since nodes generally transmit at different power and/or communication channels are not symmetric, directed links are a natural assumption.

Our goal is to design a provably convergent *decentralized* method for Problem P, over *time-varying* and *directed* graphs. To the best of our knowledge this is an open problem, as documented next.

1.1. Challenges and related works

The design of distributed algorithms for P faces the following challenges:

(i) Problem P is *non-convex* and *non-separable* in the optimization variables;

(ii) Each agent i owns exclusively $\mathbf{S}_i$ and thus can only compute its own function f_i;

(iii) Each f_i depends on a common set of variables−the dictionary $\mathbf{D}$−shared among all the agents, as well as the private variables $\mathbf{X}_i$. Shared and private variables need to be treated differently. In fact, in several applications, the size of private variables is much larger than that of the shared ones; hence, broadcasting agents' private variables over the network would result in an unaffordable communication overhead;

(iv) The gradient of each f_i is in general *neither bounded nor globally Lipschitz* on the feasible region. This represents a challenge in the design of provably convergent distributed algorithms, as boundedness and Lipschitzianity of the gradient are standard assumptions in the analysis of most distributed schemes for nonconvex problems;

(v) G and g_i's are *nonsmooth*;

(vi) The graph is *directed*, *time-varying*; no other structure is assumed (such as star or ring topology, etc.), but some long term connectivity properties (cf. Assumption B).

Centralized methods for the solution of Problem P (or some closely related variants) have been extensively studied and prominent examples are (Aharon et al., 2006; Mairal et al., 2010; Razaviyayn et al., 2014b). However, we are not aware of any *distributed* algorithm that can address challenges i)-vi) (even some subsets of them), as documented next.

Ad-hoc heuristics: Several attempts have been made to extend centralized approaches for DL problems to a distributed setting (*undirected, static* graphs), under more or less restrictive assumptions; examples include primal methods (Raja and Bajwa, 2013; Chainais and Richard, 2013; Wai et al., 2015) and (primal/)dual-based ones (Chen et al., 2015; Liang et al., 2014; Chouvardas et al., 2015). While these schemes represent good heuristics, their theoretical convergence remains an open question, and numerical results are contradictory. For instance, some schemes are shown not to converge while some others fail to reach asymptotic agreement among the local copies of the dictionary; see, e.g. (Chainais and Richard, 2013).

Recently and independently from our conference work (Daneshmand et al., 2016), Zhao et al. (2016) proposed a distributed primal-dual-based method for a class of dictionary learning problems related, but different from Problem P. Specifically, they considered: quadratic loss functions f_i, with a quadratic regularization on the dictionary (i.e., $G = 0$), and norm ball constraints on the private variables. The network is modeled as a fixed undirected graph. Asymptotic convergence of the scheme to stationary solutions is proved, but no rate analysis is reported. We remark that the scheme in (Zhao et al., 2016), in order to establish convergence, requires some penalty parameters to go to infinity, which makes the method numerically not attractive.

Distributed nonconvex optimization: Since the DL problem P is an instance of nonconvex optimization problems, we briefly discuss here the few works in the literature on distributed methods for non-convex optimization (Bianchi and Jakubowicz, 2013; Tatarenko and Touri, 2017; Wai et al., 2017; Di Lorenzo and Scutari, 2016; Sun et al., 2016; Hong et al., 2017; Scutari and Sun, 2019); we group these papers as follows. The schemes in (Bianchi and Jakubowicz, 2013; Tatarenko and Touri, 2017; Wai et al., 2017; Hong et al., 2017), while substantially different, are all applicable to *smooth, unconstrained* optimization, with (Bianchi and Jakubowicz, 2013; Wai et al., 2017) handling also *compact* constraints and (Tatarenko and Touri, 2017) implementable on (time-varying) digraphs. The distributed algorithms in (Di Lorenzo and Scutari, 2016; Sun et al., 2016; Scutari and Sun, 2019) can handle objectives with additive *nonsmooth* convex functions, with (Sun et al., 2016; Scutari and Sun, 2019) applicable to (time-varying) digraphs.

All the above schemes *cannot* adequately deal with *private* (i.e., $\mathbf{X}_i$'s) *and* shared variables (i.e., $\mathbf{D}$), which are a key feature of Problem P. Furthermore, convergence therein is proved under the assumption that the gradient of (the smooth part of) the objective function is *globally Lipschitz continuous*, a property that we do not assume and that is not satisfied in many of the applications we consider. The design of provably convergent distributed algorithms for P remains an open problem, let alone rate guarantees.

1.2. Major contributions

In this paper, we propose the *first provably convergent* distributed algorithm for the general class of DL problems P, addressing all challenges i)-vi). The proposed approach uses a general convexification-decomposition technique that hinges on recent (centralized) Successive Convex Approximation methods (Scutari et al., 2014; Facchinei et al., 2015). This technique is coupled with a perturbed push-sum consensus scheme preserving the *feasibility* of the iterates and a tracking mechanism aiming at estimating locally the gradient of $\sum_i f_i$. Both communication and tracking protocols are implementable on *time-varying* undirected or *directed* graphs (B-strongly connected). The scheme is proved to converge to stationary solutions of Problem P, under mild assumptions on the step-size employed by the algorithm; a sublinear convergence rate is also established. On the technical side, we contribute to the literature of distributed algorithms for bi-convex (nonsmooth) constrained optimization by putting forth a new non-trivial convergence analysis that, for the first time, i) avoids the assumption that the gradients ∇f_i are globally Lipschitz; and ii) deals with *private* and shared optimization variables. Numerical experiments show that the proposed schemes compare favorably with ad-hoc algorithms, proposed for special instances of Problem P.

1.3. Paper Organization

The rest of the paper is organized as follows. The problem and network setting are introduced in Sec. 2, along with some motivating applications. Sec. 3 presents the algorithm and its convergence properties; the proofs of our results are given in the Appendix, Sec. A. Extensive numerical experiments showing the effectiveness of the proposed scheme are discussed in Sec. 5 whereas Sec. 6 draws some conclusions.

1.4. Notation

Throughout the paper we use the following notation. We denote by $\mathbb{R}^n_+$ and $\mathbb{N}_+$ the non-negative orthant and the set of non-negative integers, respectively. Given $x \in \mathbb{R}$, $\lceil x \rceil$ (resp. $\lfloor x \rfloor$) denotes the smallest (resp. the largest) integer greater (resp. smaller) than or equal to x. Vectors are denoted by bold lower-case letters (e.g., $\mathbf{x}$) whereas matrices are denoted by bold capital letters (e.g., $\mathbf{A}$). The k-th canonical vector is denoted by $\mathbf{e}_k$. The inner product between two real matrices, $\mathbf{A}$ and $\mathbf{B}$, is denoted by $\langle \mathbf{A}, \mathbf{B} \rangle \triangleq \mathrm{tr}(\mathbf{A}^\intercal \mathbf{B})$, where $\mathrm{tr}(\bullet)$ is the trace operator; $\mathbf{A} \otimes \mathbf{B}$ denotes the Kronecker product. Given the real matrix $\mathbf{A}$, with ij-entries denoted by A_{ij}, we will use the following matrix norms: the Frobenius norm $||\mathbf{A}||_F \triangleq \sqrt{\sum_{i,j} |A_{ij}|^2}$; the $L_{1,1}$ norm $||\mathbf{A}||_{1,1} \triangleq \sum_{i,j} |A_{ij}|$; the $L_{2,\infty}$ norm $||\mathbf{A}||_{2,\infty} \triangleq \max_i \sqrt{\sum_j A_{ij}^2}$; the $L_{\infty,\infty}$ norm $||\mathbf{A}||_{\infty,\infty} = \max_{i,j} |A_{ij}|$; and the spectral norm $||\mathbf{A}||_2 \triangleq \sigma_{\max}(\mathbf{A})$, where $\sigma_{\max}(\mathbf{A})$ denotes the maximum singular value of $\mathbf{A}$. The matrix quantities $\nabla_D f_i(\mathbf{D}, \mathbf{X}_i)$ and $\nabla_{X_i} f_i(\mathbf{D}, \mathbf{X}_i)$ are the gradients of f_i with respect to $\mathbf{D}$ and $\mathbf{X}_i$, evaluated at $(\mathbf{D}, \mathbf{X}_i)$, respectively, with the partial derivatives arranged according to the patterns of $\mathbf{D}$ and $\mathbf{X}_i$, respectively. The same convention is adopted for subgradients of g_i and G, that are therefore written as matrices of the same dimensions of $\mathbf{X}_i$ and $\mathbf{D}$, respectively. Table 1 summarizes the main notation and symbols used in the paper.

Because of the nonconvexity of Problem P, we aim at computing stationary solutions of P, defined as follows: a tuple $(\mathbf{D}^*, \mathbf{X}^*)$, with $\mathbf{X}^* \triangleq [\mathbf{X}_1^*, \ldots, \mathbf{X}_I^*]$ is a stationarity solution of P if the following holds: $\mathbf{D}^* \in \mathcal{D}$, $\mathbf{X}_i^* \in \mathcal{X}_i$, $i = 1, \ldots, I$, and

$$\begin{aligned} &\Big\langle \nabla_D F(\mathbf{D}^*, \mathbf{X}^*), \mathbf{D} - \mathbf{D}^* \Big\rangle + G(\mathbf{D}) - G(\mathbf{D}^*) \geq 0, && \forall \mathbf{D} \in \mathcal{D}, \\ &\langle \nabla_{X_i} f_i(\mathbf{D}^*, \mathbf{X}_i^*), \mathbf{X}_i - \mathbf{X}_i^* \rangle + g_i(\mathbf{X}_i) - g_i(\mathbf{X}_i^*) \geq 0, && \forall \mathbf{X}_i \in \mathcal{X}_i,\ i = 1, \ldots, I. \end{aligned} \tag{2}$$

2. Problem Setup and Motivating Examples

In this section, we first discuss the assumptions underlying our model and then provide several examples of possible applications.

Symbol	Definition	Member of	Reference
$F(\mathbf{D},\mathbf{X})$	$\sum_{i=1}^{I} f_i(\mathbf{D},\mathbf{X}_i)$	$\mathbb{R}^{M\times K}\times\mathbb{R}^{K\times N}\to\mathbb{R}$	(P)
$U(\mathbf{D},\mathbf{X})$	$F(\mathbf{D},\mathbf{X})+\sum_{i=1}^{I} g_i(\mathbf{X}_i)+G(\mathbf{D})$	$\mathbb{R}^{M\times K}\times\mathbb{R}^{K\times N}\to\mathbb{R}$	(P)
$\mathbf{S}_i$	Local data matrix	$\mathbb{R}^{M\times n_i}$	
$\mathbf{S}$	$[\mathbf{S}_1,\mathbf{S}_2,\ldots,\mathbf{S}_I]$	$\mathbb{R}^{M\times N}$	
$\mathbf{D}$	Dictionary matrix variable	$\mathcal{D}\subseteq\mathbb{R}^{M\times K}$	
$\mathbf{D}_{(i)}$	Local copy of $\mathbf{D}$ of agent i	$\mathcal{D}\subseteq\mathbb{R}^{M\times K}$	
$\mathbf{D}_{(i)}^{\nu}$	$\mathbf{D}_{(i)}$ at iteration ν	$\mathcal{D}\subseteq\mathbb{R}^{M\times K}$	
$\mathbf{D}^{\nu}$	$[\mathbf{D}_{(1)}^{\nu\intercal},\mathbf{D}_{(2)}^{\nu\intercal},\ldots,\mathbf{D}_{(I)}^{\nu\intercal}]^{\intercal}$	$\mathbb{R}^{M\times KI}$	(25)
$\widetilde{\mathbf{D}}_{(i)}^{\nu}$	Solution of subproblem (8)	$\mathcal{D}\subseteq\mathbb{R}^{M\times K}$	(8)
$\overline{\mathbf{D}}^{\nu}$	$(1/I)\sum_{i=1}^{I}\mathbf{D}_{(i)}^{\nu}$	$\mathcal{D}\subseteq\mathbb{R}^{M\times K}$	(25)
$\mathbf{U}_{(i)}^{\nu}$	Local update of dictionary variable	$\mathcal{D}\subseteq\mathbb{R}^{M\times K}$	(10)
$\mathbf{X}_i$	Local matrix variable	$\mathcal{X}_i\subseteq\mathbb{R}^{K\times n_i}$	
$\mathbf{X}$	$[\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_I]$	$\mathcal{X}\subseteq\mathbb{R}^{K\times N}$	
$\mathbf{X}_i^{\nu}$	$\mathbf{X}_i$ at iteration ν	$\mathcal{X}_i\subseteq\mathbb{R}^{K\times n_i}$	
$\mathbf{X}^{\nu}$	$\mathbf{X}$ at iteration ν : $[\mathbf{X}_1^{\nu},\mathbf{X}_2^{\nu},\ldots,\mathbf{X}_I^{\nu}]$	$\mathcal{X}\subseteq\mathbb{R}^{K\times N}$	(25)
$\mathbf{\Theta}_{(i)}^{\nu}$	Gradient-tracking variable	$\mathcal{D}\subseteq\mathbb{R}^{M\times K}$	(13)
$\mathbf{A}^{\nu}$	$(a_{ij}^{\nu})_{i,j=1}^{I}-$ consensus weights at time ν	$\mathbb{R}^{I\times I}$	Assumption F

Table 1: Table of notation

2.1. Problem Assumptions

We consider Problem P under the following assumptions.

Assumption A (On Problem P)

(A1) *Each* $f_i:\mathcal{O}\times\mathcal{O}_i\to\mathbb{R}$ *is* $\mathcal{C}^2$*, lower bounded, and biconvex, where* $\mathcal{O}\supseteq\mathcal{D}$ *and* $\mathcal{O}_i\supseteq\mathcal{X}_i$ *are convex open sets;*

(A2) *Given* $\mathbf{D}\in\mathcal{D}$*, each* $\nabla_{X_i}f_i(\mathbf{D},\bullet)$ *is Lipschitz continuous on* $\mathcal{X}_i$*, with Lipschitz constant* $L_{\nabla X_i}(\mathbf{D})$*. Furthermore, each* $L_{\nabla X_i}:\mathcal{D}\to\mathbb{R}_+$ *is continuous;*

(A3) $\mathcal{D}$ *is compact and convex; and each* $\mathcal{X}_i$ *is closed and convex (not necessarily bounded);*

(A4) $G:\mathcal{O}\to\mathbb{R}$ *is convex (possibly non-smooth);*

(A5) *For all* $i=1,\ldots,I$*, either i)* $\mathcal{X}_i$ *is compact and* $g_i:\mathcal{O}_i\to\mathbb{R}$ *is convex; or ii)* g_i *is* μ_i*-strongly convex.*

The above assumptions are quite mild and are satisfied by several problems of practical interest; see Sec. 2.2 for several concrete examples.

Network topology

We study Problem P in the following network setting. Time is slotted and in each time-slot ν the network of the I agents is modeled as a digraph $\mathcal{G}^{\nu}=(\mathcal{V},\mathcal{E}^{\nu})$, where $\mathcal{V}=\{1,\ldots,I\}$ is the set of agents and $\mathcal{E}^{\nu}$ is the set of edges (communication links); we use $(i,j)\in\mathcal{E}^{\nu}$ to indicate that there is a directed link from node i to node j. The *in-neighborhood* of agent

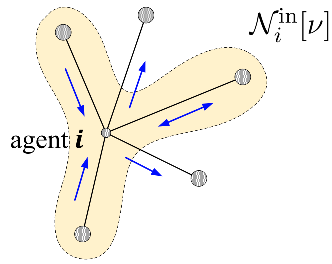


Figure 2: Illustration of in-neighborhood set of agent i at time ν.

$i \in \mathcal{V}$ at time ν is defined as $\mathcal{N}_i^{\text{in}}[\nu] = \{j \in \mathcal{V} | (j,i) \in \mathcal{E}^\nu\} \cup \{i\}$ (see Fig. 2) whereas its *out-neighborhood* is $\mathcal{N}_i^{\text{out}}[\nu] = \{j \in \mathcal{V} | \, (i,j) \in \mathcal{E}^\nu\} \cup \{i\}$. In words, agent i can receive information from its in-neighborhood members, and send information to its out-neighbors. The *out-degree* of agent i is defined as $d_i^\nu \triangleq \left|\mathcal{N}_i^{\text{out}}[\nu]\right|$, where $|\bullet|$ denotes the cardinality of a set. If the graph is undirected, the set of in-neighbors and out-neighbors coincide; in such a case we just write $\mathcal{N}_i$ to denote the set of neighbors of agent i. When the network is static, all the above quantities do not depend on the iteration index ν; hence, we will drop the superscript "ν". To let information propagate over the network, we assume that the sequence $\{\mathcal{G}^\nu\}_\nu$ possesses some "long-term" connectivity property, as stated next.

Assumption B (B-strong connectivity) *The graph sequence $\{\mathcal{G}^\nu\}_\nu$ is B-strongly connected, i.e., there exists an (arbitrarily large) integer $B > 0$ (unknown to the agents) such that the graph with edge set $\cup_{t=kB}^{(k+1)B-1} \mathcal{E}^t$ is strongly connected, for all $k \geq 0$.*

Notice that this condition is quite mild and widely used in the literature to analyze convergence of distributed algorithms over time-varying networks. Generally speaking, it permits strong connectivity to occur over time windows of length B, so that information can propagate from every node to every other node in the network. Assumption B is satisfied in several practical scenarios. For instance, commonly used settings in cloud computing infrastructures are star, ring, tree, hypercube, or n-dimensional mesh (Torus) topologies, which all satisfy Assumption B. It is worth mentioning that the multi-hop network topologies of these structures are migrating towards *high-radix mesh and Torus*, since they are scalable, low-energy consuming, and much cheaper than other topologies, like *fat-tree* topologies (Kim, 2008). These type of connected networks are generally time-invariant and undirected, and clearly they satisfy Assumption B.

2.2. Motivating examples

We conclude this section discussing some practical instances of Problem P, all satisfying Assumption A, which show the generality of the proposed model.

Elastic net sparse DL (Tosic and Frossard, 2011; Zou and Hastie, 2005)

Sparse approximation of a signal with an adaptive dictionary is one of the most studied DL problems (Tosic and Frossard, 2011). When an elastic net sparsity-inducing regularizer is used (Zou and Hastie, 2005), the problem can be written as

$$\begin{aligned}
&\min_{\mathbf{D},\{\mathbf{X}_i\}_{i=1}^I} \quad \sum_{i=1}^{I} \left\{ \frac{1}{2} \left\| \mathbf{S}_i - \mathbf{D}\mathbf{X}_i \right\|_F^2 + \lambda \left\| \mathbf{X}_i \right\|_{1,1} + \frac{\mu}{2} \left\| \mathbf{X}_i \right\|_F^2 \right\} \\
&\quad \text{s.t.} \qquad \mathbf{D} \in \mathcal{D},\ \mathbf{X}_i \in \mathbb{R}^{K \times n_i}, \quad i = 1, 2, \ldots, I,
\end{aligned} \tag{3}$$

where $\mathcal{D} \triangleq \{\mathbf{D} : ||\mathbf{D}\mathbf{e}_k||_2 \leq \alpha,\, k = 1, 2, \ldots, K\}$, and α, λ, $\mu > 0$ are the tuning parameters. Problem (3) is an instance of P, with $f_i(\mathbf{D}, \mathbf{X}_i) = (1/2) \cdot \|\mathbf{S}_i - \mathbf{D}\mathbf{X}_i\|_F^2$, $g_i(\mathbf{X}_i) = \lambda \|\mathbf{X}_i\|_{1,1} + \frac{\mu}{2} \|\mathbf{X}_i\|_F^2$, $G(\mathbf{D}) = 0$, and $\mathcal{X}_i = \mathbb{R}^{K \times n_i}$. It is not difficult to check that (3) satisfies Assumption A, and the Lipschitz constant in A2 is given by $L_{\nabla X_i}(\mathbf{D}) = (\sigma_{\max}(\mathbf{D}))^2$.

Supervised DL (Mairal et al., 2008)

Consider a classification problem with training set $\{\mathbf{s}_n, y_n\}_{n=1}^N$, where $\mathbf{s}_n$ is the feature vector with associated binary label y_n. The discriminative DL problem aims at simultaneously learning a dictionary $\mathbf{D}^{(1)} \in \mathbb{R}^{M \times K}$ such that $\mathbf{s}_n = \mathbf{D}^{(1)}\mathbf{x}_n$, for some sparse $\mathbf{x}_n \in \mathbb{R}^K$, and finding a bilinear classifier $\zeta_n(\mathbf{D}^{(2)}, \mathbf{x}_n, \mathbf{s}_n) \triangleq \mathbf{s}_n^\intercal \mathbf{D}^{(2)} \mathbf{x}_n$ that best separates the coded data with distinct labels (Mairal et al., 2008). Assume that each agent i owns $\{(\mathbf{s}_n, y_n) : n \in \mathcal{S}_i\}$, with $\{\mathcal{S}_i\}_{i=1}^I$ being a partition of $\{1, \ldots, N\}$, then the discriminative DL reads

$$\begin{aligned}
&\min_{\substack{\mathbf{D}^{(1)},\mathbf{D}^{(2)} \\ \{\mathbf{x}_n\}_{n=1}^N}} \quad \sum_{i=1}^{I} \sum_{n \in \mathcal{S}_i} \left[\ell\left(y_n \zeta_n \left(\mathbf{D}^{(2)}, \mathbf{x}_n, \mathbf{s}_n \right) \right) + \frac{1}{2} \left\| \mathbf{s}_n - \mathbf{D}^{(1)}\mathbf{x}_n \right\|_2^2 + g_n(\mathbf{x}_n) \right] \\
&\quad \text{s.t.} \qquad \mathbf{D}^{(1)} \in \mathcal{D}^{(1)},\ \mathbf{D}^{(2)} \in \mathcal{D}^{(2)},\ \mathbf{x}_n \in \mathbb{R}^K, \quad n = 1, 2, \ldots, N,
\end{aligned} \tag{4}$$

where $\ell(x) \triangleq \log(1 + e^{-x})$ is the *logistic* loss function; and $g_n(\mathbf{x}_n) \triangleq \lambda \|\mathbf{x}_n\|_1 + (\mu/2) \cdot \|\mathbf{x}_n\|_2^2$ is the elastic net regularizer. The dictionary $\mathbf{D}^{(1)}$ and classifier parameter $\mathbf{D}^{(2)}$ are constrained to belong to the convex compact sets $\mathcal{D}^{(1)}$ and $\mathcal{D}^{(2)}$, respectively. Problem (4) is an instance of Problem P, with $\mathbf{D} \triangleq \left[\mathbf{D}^{(1)}, \mathbf{D}^{(2)}\right]$, $\mathbf{S}_i \triangleq [\mathbf{s}_n]_{n \in \mathcal{S}_i}$ and $\mathbf{X}_i \triangleq [\mathbf{x}_n]_{n \in \mathcal{S}_i}$. Note that Assumption A is satisfied, and the Lipschitz constant in A2 is given by $L_{\nabla X_i}(\mathbf{D}) = (1/4) \cdot \|y_i \cdot \mathbf{s}_i^\intercal \mathbf{D}^{(2)}\|_2^2 + (\sigma_{\max}(\mathbf{D}^{(1)}))^2$.

DL for low-rank plus sparse representation (Bouwmans et al., 2017)

The low-rank plus sparse decomposition problems cover many applications in signal processing and machine learning (Bouwmans et al., 2017), including matrix completion, image denoising, deblurring, superresolution, and Principal Component Pursuit (PCP) (Candès et al., 2011). Consider the bi-linear model $\mathbf{S} \approx \mathbf{L} + \mathbf{H}\mathbf{Q}\mathbf{U}$: the data matrix $\mathbf{S}$ is decomposed as the superposition of a low-rank matrix $\mathbf{L}$ (capturing the correlations among data) and $\mathbf{H}\mathbf{Q}\mathbf{U}$, where $\mathbf{Q} \in \mathbb{R}^{M \times \tilde{K}}$ is an over-complete dictionary (capturing the representative modes of the data), $\mathbf{U} \in \mathbb{R}^{\tilde{K} \times N}$ is a sparse matrix (representing the data parsimoniously), and $\mathbf{H}$ is a given *degradation* matrix, which accounts for tasks such as denoising, superresolution, and deblurring. To enforce $\mathbf{L}$ to be low-rank, we employ the nuclear norm $\|\mathbf{L}\|_*$ regularizer, which can be equivalently rewritten as $\|\mathbf{L}\|_* = \inf\{\frac{1}{2} \|\mathbf{P}\|_F^2 + \frac{1}{2} \|\mathbf{V}\|_F^2 : \mathbf{L} = \mathbf{P}\mathbf{V}\}$, where $\mathbf{P} \in \mathbb{R}^{M \times L}$, $\mathbf{V} \in \mathbb{R}^{L \times N}$, and $L \ll \min(M, N)$ (Srebro and Shraibman, 2005; Recht et al., 2010). Partitioning $\mathbf{V}$ and $\mathbf{U}$ according to $\mathbf{S}$, i.e., $\mathbf{S}_i = \mathbf{P}\mathbf{V}_i + \mathbf{H}\mathbf{Q}\mathbf{U}_i$, the problem reads

$$\begin{aligned} \min_{\mathbf{P},\mathbf{Q},(\mathbf{V}_i,\mathbf{U}_i)_{i=1}^{I}} \quad & \sum_{i=1}^{I} \left[\frac{1}{2} \left\| \mathbf{S}_i - [\mathbf{P}\ \mathbf{H}\mathbf{Q}] \begin{bmatrix} \mathbf{V}_i \\ \mathbf{U}_i \end{bmatrix} \right\|_F^2 \right. \\ & \qquad \left. + \frac{\zeta}{2I} \left(\|\mathbf{P}\|_F^2 + I \cdot \|\mathbf{V}_i\|_F^2 \right) + \lambda \|\mathbf{X}_i\|_{1,1} + \frac{\mu}{2} \|\mathbf{X}_i\|_F^2 \right] \\ \text{s.t.} \quad & \mathbf{D} \in \mathcal{D}, \quad \mathbf{X}_i \in \mathbb{R}^{(L+\tilde{K}) \times n_i}, \quad i = 1, 2, \ldots, I, \end{aligned} \tag{5}$$

where $\mathcal{D}$ is some compact set; $\zeta > 0$ is a constant used to promote the low-rank structure on $\mathbf{L}$ while sparsity on $\mathbf{X}$ is enforced by the elastic net regularization, with constants λ, $\mu > 0$. Problem (5) is clearly an instance of Problem P wherein f_i is the quadratic loss, and $[\mathbf{P}, \mathbf{Q}]$ and $[\mathbf{V}_i^\intercal, \mathbf{U}_i^\intercal]^\intercal$ are the shared and private variables ($K = L + \tilde{K}$), respectively. Assumption A is satisfied, and the Lipschitz constant in A2 is given by $L_{\nabla X_i}(\mathbf{D}) = (\sigma_{\max}(\mathbf{PHQ}))^2$.

A variant of this problem, which still is a particular case of Problem (5), is obtained by replacing the quadratic loss function with the smoothed Huber function to achieve robustness against outliers (Aravkin et al., 2014).

Sparse SVD/PCA (Lee et al., 2010; Udell et al., 2016; Mairal et al., 2010)

Computing the SVD of a set of data with sparse singular vectors (Sparse SVD) is the foundation of many applications in multivariate analysis, e.g., biclustering (Lee et al., 2010). As proposed in (Mairal et al., 2010), Problem P can be used to accomplish this task by imposing sparsity on the factors $\mathbf{D}$ and $\mathbf{X}$ of $\mathbf{S}$. More specifically, we have

$$\begin{aligned} \min_{\mathbf{D},(\mathbf{X}_i)_{i=1}^{I}} \quad & \sum_{i=1}^{I} \left\{ \frac{1}{2} \ \|\mathbf{S}_i - \mathbf{D}\mathbf{X}_i\|_F^2 + \lambda_X \|\mathbf{X}_i\|_{1,1} + \frac{\mu_X}{2} \|\mathbf{X}_i\|_F^2 \right\} + \lambda_D \|\mathbf{D}\|_{1,1} + \frac{\mu_D}{2} \|\mathbf{D}\|_F^2 \\ \text{s.t.} \quad & \mathbf{D} \in \mathcal{D} \triangleq \{\mathbf{D} \in \mathbb{R}^{M \times K} \ : \ ||\mathbf{D}||_{2,\infty} \leq \alpha\}, \quad \mathbf{X}_i \in \mathbb{R}^{K \times n_i}, \quad i = 1, 2, \ldots, I, \end{aligned} \tag{6}$$

where λ_D, λ_X, μ_D, μ_X, $\alpha > 0$ are given constants. Problem (6) is an instance of P, with $f_i(\mathbf{D}, \mathbf{X}_i) = (1/2) \cdot ||\mathbf{S}_i - \mathbf{D}\mathbf{X}_i||_F^2$; $G(\mathbf{D}) = \lambda_D ||\mathbf{D}||_{1,1} + (\mu_D/2) \cdot ||\mathbf{D}||_F^2$, and $g_i(\mathbf{X}_i) = \lambda_X ||\mathbf{X}_i||_{1,1} + (\mu_X/2) \cdot ||\mathbf{X}_i||_F^2$. Note that orthonormality of factors are relaxed for sake of simplicity. A related formulation, termed Sparse PCA, has also been used in (Udell et al., 2016). It is not difficult to show that Assumption A is satisfied, and the Lipschitz constant in A2 is given by $L_{\nabla X_i}(\mathbf{D}) = (\sigma_{\max}(\mathbf{D}))^2$.

Non-negative Sparse Coding (NNSC) (Hoyer, 2004)

Non-negative Matrix Factorization (NMF) was primarily proposed by (Lee and Seung, 1999) as a better alternative to the classic SVD in learning localized features of image datasets, such as face images. The formulation enforces non-negativity of the entries of $\mathbf{D}$ and $\mathbf{X}$. This has been shown to empirically lead to sparse solutions; however no explicit control on sparsity is employed in the model. To overcome this shortcoming, (Hoyer, 2004) proposed a non-negative sparse coding (NNSC) formulation which extends NMF by adding a sparsity-inducing penalty function of $\mathbf{X}$. The problem reads

$$\begin{aligned} \min_{\mathbf{D},(\mathbf{X}_i)_{i=1}^I} \quad & \sum_{i=1}^{I} \left\{ \frac{1}{2} \left\| \mathbf{S}_i - \mathbf{D}\mathbf{X}_i \right\|_F^2 + \lambda \left\| \mathbf{X}_i \right\|_{1,1} + \frac{\mu}{2} \left\| \mathbf{X}_i \right\|_F^2 \right\} \\ \text{s.t.} \quad & \mathbf{D} \in \mathcal{D} \triangleq \{ \mathbf{D} \in \mathbb{R}_+^{M\times K} \mid ||\mathbf{D}||_{2,\infty} \leq \alpha \}, \; \mathbf{X}_i \in \mathbb{R}_+^{K\times n_i}, \quad i = 1, 2, \ldots, I, \end{aligned} \tag{7}$$

for some λ, μ, $\alpha > 0$. Problem (7) is another instance of P, with $f_i(\mathbf{D}, \mathbf{X}_i) = (1/2) \cdot ||\mathbf{S}_i - \mathbf{D}\mathbf{X}_i||_F^2$, $g_i(\mathbf{X}_i) = \lambda \|\mathbf{X}_i\|_{1,1} + (\mu/2) \cdot \|\mathbf{X}_i\|_F^2$, $G(\mathbf{D}) = 0$, $\mathcal{D} = \{\mathbf{D} \in \mathbb{R}_+^{M\times K} \mid ||\mathbf{D}||_{2,\infty} \leq \alpha\}$, and $\mathcal{X}_i = \mathbb{R}_+^{K\times n_i}$. Assumption A is satisfied, and the Lipschitz constant in A2 is given by $L_{\nabla X_i}(\mathbf{D}) = (\sigma_{\max}(\mathbf{D}))^2$.

3. Algorithmic Design

We introduce now our algorithmic framework. To shed light on the core idea behind the proposed scheme, we begin introducing an informal and constructive description of the algorithm, followed by its formal description along with its convergence properties.

Each agent i controls its private variable $\mathbf{X}_i$ and maintains a local copy of the shared variables $\mathbf{D}$, denoted by $\mathbf{D}_{(i)}$, along with an auxiliary variable $\mathbf{\Theta}_{(i)}$; we anticipate that $\mathbf{\Theta}_{(i)}$ aims at *locally* estimating the gradient sum $\sum_j \nabla_D f_j(\mathbf{D}_{(i)}, \mathbf{X}_j)$, an information that is not available at agent i's side. The value of these variables at iteration ν is denoted by $\mathbf{X}_i^\nu$, $\mathbf{D}_{(i)}^\nu$, and $\mathbf{\Theta}_{(i)}^\nu$, respectively. Roughly speaking, the update of these variables is designed so that asymptotically i) all the $\mathbf{D}_{(i)}$ will be consensual, i.e., $\mathbf{D}_{(i)} = \mathbf{D}_{(j)}$, $\forall i \neq j$; and ii) the tuples $(\mathbf{D}_{(i)}, (\mathbf{X}_j)_{j=1}^I)$ will be a stationary solutions of Problem P. This is accomplished throughout the following two steps, which are performed iteratively and in parallel across the agents.

Step 1: Local Optimization

The nonconvexity of f_i together with the lack of knowledge of $\sum_{j\neq i} f_j$ in F prevents agent i to solve directly Problem P with respect to $(\mathbf{D}_{(i)}, \mathbf{X}_i)$. Since f_i is *bi-convex* in $(\mathbf{D}_{(i)}, \mathbf{X}_i)$, a natural approach is then to update $\mathbf{D}_{(i)}$ and $\mathbf{X}_i$ in an *alternating* fashion by solving a *local* approximation of P. Specifically, at iteration ν, given the iterates $\mathbf{X}_i^\nu$, $\mathbf{D}_{(i)}^\nu$, and $\mathbf{\Theta}_{(i)}^\nu$, agent i fixes $\mathbf{X}_i = \mathbf{X}_i^\nu$ and solves the following strongly convex problem in $\mathbf{D}_{(i)}$:

$$\widetilde{\mathbf{D}}_{(i)}^\nu \triangleq \operatorname*{argmin}_{\mathbf{D}_{(i)}\in\mathcal{D}} \tilde{f}_i\big(\mathbf{D}_{(i)}; \mathbf{D}_{(i)}^\nu, \mathbf{X}_i^\nu\big) + \Big\langle I \cdot \mathbf{\Theta}_{(i)}^\nu - \nabla_D f_i(\mathbf{D}_{(i)}^\nu, \mathbf{X}_i^\nu), \mathbf{D}_{(i)} - \mathbf{D}_{(i)}^\nu \Big\rangle + G\left(\mathbf{D}_{(i)}\right), \tag{8}$$

where $\tilde{f}_i(\bullet; \mathbf{D}_{(i)}^\nu, \mathbf{X}_i^\nu)$ is a suitably chosen strongly convex approximation of $f_i(\bullet, \mathbf{X}_i^\nu)$ at $(\mathbf{D}_{(i)}^\nu, \mathbf{X}_i^\nu)$ (cf. Assumption C, Sec. 3.1); and $\mathbf{\Theta}_{(i)}^\nu$, as anticipated, is used to track the gradient of F, with $\lim_{\nu\to\infty} \|I \cdot \mathbf{\Theta}_{(i)}^\nu - \sum_{j=1}^I \nabla_D f_j(\mathbf{D}_{(i)}^\nu, \mathbf{X}_j^\nu)\| = 0$; which would lead to

$$\lim_{\nu\to\infty} \left\| \Big(I \cdot \mathbf{\Theta}_{(i)}^\nu - \nabla_D f_i(\mathbf{D}_{(i)}^\nu, \mathbf{X}_i^\nu) \Big) - \sum_{j\neq i} \nabla_D f_j(\mathbf{D}_{(i)}^\nu, \mathbf{X}_j^\nu) \right\| = 0. \tag{9}$$

This sheds light on the role of the linear term in (8): it can be regarded as a proxy of the sum-gradient $\sum_{j\neq i} \nabla_D f_j(\mathbf{D}_{(i)}^\nu, \mathbf{X}_j^\nu)$, which is not available at agent i's side. In Step 2 below we show how to update $\mathbf{\Theta}_{(i)}^\nu$ using only local information, so that (9) holds.

Given $\widetilde{\mathbf{D}}^{\nu}_{(i)}$, a step-size is employed in the update of $\mathbf{D}_{(i)}$, generating the iterate $\mathbf{U}^{\nu}_{(i)}$:

$$\mathbf{U}^{\nu}_{(i)} = \mathbf{D}^{\nu}_{(i)} + \gamma^{\nu}(\widetilde{\mathbf{D}}^{\nu}_{(i)} - \mathbf{D}^{\nu}_{(i)}), \tag{10}$$

where γ^{ν} is the step-size, to be properly chosen (see Assumption E, Sec. 3.1).

Let us now consider the update of the private variables $\mathbf{X}_i$. Fixing $\mathbf{D}_{(i)} = \mathbf{U}^{\nu}_{(i)}$, agent i computes the new update $\mathbf{X}_i^{\nu+1}$ by solving the following strongly convex optimization problem:

$$\mathbf{X}_i^{\nu+1} \triangleq \underset{\mathbf{X}_i \in \mathcal{X}_i}{\operatorname{argmin}} \quad \tilde{h}_i(\mathbf{X}_i; \mathbf{U}^{\nu}_{(i)}, \mathbf{X}_i^{\nu}) + g_i(\mathbf{X}_i), \tag{11}$$

where $\tilde{h}_i(\bullet; \mathbf{U}^{\nu}_{(i)}, \mathbf{X}_i^{\nu})$ is a strongly convex function of $\mathbf{X}_i$, approximating $f_i(\mathbf{U}^{\nu}_{(i)}, \bullet)$ at $(\mathbf{U}^{\nu}_{(i)}, \mathbf{X}_i^{\nu})$; see Assumption C (cf. Sec. 3.1) for specific instances of $\tilde{h}_i$.

Step 2: Local Communications

Let us design now a local communication mechanism ensuring asymptotic consensus over the local copies $\mathbf{D}_{(i)}$'s and property (9). To do so, we build on the (perturbed) push-sum protocol proposed in (Sun et al., 2016) (see also Kempe et al. (2003)). Specifically, an extra scalar variable ϕ_i is introduced at each agent's side to deal with the directed nature of the graph; given ϕ_i^{ν} and $\mathbf{U}^{\nu}_{(j)}$ from its in-neighbors $j \in \mathcal{N}_i$, each agent i updates its own local estimate $\mathbf{D}^{\nu}_{(i)}$ and ϕ_i^{ν} according to:

$$\phi_i^{\nu+1} = \sum_{j \in \mathcal{N}_i^{\text{in}}[\nu]} a_{ij}^{\nu}\, \phi_j^{\nu} \quad \text{and} \quad \mathbf{D}^{\nu+1}_{(i)} = \frac{1}{\phi_i^{\nu+1}} \sum_{j \in \mathcal{N}_i^{\text{in}}[\nu]} a_{ij}^{\nu}\, \phi_j^{\nu} \mathbf{U}^{\nu}_{(j)}, \tag{12}$$

where a_{ij}^{ν}'s are some weights (to be properly chosen, see Assumption F, Sec. 3.1); and $\phi_i^0 = 1$, for all $i = 1, \ldots, I$.

Note that the updates in (12) can be implemented locally: all agents only need to (i) send their local variable $\mathbf{U}^{\nu}_{(j)}$ and the scalar weight $a_{ij}^{\nu}\, \phi_j^{\nu}$ to their neighbors; and (ii) collect locally the information coming from the neighbors.

To update the $\mathbf{\Theta}^{\nu}_{(i)}$ variables we leverage the gradient tracking mechanism first introduced in (Di Lorenzo and Scutari, 2016), coupled with the push-sum consensus scheme (Sun et al., 2016), resulting in the following perturbed push-sum scheme:

$$\mathbf{\Theta}^{\nu+1}_{(i)} = \frac{1}{\phi_i^{\nu+1}} \sum_{j \in \mathcal{N}_i^{\text{in}}[\nu]} a_{ij}^{\nu} \phi_j^{\nu} \mathbf{\Theta}^{\nu}_{(j)} + \frac{1}{\phi_i^{\nu+1}} \left(\nabla_D f_i(\mathbf{D}^{\nu+1}_{(i)}, \mathbf{X}_i^{\nu+1}) - \nabla_D f_i(\mathbf{D}^{\nu}_{(i)}, \mathbf{X}_i^{\nu}) \right), \tag{13}$$

with $\mathbf{\Theta}^0_{(i)} \triangleq \nabla_D f_i(\mathbf{D}^0_{(i)}, \mathbf{X}_i^0)$, for all $i = 1, \ldots, I$. The update (13) follows similar logic as that of $\mathbf{D}^{\nu}_{(i)}$ in (12), with the difference that (13) contains a perturbation [the second term in the RHS of (13)], which employs $\mathbf{\Theta}^{\nu}_{(i)}$ and ensures the desired tracking properties (otherwise $\mathbf{\Theta}^{\nu}_{(i)}$ would converge to the average of their initial values). Note that (13) can be performed locally by agent i, following the same procedure as described for (12).

Combining the above steps, we can now formally introduce the proposed distributed algorithm for the DL problems P, as described in Algorithm 1, and termed D^4L (Decentralized Dictionary Learning over Dynamic Digraphs) *Algorithm.*

Algorithm 1 : Decentralized Dictionary Learning over Dynamic Digraphs ($\mathrm{D}^4\mathrm{L}$)

Initialization : set $\nu = 0$ and $\phi_i^0 = 1$, $\mathbf{D}_{(i)}^0 \in \mathcal{D}$, $\mathbf{X}_i^0 \in \mathcal{X}_i$, $\mathbf{\Theta}_{(i)}^0 = \nabla_D f_i(\mathbf{D}_{(i)}^0, \mathbf{X}_i^0)$, for all $i = 1, 2, \ldots, I$.

`S1.` If $(\mathbf{D}_{(i)}^\nu, \mathbf{X}_i^\nu)$ satisfies a suitable stopping criterion: `STOP`;

`S2.` **Local Optimization:** Each agent i computes:

(a) $\widetilde{\mathbf{D}}_{(i)}^\nu$ and $\mathbf{U}_{(i)}^\nu$ according to (8) and (10);

(b) $\mathbf{X}_i^{\nu+1}$ according to (11);

`S3.` **Local Communications:** Each agent i collects data from its current neighbors and updates:

(a) $\phi_i^{\nu+1}$ and $\mathbf{D}_{(i)}^{\nu+1}$ according to (12);

(b) $\mathbf{\Theta}_{(i)}^{\nu+1}$ according to (13);

`S4.` Set $\nu + 1 \to \nu$, and go to `S1`.

3.1. Algorithmic Assumptions

Before stating the main convergence result for the $\mathrm{D}^4\mathrm{L}$ Algorithm, we discuss the main assumptions governing the choices of the free parameters of the algorithm, namely: the surrogate functions $\tilde{f}_i$ and $\tilde{h}_i$, the step-size γ^ν, and the consensus weights $(a_{ij}^\nu)_{i,j=1}^I$.

3.1.1. On the choice of $\tilde{f}_i$ and $\tilde{h}_i$.

The surrogate functions are chosen to satisfy the following assumption.

Assumption C (On $\tilde{f}_i$ and $\tilde{h}_i$) *Given $\mathbf{D}_{(i)}^\nu$ and $\mathbf{X}_i^\nu$, $\tilde{f}_i(\bullet; \mathbf{D}_{(i)}^\nu, \mathbf{X}_i^\nu)$ in (8) is either*

$$\tilde{f}_i(\mathbf{D}_{(i)}; \mathbf{D}_{(i)}^\nu, \mathbf{X}_i^\nu) = f_i(\mathbf{D}_{(i)}, \mathbf{X}_i^\nu) + \frac{\tau_{D,i}^\nu}{2} ||\mathbf{D}_{(i)} - \mathbf{D}_{(i)}^\nu||_F^2, \tag{14}$$

or

$$\tilde{f}_i(\mathbf{D}_{(i)}; \mathbf{D}_{(i)}^\nu, \mathbf{X}_i^\nu) = \left\langle \nabla_D f_i(\mathbf{D}_{(i)}^\nu, \mathbf{X}_i^\nu), \mathbf{D}_{(i)} - \mathbf{D}_{(i)}^\nu \right\rangle + \frac{\tau_{D,i}^\nu}{2} \left\| \mathbf{D}_{(i)} - \mathbf{D}_{(i)}^\nu \right\|_F^2, \tag{15}$$

where $\tau_{D,i}^\nu$ is a positive scalar satisfying Assumption D.

Given $\mathbf{U}_{(i)}^\nu$ and $\mathbf{X}_i^\nu$, $\tilde{h}_i(\bullet; \mathbf{U}_{(i)}^\nu, \mathbf{X}_i^\nu)$ in (11) is either

$$\tilde{h}_i(\mathbf{X}_i; \mathbf{U}_{(i)}^\nu, \mathbf{X}_i^\nu) \triangleq f_i(\mathbf{U}_{(i)}^\nu, \mathbf{X}_i) + \frac{\tau_{X,i}^\nu}{2} \|\mathbf{X}_i - \mathbf{X}_i^\nu\|_F^2\,, \tag{16}$$

or

$$\tilde{h}_i(\mathbf{X}_i; \mathbf{U}_{(i)}^\nu, \mathbf{X}_i^\nu) = \left\langle \nabla_{X_i} f_i(\mathbf{U}_{(i)}^\nu, \mathbf{X}_i^\nu), \mathbf{X}_i - \mathbf{X}_i^\nu \right\rangle + \frac{\tau_{X,i}^\nu}{2} \left\| \mathbf{X}_i - \mathbf{X}_i^\nu \right\|_F^2, \tag{17}$$

where $\tau_{X,i}^\nu$ is a positive scalar satisfying Assumption D.

Assumption D (On $\tau_{X,i}^{\nu}$ and $\tau_{D,i}^{\nu}$) *The parameters* $(\tau_{X,i}^{\nu})_{i=1}^{I}$ *and* $(\tau_{D,i}^{\nu})_{i=1}^{I}$ *are chosen such that*

(D1) $\{\tau_{D,i}^{\nu}\}_{\nu}$ *and* $\{\tau_{X,i}^{\nu}\}_{\nu}$ *satisfy*

$$0 < \inf_{\nu} \tau_{D,i}^{\nu} \leq \sup_{\nu} \tau_{D,i}^{\nu} < +\infty, \tag{18}$$

and

$$\begin{aligned} &\sup_{\nu} \tau_{X,i}^{\nu} < +\infty, \\ &\tau_{X,i}^{\nu} \geq \tfrac{1}{2} L_{\nabla X_i}(\mathbf{U}_{(i)}^{\nu}) + \epsilon, \quad \forall \nu \geq 1, \end{aligned} \tag{19}$$

for all $i = 1, 2, \ldots, I$, *where* $\epsilon > 0$ *is an arbitrarily small constant, and* $L_{\nabla X_i}$ *is defined in Assumption A2.*

(D2) *Stronger convergence results [cf. Theorem 2] can be obtained if, under Assumption A5(ii), the sequences* $\{\tau_{D,i}^{\nu}\}_{\nu}$ *and* $\{\tau_{X,i}^{\nu}\}_{\nu}$*, in addition to D1, also satisfy*

$$\sum_{t=0}^{\infty} \left| \tau_{D,i}^{t+1} - \tau_{D,i}^{t} \right| < \infty, \tag{20}$$

and

$$\limsup_{\nu} \left| \tau_{X,i}^{\nu} - \tau_{X,i}^{\nu-1} \right| < \mu, \tag{21}$$

where $\mu \triangleq \min_i \mu_i$ *and* μ_i *is the strongly convexity constant of* f_i.

Discussion. Several comments are in order.

• *On the choice of* $\tilde{f}_i$ *and* $\tilde{h}_i$. Since f_i (resp. h_i) is convex in $\mathbf{D}_{(i)}$ (resp. $\mathbf{X}_i$), (14) [resp. (16)] is a natural choice for the surrogate $\tilde{f}_i$ (resp. $\tilde{h}_i$): the structure of f_i (resp. h_i) is preserved while a quadratic term is added to make the overall surrogate strongly convex. The non-smooth strongly convex subproblems (8) and (11) resulting from (14) and (16) can be solved using standard solvers, e.g., projected subgradient methods. When dealing with large-scale instances, effective methods are also (Facchinei et al., 2015; Daneshmand et al., 2015).

The alternative surrogates $\tilde{f}_i$ and $\tilde{h}_i$ as given in (15) and (17), respectively, are based on the the linearization of the original f_i and h_i. This option is motivated by the fact that, for specific instances of f_i and h_i, (15) and (17) lead to subproblems (8) and (11) whose solution can be computed in closed form. For instance, consider the *elastic net sparse DL* problem (3) in Sec. 2.2, where $f_i(\mathbf{D}, \mathbf{X}_i) = \frac{1}{2}||\mathbf{S}_i - \mathbf{D}\mathbf{X}_i||_F^2$; $G(\mathbf{D}) = 0$; and $g_i(\mathbf{X}_i) = \lambda\,||\mathbf{X}_i||_1 + \frac{\mu}{2}\,||\mathbf{X}_i||_F^2$, with $\lambda, \mu > 0$. By using (15), the resulting subproblem (8) admits the following closed form solution:

$$\widetilde{\mathbf{D}}_{(i)}^{\nu} = P_{\mathcal{D}}\left[\mathbf{D}_{(i)}^{\nu} - \frac{I}{\tau_{D,i}^{\nu}} \mathbf{\Theta}_{(i)}^{\nu}\right]. \tag{22}$$

Referring to the sparse coding subproblem (11), if $\tilde{h}_i$ is chosen according to (16), computing the update $\mathbf{X}_i^{\nu+1}$ results in solving a LASSO problem. If instead one uses the surrogate in (17), the solution of (11) can be computed in closed form as

$$\mathbf{X}_i^{\nu+1} = \frac{\tau_{X,i}^{\nu}}{\mu + \tau_{X,i}^{\nu}} \mathcal{T}_{\frac{\lambda}{\tau_{X,i}^{\nu}}}\left(\mathbf{X}_i^{\nu} - \frac{1}{\tau_{X,i}^{\nu}} \nabla_{X_i} f_i(\mathbf{U}_{(i)}^{\nu}, \mathbf{X}_i^{\nu})\right), \tag{23}$$

where $\mathcal{T}$ is the soft-thresholding operator $\mathcal{T}_\theta(x) \triangleq \max(|x| - \theta, 0) \cdot \text{sign}(x)$ [with $\text{sign}(\cdot)$ denoting the sign function], applied to the matrix argument component-wise.

• *On the choice of* $\tau^\nu_{X,i}$ *and* $\tau^\nu_{D,i}$. These coefficients must satisfy Assumption D. Roughly speaking, D1 ensures that $(\tau^\nu_{X,i})_{i=1}^I$ and $(\tau^\nu_{D,i})_{i=1}^I$ are bounded (both from below and above) while D2 guarantees that these parameters are asymptotically "stable". A trivial choice for $\tau^\nu_{D,i}$ satisfying both (18) and (20) is $\tau^\nu_{D,i} = c$, for some $c > 0$; some practical rules for $\tau^\nu_{X,i}$ satisfying both (19) and (21) are the following:

(a) Use a constant $\tau^\nu_{X,i}$, that is,

$$\tau^\nu_{X,i} = \max_{\mathbf{D}\in\mathcal{D}}\left[\max\left(\sigma_{\max}(\nabla^2_{X_i} f_i(\mathbf{D}, \mathbf{X}^\nu_i)), \tilde{\epsilon}\right)\right],$$

for some $\tilde{\epsilon} > 0$. The above value can be, however, much larger than any $\sigma_{\max}(\nabla^2_{X_i} f_i(\mathbf{U}^\nu_{(i)}, \mathbf{X}^\nu_i))$, which can slow down the practical convergence of the algorithm;

(b) A less conservative choice is to satisfy (19) iteratively, while guaranteeing that $\tau^\nu_{X,i}$ is uniformly positive:

$$\tau^\nu_{X,i} = \max(L_{\nabla X_i}(\mathbf{U}^\nu_{(i)}), \tilde{\epsilon}), \tag{24}$$

where $\tilde{\epsilon}$ is any positive (possibly small) constant;

(c) A generalization of (b) is

$$\tau^\nu_{X,i} \in \left[\max(L_{\nabla X_i}(\mathbf{U}^\nu_{(i)}), \tilde{\epsilon}),\ L_{\nabla X_i}(\mathbf{U}^\nu_{(i)}) + \tilde{\mu}\right],$$

for some $\tilde{\epsilon}$ and $\tilde{\mu}$ such that $0 < \tilde{\epsilon} \leq \tilde{\mu} < \mu$.

Remark 1 *While the choices (a)-(c) above clearly satisfy* (19), *it can be shown that* (21) *also holds, as a consequence of the continuity of* $L_{\nabla X_i}(\cdot)$ *and Proposition 5 (cf. Appendix A.4).*

Note that all the above rules do not require any coordination among the agents, but are implementable in a fully distributed manner, using only local information.

3.1.2. On the choice of γ^ν

The step-size can be chosen according to the following assumption.

Assumption E (On γ^ν) $\{\gamma^\nu\}_\nu$ *satisfies:* $\gamma^\nu \in (0, 1]$, *for all* ν; $\sum_{\nu=0}^\infty \gamma^\nu = \infty$; *and* $\sum_{\nu=0}^\infty (\gamma^\nu)^2 < \infty$.

The above assumption is the standard diminishing-rule; see, e.g., (Bertsekas and Tsitsiklis, 1997). Here, we only recall one rule, satisfying Assumption E, that we found very effective in our experiments, namely (Facchinei et al., 2015): $\gamma^\nu = \gamma^{\nu-1}(1 - \epsilon_0\gamma^{\nu-1})$ with $\gamma^0 \in (0, 1]$ and $\epsilon_0 \in (0, 1/\gamma^0)$.

3.1.3. On the choice of the weigh coefficients $\{a_{ij}^\nu\}$.

We denote by $\mathbf{A}^\nu$ the matrix whose entries are the weights a_{ij}^ν's, i.e., $[\mathbf{A}^\nu]_{i,j} = a_{ij}^\nu$. This matrix is chosen so that the following conditions are satisfied.

Assumption F (On the weighting matrix) *Given the digraph $\mathcal{G}^\nu = (\mathcal{V}, \mathcal{E}^\nu)$, each matrix $\mathbf{A}^\nu$, with $[\mathbf{A}^\nu]_{ij} = a_{ij}^\nu$, satisfies*

(F1) $a_{ii}^\nu \geq \kappa > 0$ *for all* $i = 1, \ldots, I$*;*

(F2) $a_{ij}^\nu \geq \kappa > 0$*, if* $(j, i) \in \mathcal{E}^\nu$*; and* $a_{ij}^\nu = 0$ *otherwise;*

(F3) $\mathbf{A}^\nu$ *is column stochastic, i.e.,* $\mathbf{1}^\intercal \mathbf{A}^\nu = \mathbf{1}^\intercal$.

When the graph $\mathcal{G}^\nu$ is directed, a valid choice of $\mathbf{A}^\nu$ is (Kempe et al., 2003): $a_{ij}^\nu = 1/d_j^\nu$ if $j \in \mathcal{N}_i^{\text{in}}[\nu]$, and $a_{ij}^\nu = 0$ otherwise, where d_j^ν is the out-degree of agent j at time ν. The resulting communication protocols (12)–(13) can be easily implemented in a distributed fashion:each agent i) broadcasts its local variable normalized by its current out-degree; and ii) collects locally the information coming from its neighbors. When the graph is undirected, several options are available in the literature, including: the Laplacian, Metropolis-Hastings, and maximum-degree weights; see, e.g., (Xiao et al., 2005).

4. Convergence of D^{4}L

In this section, we provide the main convergence results for the D^{4}L Algorithm. We begin introducing some definitions, instrumental to state our results. Let

$$\mathbf{D}^\nu \triangleq [\mathbf{D}_{(1)}^{\nu\intercal}, \mathbf{D}_{(2)}^{\nu\intercal}, \ldots, \mathbf{D}_{(I)}^{\nu\intercal}]^\intercal, \quad \mathbf{X}^\nu \triangleq [\mathbf{X}_1^\nu, \mathbf{X}_2^\nu, \ldots, \mathbf{X}_I^\nu], \quad \text{and} \quad \overline{\mathbf{D}}^\nu \triangleq \frac{1}{I}\sum_{i=1}^{I} \mathbf{D}_{(i)}^\nu. \tag{25}$$

Given the sequence $\{(\mathbf{D}^\nu, \mathbf{X}^\nu)\}_\nu$ generated by the D^{4}L Algorithm, convergence is stated measuring the distance of the sequence $\{(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu\}_\nu$ from optimality as well as the consensus disagreement among the local variables $\mathbf{D}_{(i)}^\nu$'s. Distance from stationarity is measured by

$$\Delta^\nu \triangleq \max(\Delta_D(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu), \Delta_X(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu)) \tag{26}$$

where

$$\Delta_D(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu) \triangleq ||\widehat{\mathbf{D}}(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu) - \overline{\mathbf{D}}^\nu||_{\infty,\infty}, \quad \Delta_X(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu) \triangleq ||\widehat{\mathbf{X}}(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu) - \mathbf{X}^\nu||_{\infty,\infty}, \tag{27}$$

with the functions $\widehat{\mathbf{D}}(\bullet, \bullet)$ and $\widehat{\mathbf{X}}(\bullet, \bullet)$ defined as:

$$\widehat{\mathbf{D}}(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu) \triangleq \underset{\mathbf{D}' \in \mathcal{D}}{\operatorname{argmin}} \left\langle \nabla_D F(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu), \mathbf{D}' - \overline{\mathbf{D}}^\nu \right\rangle + \frac{\hat{\tau}_D}{2} \left\| \mathbf{D}' - \overline{\mathbf{D}}^\nu \right\|^2 + G(\mathbf{D}'), \tag{28}$$

$$\widehat{\mathbf{X}}(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu) \triangleq [\widehat{\mathbf{X}}_1(\overline{\mathbf{D}}^\nu, \mathbf{X}_1^\nu), \ldots, \widehat{\mathbf{X}}_I(\overline{\mathbf{D}}^\nu, \mathbf{X}_I^\nu)],$$

$$\text{with } \widehat{\mathbf{X}}_i(\overline{\mathbf{D}}^\nu, \mathbf{X}_i^\nu) \triangleq \underset{\mathbf{X}_i' \in \mathcal{X}_i}{\operatorname{argmin}} \left\langle \nabla_{X_i} f_i(\overline{\mathbf{D}}^\nu, \mathbf{X}_i^\nu), \mathbf{X}_i' - \mathbf{X}_i^\nu \right\rangle + \frac{\hat{\tau}_X}{2} \left\| \mathbf{X}_i' - \mathbf{X}_i^\nu \right\|^2 + g_i(\mathbf{X}_i'), \tag{29}$$

for some given constants $\hat{\tau}_D > 0$ and $\hat{\tau}_X > 0$. Note that $\Delta_D(\bullet, \bullet)$ and $\Delta_X(\bullet, \bullet)$ are valid merit functions, in the sense that they are continuous and $\Delta_D(\overline{\mathbf{D}}^\infty, \mathbf{X}^\infty) = \Delta_X(\overline{\mathbf{D}}^\infty, \mathbf{X}^\infty) = 0$ if and only if $(\overline{\mathbf{D}}^\infty, \mathbf{X}^\infty)$ is a stationary solution of Problem (P) (Facchinei et al., 2015).

The consensus error at iteration ν is measured by the function

$$e^\nu \triangleq ||\mathbf{D}^\nu - \mathbf{1} \otimes \overline{\mathbf{D}}^\nu||_{\infty,\infty}. \tag{30}$$

Asymptotic convergence of D^{4}L to stationary solutions of (P) is stated in Theorem 2 below while the convergence rate is studied in Theorem 3.

Theorem 2 *Given Problem P under Assumption A, let $\{(\mathbf{D}^\nu, \mathbf{X}^\nu)\}_\nu$ be the sequence generated by the D^{4}L Algorithm for a given initial point $(\mathbf{D}^0, \mathbf{X}^0)$ and under Assumptions B, C, D1, E, F. Then,*

(a) `[Consensus]`*: All $\mathbf{D}^\nu_{(i)}$'s are asymptotically consensual, i.e., $\lim_{\nu\to\infty} e^\nu = 0$;*

(b) `[Convergence]`*: i) $\{(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu)\}_\nu$ is bounded; ii) $\{U(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu)\}_\nu$ converges to a finite value; iii) $\lim_{\nu\to\infty} \Delta_X(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu) = 0$; and iv) $\liminf_{\nu\to\infty} \Delta_D(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu) = 0$. Therefore, $\{(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu)\}_\nu$ has at least one limit point which is a stationary solution of P.*

If, in particular, Assumption A5(ii) holds and D1 is reinforced by D2, then convergence in (b) can be strengthened as follows:

(b') Case (b) holds and $\lim_{\nu\to\infty} \Delta^\nu = 0$, implying that all the limit points of $\{(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu)\}_\nu$ are stationary solutions of P.

Proof The proof is quite involved and is given in Appendix A.2. ■

The above theorem states two main convergence results under Assumptions B, C, D1, E, F: i) existence of at least a subsequence of $(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu)$ converging to a stationary solution of Problem P; and ii) asymptotic consensus of all $\mathbf{D}^\nu_{(i)}$ to a common value $\overline{\mathbf{D}}^\nu$. If Assumption A5(ii) is also assumed and D1 is reinforced by D2 the stronger results in (b') can be proven, showing that every limit point is a stationary solution. Note that from a practical point of view the weaker result guaranteeing existence of at least a subsequence converging to a stationary solution is perfectly satisfying, since it guarantees that the algorithm can be terminated after a finite number of iterations with an approximate solution.

Theorem 3 *Consider either settings of Theorem 2, with the additional assumption that the step-size sequence $\{\gamma^\nu\}_\nu$ is non-increasing. For any given $\epsilon > 0$, let $T_{D,\epsilon} \triangleq \min\{\nu \in \mathbb{N}_+ : \Delta_D(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu) \leq \epsilon\}$ and $T_{X,\epsilon} \triangleq \min\{\nu \in \mathbb{N}_+ : \Delta_X(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu) \leq \epsilon\}$. Then,*

(a) `[Rate of consensus error]`*:*

for every $\theta \in (0,1)$;

$$e^\nu = \mathcal{O}\left(\gamma^{\lceil \theta\nu \rceil}\right), \tag{31}$$

(b) `[Rate of optimization errors]`*:*

$$T_{X,\epsilon} = \mathcal{O}\left(\frac{1}{\epsilon^2}\right). \tag{32}$$

Let $\gamma^\nu = K/\nu^p$ with some constant $K > 0$ and $p \in (1/2, 1)$. Then,

$$T_{D,\epsilon} = \mathcal{O}\left(\frac{1}{\epsilon^{2/(1-p)}}\right). \tag{33}$$

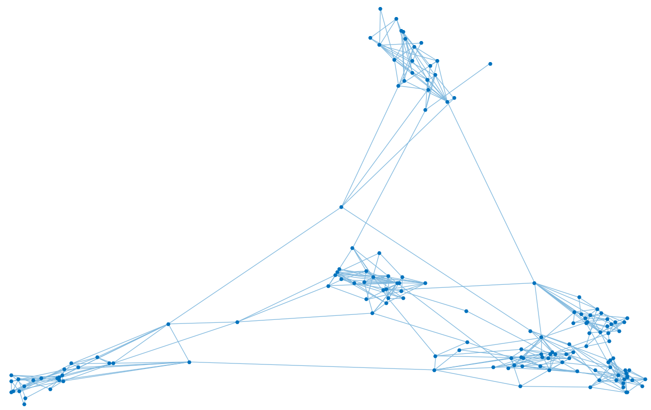


Figure 3: Topology of a simulated (sparse) network

Proof See Appendix A.3 ∎

We remark that, while a convergence rate has been established in the literature (see, e.g., Razaviyayn et al. (2014a)) for certain *centralized* algorithms applied to special classes of DL problems, Theorem 3 represents the first rate result for a *distributed* algorithm tackling the class of DL problems P.

5. Numerical Experiments

In this section, we test numerically our algorithmic framework on several classes of problems, namely: (i) Image denoising, (ii) Biclustering, (iii) Sparse PCA, and (iv) Non-negative sparse coding. We recall that $\mathrm{D^4L}$ is the first provably convergent distributed algorithm for Problem P; comparisons are thus not simple. To give the sense of the performance of $\mathrm{D^4L}$, in our experiments,

(i) when available, we implemented, *centralized* algorithms tailored to the specific problems under consideration and used the results as *benchmarks*;

(ii) for undirected graphs, we extended the (distributed) Prox-PDA-IP (Zhao et al., 2016) algorithm to the simulated instances of Problem P (generalizations of this method to directed graphs seem not possible);

(iii) for both undirected and directed graphs, we implemented a suitable version of the Adapt-Then-Combine (ATC) Algorithm (Chainais and Richard, 2013). Note that ATC has no formal convergence proof, and is originally designed to handle only undirected graphs, but we managed to make a sensible extension of this method to directed graphs too, by using some of the ideas developed in this paper.

All codes are written in MATLAB 2016b, and implemented on a computer with Intel Xeon (E5-1607 v3) quad-core 3.10GHz processor and 16.0 GB of DDR4 main memory.

5.1. Image Denoising

Problem formulations: We consider denoising a 512×512 pixels image of a fishing boat (USC, 1997)−see Fig. 5(a). We simulate a cluster computer network composed of 150 nodes (computers). Denoting by $\mathbf{F}_0$ and $\mathbf{F}$ the noise-free and corrupted image, respectively, the SNR (in dB) is defined as $\mathrm{SNR} \triangleq 20 \cdot \log(\|\mathrm{vec}(\mathbf{F}_0)\|_2/\sqrt{\mathrm{MSE}})$ while the Peak SNR (in dB) is defined as PSNR $\triangleq 20 \cdot \log(\max_i(\text{vec}(\mathbf{F}_0))_i/\sqrt{\text{MSE}})$, where MSE is the Mean-Squared-Error between $\mathbf{F}_0$ and $\mathbf{F}$. The fishing boat image is corrupted by additive white Gaussian noise, so that SNR = 15 dB and PSNR = 20.34 dB.

To perform the denoising task, we consider the elastic net sparse DL formulation (3). We extract 255,150 square sliding $s \times s$ pixel patches ($s = 8$) and aggregate the vectorized extracted patches in a single data matrix $\mathbf{S}$ of size $64 \times 255,150$. The size of the dictionary is $s^2 \times s^2 = 64 \times 64$; the data matrix is equally distributed across the 150 nodes, resulting in sparse representation matrices $\mathbf{X}_i$ of size 64×1701 ($K = 64$ and $n_i = 1701$). The total number of optimization variables is then $16,333,696$. The free parameters λ, μ and α in (3) are set to $\lambda = 1/s$, $\mu = \lambda$ and $\alpha = 1$, respectively.

Algorithms and tuning: We tested: i) two instances of the D^{4}L Algorithm, corresponding to two alternative choices of the surrogate functions; ii) the Prox-PDA-IP algorithm (Zhao et al., 2016), adapted to problem (3) (only on undirected networks); iii) the ATC algorithm (Chainais and Richard, 2013); and iv) the *centralized* K-SVD algorithm (Elad and Aharon, 2006) (`KSVD-Box v13` package), used as a benchmark. More specifically, the two instances of the D^{4}L Algorithm are:

- `Plain D`^{4}`L`: $\tilde{h}_i$ is chosen as in (16) (the original function) and $\tilde{f}_i$ as in (15);

- `Linearized D`^{4}`L`: $\tilde{h}_i$ is given by (17) (first-order approximation) and $\tilde{f}_i$ is given by (15).

The rest of the parameters in both instances of D^{4}L is set as: $\gamma^\nu = \gamma^{\nu-1}(1 - \epsilon\gamma^{\nu-1})$, with $\gamma^0 = 0.5$ and $\epsilon = 10^{-2}$; $\tau^\nu_{D,i} = 10$; and $\tau^\nu_{X,i} = \max(L_{\nabla X_i}(\mathbf{U}^\nu_{(i)}), 1)$ [cf. (24)].

Our adaptation of the Prox-PDA-IP algorithm to Problem (3) is summarized in Algorithm 2. The difference with the original version in (Zhao et al., 2016) are: i) the elastic net penalty is used in the objective function for the $\mathbf{X}_i$'s variables, instead of the ℓ_1-norm and ℓ_2-norm ball constraints; and ii) the variables $\mathbf{D}_{(i)}$'s are constrained in $\mathcal{D} \triangleq \{\mathbf{D} : ||\mathbf{D}\mathbf{e}_k||_2 \leq \alpha, k = 1, 2, \ldots, K\}$ rather than using the ℓ_2-norm regularization in the objective function. The other symbols used in Algorithm 2 are: i) the incidence matrix of $\mathcal{G}$, denoted by $\mathbf{M} = (M_{ei})_{e,i} \in \mathbb{R}^{E \times I}$, with $E \triangleq |\mathcal{E}|$; ii) the matrices $\mathbf{\Omega}^\nu_e \in \mathbb{R}^{M \times K}$, $e = 1, ..., E$, which are the ν-th iterate of the dual matrix variables $\mathbf{\Omega}_e \in \mathbb{R}^{M \times K}$, as introduced in the original Prox-PDA-IP; and iii) $\{\beta^\nu\}_{\nu \in \mathbb{N}_+}$ is the increasing penalty parameter, set to $\beta^\nu = 0.002\nu$.

All the algorithms are initialized to the same value: $\mathbf{D}^0_{(i)}$'s coincide with randomly (uniformly) chosen columns of $\mathbf{S}_{(i)}$'s whereas all $\mathbf{X}^0_i$'s are set to zero.

While the subproblems solved at each iteration ν in Linearized D^{4}L admit a closed-form−see (23) and (22)−in both Plain D^{4}L and ATC, the update of the dictionary has the closed form expression (22), but the update of the private variables calls for the solution of a LASSO problem (cf. Sec. 3.1). For both Plain D^{4}L and ATC, the LASSO subproblems at iteration ν are solved using the (sub)gradient algorithm, with the following tuning. A diminishing step-size is used, set to $\gamma^r = \gamma^{r-1}(1 - \epsilon\gamma^{r-1})$, where $\gamma^0 = 0.9$, $\epsilon = 10^{-3}$, and r denotes the inner iteration index. A warm start is used for the subgradient algorithm: the initial points are set to $\mathbf{X}^\nu_i$, where ν is the iteration index of the outer loop. We terminate

---

Algorithm 2 : Prox-PDA-IP algorithm (Zhao et al., 2016)

Initialization : $\mathbf{D}_{(i)}^0 \in \mathcal{D},\ \mathbf{X}_i^0 \in \mathcal{X}_i,\ \mathbf{\Omega}^0 = \mathbf{0}$;

`S1`. If $(\mathbf{D}_{(i)}^{\nu}, \mathbf{X}_i^{\nu})_i$ satisfies stopping criterion: `STOP`;

`S2`. Each agent i computes $\theta_i^{\nu} = ||\mathbf{D}_{(i)}^{\nu}\mathbf{X}_i^{\nu} - \mathbf{S}_i||_F^2$ and:

(a) $\mathbf{X}_i^{\nu+1} = \underset{\mathbf{X}_i \in \mathbb{R}^{K\times n_i}}{\operatorname{argmin}}\ f_i(\mathbf{D}_{(i)}^{\nu}, \mathbf{X}_i) + g_i(\mathbf{X}_i) + \dfrac{\beta^{\nu+1}\theta_i^{\nu}}{2}||\mathbf{X}_i - \mathbf{X}_i^{\nu}||_F^2 + \dfrac{\beta^{\nu+1}}{2}||\mathbf{D}_{(i)}^{\nu}(\mathbf{X}_i - \mathbf{X}_i^{\nu})||_F^2$;

(b) $\mathbf{D}_{(i)}^{\nu+1} = \underset{\mathbf{D}_{(i)} \in \mathcal{D}}{\operatorname{argmin}}\ \ f_i(\mathbf{D}_{(i)}, \mathbf{X}_i^{\nu+1}) + \sum_{e=1}^{E} M_{ei}\left\langle \mathbf{\Omega}_e^{\nu}, \mathbf{D}_{(i)}\right\rangle$
$\qquad + \beta^{\nu+1}\Big(d_i||\mathbf{D}_{(i)}||_F^2 - \big\langle \mathbf{D}_{(i)}, (d_i - 1)\mathbf{D}_{(i)}^{\nu} + \sum_{j\in\mathcal{N}_i}\mathbf{D}_{(j)}^{\nu}\big\rangle\Big)$;

(c) $\mathbf{\Omega}_e^{\nu+1} = \mathbf{\Omega}_e^{\nu} + \beta^{\nu+1}\sum_{i=1}^{I} M_{ei}\mathbf{D}_{(i)}^{\nu+1}, \quad \forall e = (i,j) \in \mathcal{E}$;

`S3`. Set $\nu + 1 \to \nu$, and go to `S1`.

---

the subgradient algorithm in the inner loop when $J_i^r \leq 10^{-6}$, with

$$J_i^r \triangleq \left\|\mathbf{X}_i^{\nu,r} - \frac{s}{1+s}\,\mathcal{T}_{\frac{1}{s}}\left(\mathbf{X}_i^{\nu,r} - \left(\nabla_{X_i} f_i(\mathbf{U}_{(i)}^{\nu}, \mathbf{X}_i^{\nu,r}) + \tau_{X,i}^{\nu}\left(\mathbf{X}_i^{\nu,r} - \mathbf{X}_i^{\nu}\right)\right)\right)\right\|_{\infty,\infty},$$

where $\mathbf{X}_i^{\nu,r}$ denotes the value of $\mathbf{X}_i$ at the r-th inner iteration and outer iteration ν; and $\mathcal{T}_{\theta}(x) \triangleq \max(|x| - \theta, 0)\cdot \operatorname{sign}(x)$ is the soft-thresholding operator, applied to the matrix argument componentwise. In all our simulations, we observed that the above accuracy was reached within 30 (inner) iterations of the subgradient algorithm.

In the Prox-PDA-IP scheme, Step `S2` (cf. Algorithm 2) calls for the solution of two subproblems, including a LASSO problem. As for Plain D^{4}L and ATC, we used the (projected) (sub)-gradient algorithm (with the same diminishing step-size rule) to solve the subproblems; we terminated the inner loop when the length between two consecutive iterates of the (projected) (sub)-gradient algorithm goes for the first time below 10^{-6}.

We simulated both undirected and directed static graphs. In the former case, there is no need of the ϕ-variables and, in the second equation of (12) [and (13)], the terms $(\phi_j^{\nu}\, a_{ij}^{\nu})/\phi_i^{\nu+1}$ reduce to a_{ij}. The weights a_{ij} are chosen according to the Metropolis-Hasting rule (Xiao et al., 2007); the resulting matrix $\mathbf{A}^{\nu} = [a_{ij}]_{ij}$ is thus time-invariant and doubly stochastic. When the graph is directed, we use the update of the ϕ_i^{ν}'s as in (12), with the weights a_{ij}^{ν} chosen according to the push-sum protocol (Kempe et al., 2003) (cf. Sec. 3.1.3).

Convergence speed and quality of the reconstruction: In the first set of simulations, we considered an undirected graph composed of 150 nodes, clustered in 6 groups of 25 (see Fig. 3). Starting from this topology, we kept adding random edges till a connected graph was obtained. Specifically, an arc is added between two nodes in the same cluster (resp. different clusters) with probability $p_1 = 0.2$ (resp. $p_2 = 2 \times 10^{-3}$).

In Fig. 4 we plot the objective function value [subplot on the left], the consensus disagreement e^{ν} as in (30) [subplot in the center], and the distance from stationarity Δ^{ν} as in

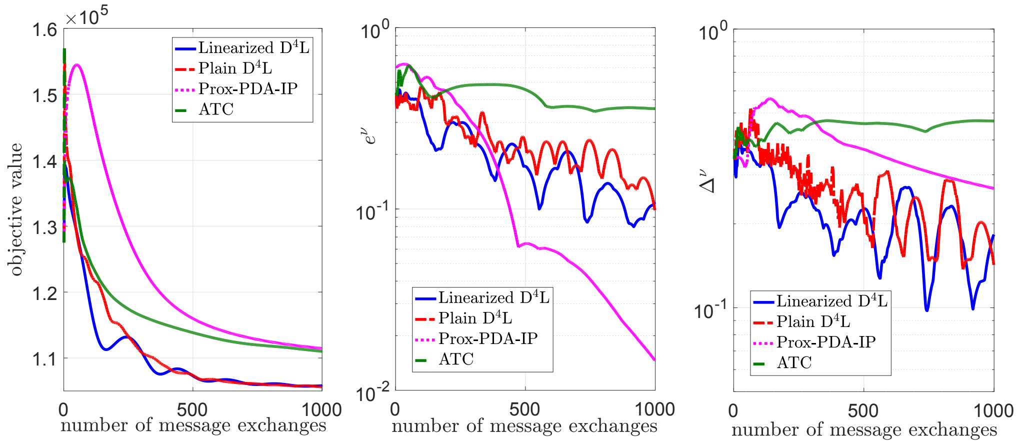


Figure 4: Denoising problem – D^{4}L, Prox-PDA-IP and ATC algorithms: objective value [subplot on the left], consensus disagreement [subplot in the center], and distance from stationarity Δ^ν [cf. (26)] [subplot on the right] vs. number of message exchanges.

(28) [subplot on the right] versus the *number of message exchanges*, achieved by Plain D^{4}L, Linearized D^{4}L, Prox-PDA-IP, and ATC. Note that the number of messages exchanged in the ATC algorithm at iteration ν coincides with ν whereas for Prox-PDA-IP and the D^{4}L schemes is 2ν (recall that the latter schemes employ two steps of communications per iteration). The figures clearly show that both versions of D^{4}L are much faster than Prox-PDA-IP and ATC (or, equivalently, they require fewer information exchanges). Moreover, ATC does not seem to reach a consensus on the local copies of the dictionary, while Prox-PDA-IP and D^{4}L schemes reach an agreement quite soon. In Fig. 5, we plot the reconstructed images along with their PSNR and MSE, obtained by the algorithms, when terminated after 1000 message exchanges. The figures clearly show superior performance of D^{4}L over its competitors. Also, the values of PSNR and MSE achieved by D^{4}L are comparable with those obtained by (centralized) K-SVD (`KSVD-Box v13` package).

A closer look at Fig. 4 shows that a significant decay on the objective function occurs in the first 200 message exchanges. It is then interesting to check the quality of the reconstructed images, achieved by the algorithms if terminated then. In Fig. 6, we report the images and values of PSNR and MSE obtained by terminating the schemes after 200 message exchanges (we also plot the benchmark obtained by K-SVD, run till optimality). The figure shows that both versions of D^{4}L attain high quality solutions even if terminated after few message exchanges while ATC and Prox-PDA-IP lag behind. This means that, in practice, there is no need to run D^{4}L till very low values of e^ν and Δ^ν are achieved.

Since the algorithms do not have the same cost-per-iteration, to get further insights into the performance of these schemes, we also compare them in terms of running time. In Table 2, we report the *averaged elapsed time* to execute one iteration of all algorithms. We considered the same setting as in the previous figures, but we terminated all algorithms after 273 seconds, which corresponds to the time for the fastest algorithm (i.e. Linearized D^{4}L) to perform 200 message exchanges [cf. Fig. 6]. The associated reconstructed images are shown in Fig. 7. Once again, these results clearly show that the linearized D^{4}L scheme significantly outperforms Prox-PDA-IP and ATC. Also, Linearized D^{4}L performs considerably better

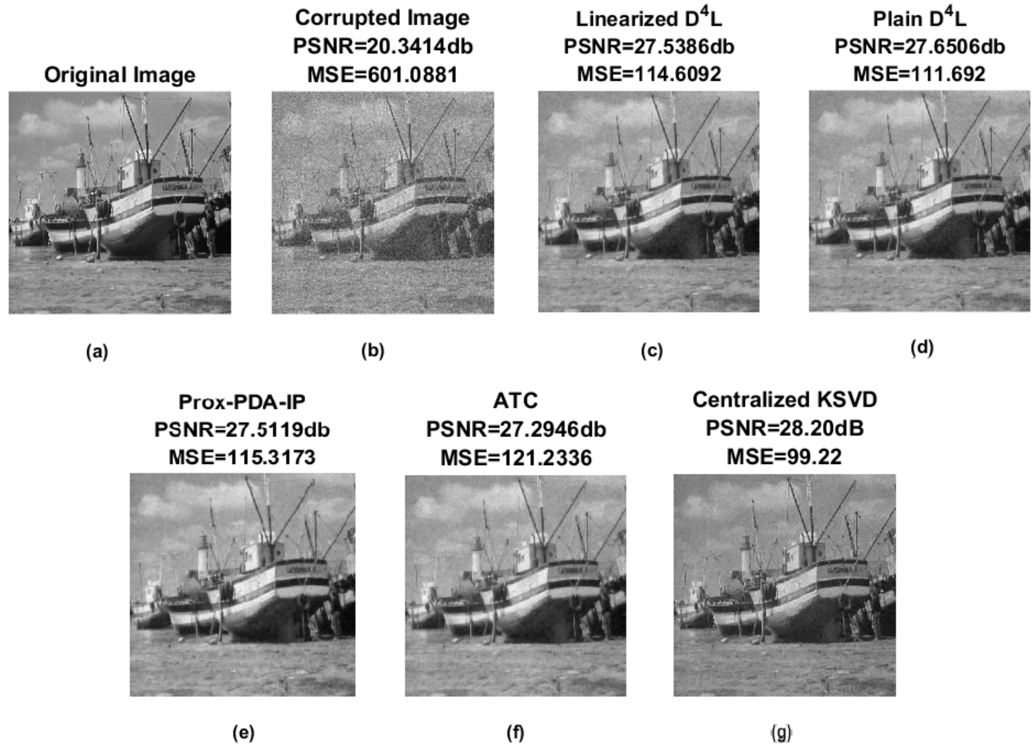


Figure 5: Denoising outcome. (a): original image; (b): corrupted image; (c)-(f): denoising achieved by D^{4}L, Prox-PDA-IP and ATC terminated after 1000 message exchanges; and (g): denoising achieved by centralized K-SVD (`KSVD-Box v13`).

than Plain D^{4}L, when terminated early; the explanation is in Table 2 which shows that the time of one iteration of the former algorithm is much shorter than that of Plain D^{4}L.

Algorithm	**Average Time per Iteration (sec)**
Linearized D^{4}L	2.862
Plain D^{4}L	11.328
Prox-PDA-IP	30.98
ATC	9.838

Table 2: D^{4}L vs. Prox-PDA-IP and ATC: Average computation time per iteration

Impact of the graph topology and connectivity: We study now the influence of the topology and graph connectivity on the performance of the algorithms. We consider *directed, static* graphs. We generated 5 instances of digraphs, with different connectivity, according to the following procedure. There are 500 nodes ($I = 500$), which are clustered in $n_c = 50$ clusters, each of them containing $10 = I/n_c$ nodes. Each node has an outgoing arc to another node in the same cluster with probability p_1 while p_2 is the probability of an outgoing arc to a node in a different cluster. We chose the values of p_1 and p_2, as in Table 3; we simulated three scenarios, namely: N1 corresponds to a “highly” connected network, N3 describes a “poorly” connected scenario, and N2 is an intermediate case. For

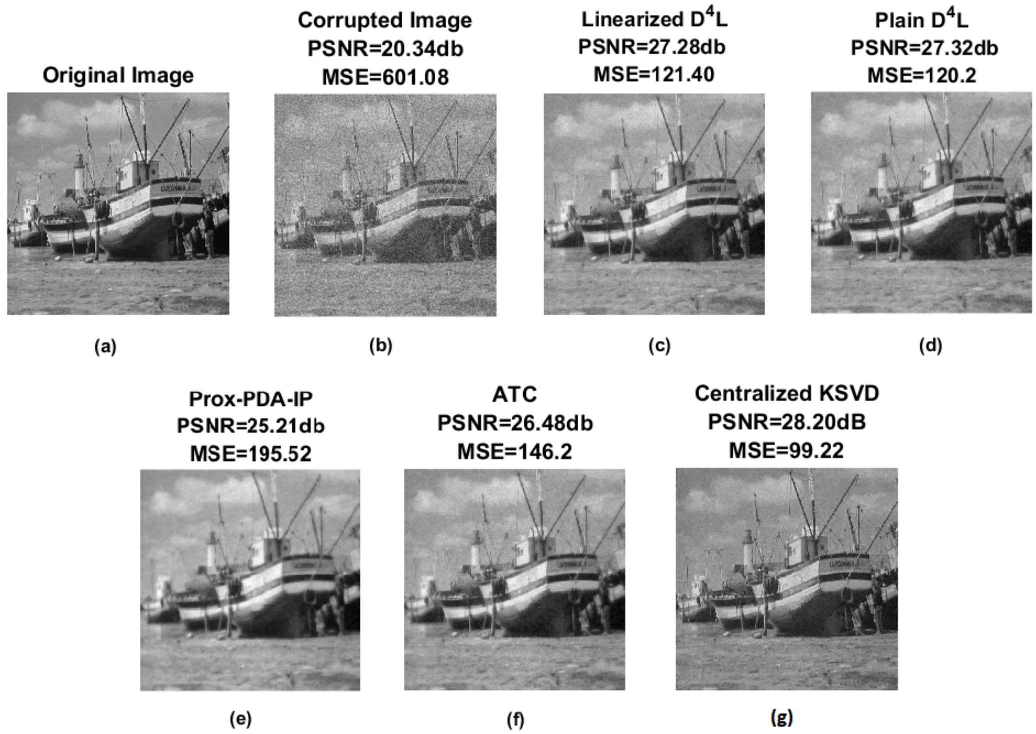


Figure 6: Denoising outcome. (a): original image; (b): corrupted image; (c)-(f): denoising achieved by D^{4}L, Prox-PDA-IP and ATC terminated after 200 message exchanges; and (g): denoising achieved by centralized K-SVD (`KSVD-Box v13`).

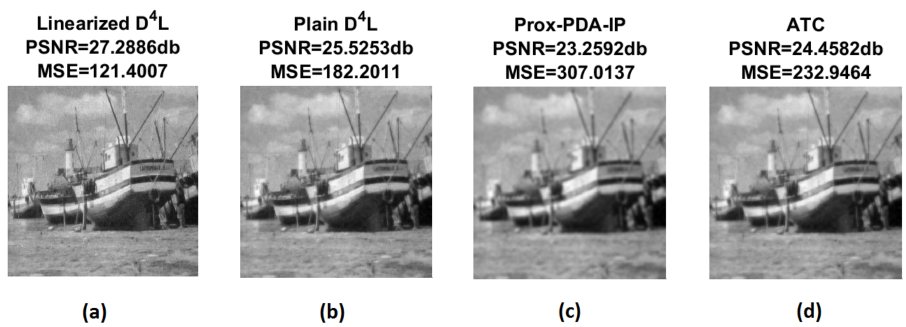


Figure 7: Denoising outcome. (a): Linearized D^{4}L; (b): Plain D^{4}L; (c): Prox-PDA-IP; (d) ATC; all terminated after after 273 seconds run-time (corresponding to 200 message exchanges of the Linearized D^{4}L).

each scenario, we generated 5 random instances (if a generated graph was not strongly connected we discarded it and generated a new one) and then ran Plain and Linearized D^{4}L and ATC on the resulting 15 graphs. Recall that ATC was not designed to work on directed networks. We thus modified it by using our new consensus protocol (but not the gradient tracking mechanism); we term it *Modified* ATC.

In Fig. 8 we plot the *average* value of the objective function [subplot on the left], the consensus disagreement e^ν [subplot in the center], and the distance from stationarity Δ^ν [subplot on the right], achieved by Plain D^{4}L, Linearized D^{4}L, and Modified ATC, versus the number of message exchanges, for the three scenarios N1 [subplot (a)], N2 [subplot (b)]

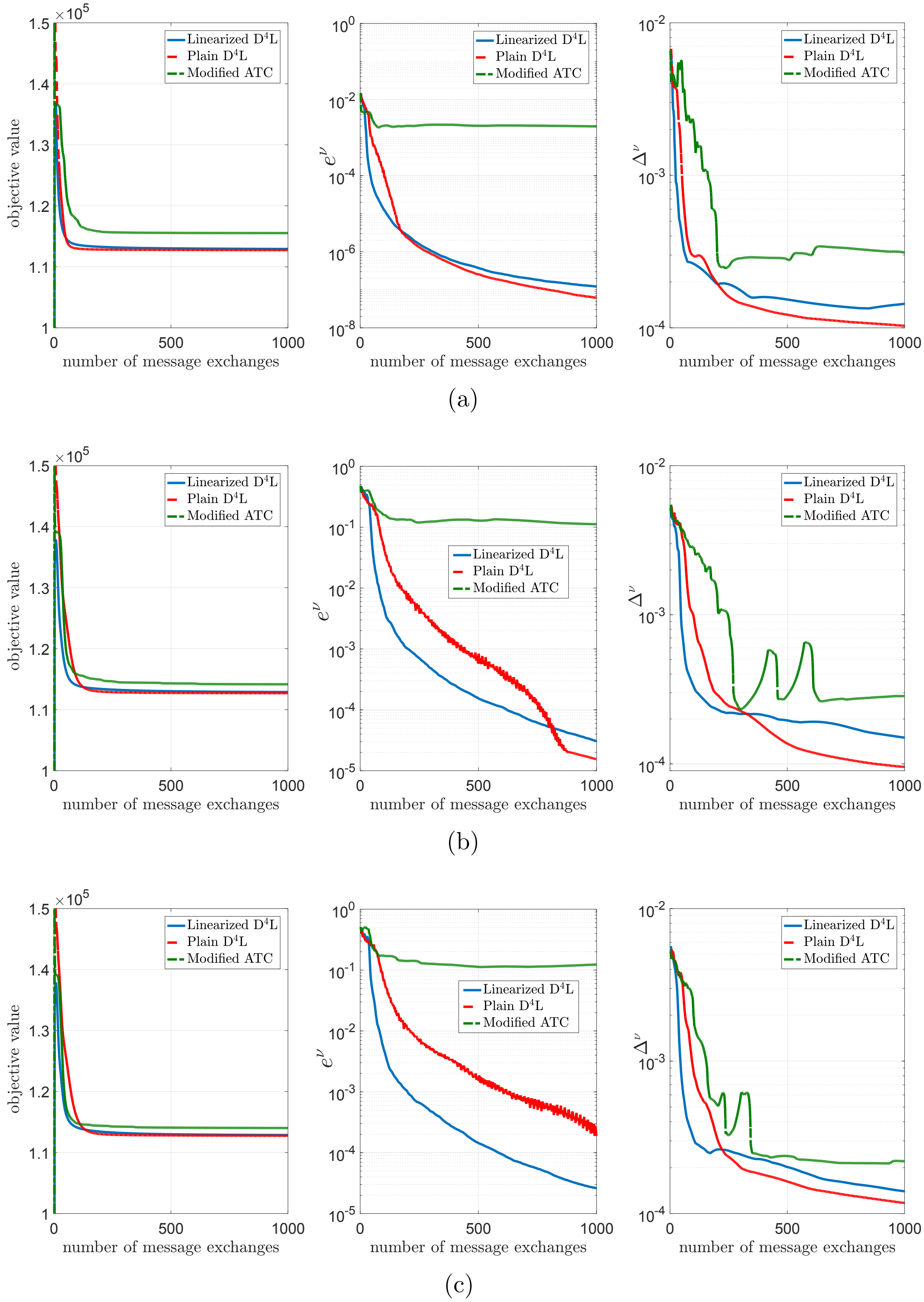


Figure 8: Denoising problem–D^{4}L and Modified ATC algorithms: objective value [subplots on the left], consensus disagreement [subplots in the center], and distance from stationarity Δ^{ν} [cf. (26)] [subplots on the right] vs. number of message exchanges. Comparison over three network settings [cf. Table 3]: N1 [subplots (a)], N2 [subplots (b)], and N3 [subplots (c)].

Network #	I	n_c	p_1	p_2
N1	500	50	0.9	0.9
N2	500	50	0.1	0.01
N3	500	50	0.05	0.01

Table 3: Network setting

and N3 [subplot (c)]. The average is taken over the aforementioned 5 digraph realizations. While also this batch of tests confirms the better behavior of D^{4}L schemes over ATC, it is interesting to observe that there seems to be little influence of the degree of connectivity on the behavior of Linearized and Plain D^{4}L. The only aspect for which a reasonable influence can be seen is on consensus. In fact, with respect to consensus, Linearized D^{4}L seems to improve over Plain D^{4}L, when connectivity decreases. This has a natural interpretation. Plain D^{4}L solves much more accurate subproblems at each iteration and this is, in some sense useless, especially in early iterations, when information has not spread across the network. It seems clear that the less connected the graph, the more time information needs to spread. Therefore, in scenario N1, the two methods are almost equivalent and, looking at consensus error, we see that initially Linearized D^{4}L is better than Plain D^{4}L, but soon, as information spreads, Plain D^{4}L becomes, even if slightly, better than Linearized D^{4}L. The same behavior can be observed for scenario N2, but this time the initial advantage of Linearized D^{4}L is larger and the switching point is reached much later. This is consistent with the fact that information needs more time to spread and therefore solving the accurate subproblem is not advantageous. If one passes to N3, where connectivity is very loose, there is no switching point within the first 1000 message exchanges.

5.2. Biclustering

Biclustering has been shown to be useful in several applications, including biology, information retrieval, and data mining; see, e.g., (Madeira and Oliveira, 2004).

Problem Formulation: We consider a Biclustering problem in the form (6), applied to genetic information. We solved the problem simulating a networked computer cluster composed of 500 nodes (see Table 3). The genetic data is borrowed from (Lee et al., 2010) (centered and normalized): the data matrix $\mathbf{S}$ of size $56 \times 12,625$ ($M = 56$ and $N = 12,625$) contains microarray gene expressions of 56 patients (rows); each patient is either identified to be normal (Normal) or belonging to one of the following three types of lung cancer: pulmonary carcinoid tumors (Carcinoid), colon metastases (Colon), and small cell carcinoma (SmallCell). We considered the *unsupervised* instance of the problem, meaning that none of the a-priori information about the type of patients' cancer is used to perform biclustering. Following the numerical experiments of (Lee et al., 2010), we seek rank-3 sparse matrices $\mathbf{X}_i$, and the data matrix $\mathbf{S}$ is equally distributed across the 500 nodes, resulting thus in $K = 3$ and $n_i = 26$. The total number of variables is then $39,168$. The other parameters are set as follows: $\alpha = 1$, $\lambda_X = \mu_X = 0.1$, and $\lambda_D = \mu_D = 0.1$.

Algorithms and tuning: We tested the instance of D^{4}L where $\tilde{f}_i$ and $\tilde{h}_i$ are chosen according to (14) and (16), respectively. The rationale behind this choice is to exploit the extra structure of the original function f_i, plus, in case of $\tilde{f}_i$, there is no certain benefit in using the linear approximation (15) as it does *not* lead to any closed form solution of the subproblem (8). Note that the Prox-PDA-IP scheme is not applicable here since the network is directed.

The other parameters of the algorithm are set to: $\gamma^\nu = \gamma^{\nu-1}(1-\epsilon\gamma^{\nu-1})$, with $\gamma^0 = 0.2$ and $\epsilon = 10^{-2}$; and $\tau^\nu_{D,i} = 100$ and $\tau^\nu_{X,i} = \max(L_{\nabla X_i}(\mathbf{U}^\nu_{(i)}), 100)$. We term such an instance of D^{4}L *Plain D^{4}L*. We compared Plain D^{4}L with the following algorithms: i) (a modified version of) the distributed ATC algorithm (Chainais and Richard, 2013), where the optimization of $\mathbf{D}$ is adjusted to solve (6) (the elastic-net penalty is added), and the consensus mechanism is modified with our new consensus protocol to handle directed network topologies; we termed this instance *Modified ATC*; and ii) the centralized SSVD algorithm proposed in Lee et al. (2010) (implemented using the MATLAB code provided by the authors), to benchmark the results obtained by the distributed algorithms. All the distributed algorithm are initialized setting each $\mathbf{X}^0_i = \mathbf{0}$, and each $\mathbf{D}^0_{(i)}$ equal to some randomly chosen columns of $\mathbf{S}_i$.

In D^{4}L, the subproblems (8) and (11) at iteration ν do not have a closed form solution; they are solved using the projected (sub)gradient algorithm, with diminishing step-size $\gamma^r = \gamma^{r-1}(1-\epsilon\gamma^{r-1})$, where $\gamma^0 = 0.9$, $\epsilon = 10^{-3}$, and r denotes the inner iteration index. A warm start is used for the projected subgradient algorithm; the initial points are set to $\mathbf{D}^\nu_{(i)}$ and $\mathbf{X}^\nu_i$ in problems (8) and (11), respectively, where ν is the iteration index of the outer loop. We terminate the projected subgradient algorithm solving (8) and (11) when $J^r_{D,i} \triangleq \|\widehat{\mathbf{D}}^{\nu,r}_{(i)} - \mathbf{D}^{\nu,r}_{(i)}\|_{\infty,\infty} \leq 10^{-6}$ and $J^r_{X,i} \triangleq \|\widehat{\mathbf{X}}^{\nu,r}_i - \mathbf{X}^{\nu,r}_i\|_{\infty,\infty}, \leq 10^{-6}$, respectively, where

$$\begin{aligned}
\widehat{\mathbf{D}}^{\nu,r}_{(i)} \triangleq \underset{\mathbf{D}_{(i)}\in\mathcal{D}}{\operatorname{argmin}} & \left\langle \nabla_D f_i(\mathbf{D}^{\nu,r}_{(i)}, \mathbf{X}^\nu_i) + I\boldsymbol{\Theta}^\nu_{(i)} - \nabla_D f_i(\mathbf{D}^\nu_{(i)}, \mathbf{X}^\nu_i) + \tau^\nu_{D,i}\big(\mathbf{D}^{\nu,r}_{(i)} - \mathbf{D}^\nu_{(i)}\big), \mathbf{D}_{(i)} - \mathbf{D}^{\nu,r}_{(i)} \right\rangle \\
& + \frac{100}{2}\left\|\mathbf{D}_{(i)} - \mathbf{D}^{\nu,r}_{(i)}\right\|^2 + G\left(\mathbf{D}_{(i)}\right), \\
\widehat{\mathbf{X}}^{\nu,r}_i \triangleq \underset{\mathbf{X}_i\in\mathbb{R}^{K\times n_i}}{\operatorname{argmin}} & \left\langle \nabla_{X_i} f_i(\mathbf{U}^\nu_{(i)}, \mathbf{X}^{\nu,r}_i) + \tau^\nu_{X,i}\left(\mathbf{X}^{\nu,r}_i - \mathbf{X}^\nu_i\right), \mathbf{X}_i - \mathbf{X}^{\nu,r}_i \right\rangle \\
& + \frac{100}{2}\|\mathbf{X}_i - \mathbf{X}^{\nu,r}_i\|^2 + g_i\left(\mathbf{X}_i\right),
\end{aligned}$$

with $\mathbf{D}^{\nu,r}_{(i)}$ and $\mathbf{X}^{\nu,r}_i$ denoting the value of $\mathbf{D}_{(i)}$ and $\mathbf{X}_i$ at the ν-th outer and r-th inner iteration, respectively. In all our simulations, the above accuracy was reached within 50 (inner) iterations of the projected subgradient algorithm.

Convergence speed and quality of the reconstruction: We simulated 3 *directed static* network topologies, namely: N1-N3, as given in Table 3. In Fig. 9 we plot he *average* value of the objective function [subplot on the left], the consensus disagreement e^ν [subplot in the center], and the distance from stationarity Δ^ν [subplot on the right], achieved by Plain D^{4}L and Modified ATC, versus the number of message exchanges, for the three scenarios N1 [subplot (a)], N2 [subplot (b)] and N3 [subplot (c)]. The average is taken over 5 digraph realizations. Fig. 9 shows that Plain D^{4}L algorithm attains satisfactory merit values in all network scenarios, while Modified ATC fails to reach consensus/convergence, even in highly connected networks. The poor performance of Modified ATC seem mainly due to the incapability of locking the consensus.

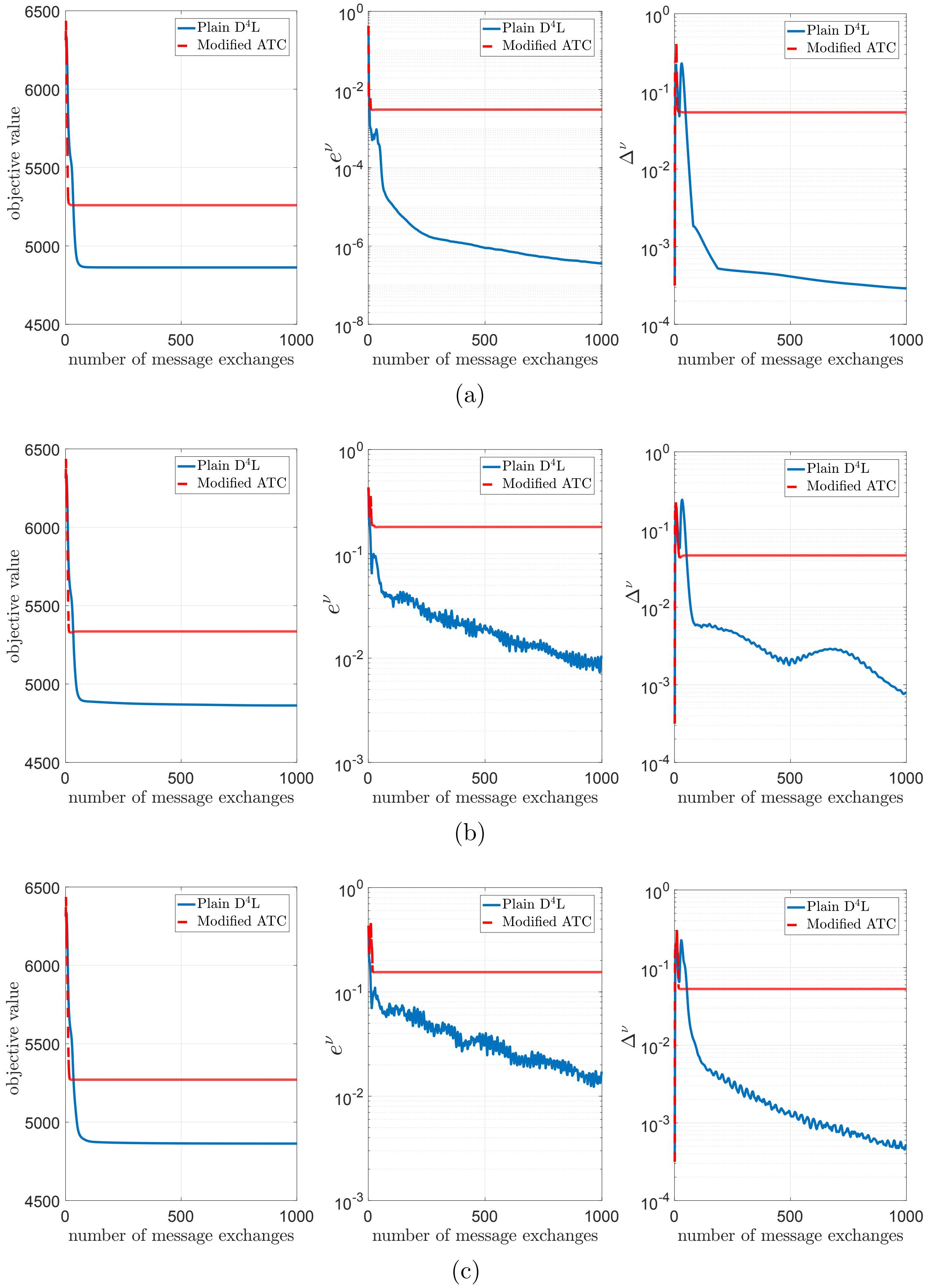


Figure 9: Biclustering problem–Plain D^{4}L and Modified ATC algorithms: objective value [subplots on the left], consensus disagreement [subplots in the center], and distance from stationarity Δ^ν [cf. (26)] [subplots on the right] vs. number of message exchanges. Comparison over three network settings [cf. Table 3]: N1 [subplots (a)], N2 [subplots (b)], and N3 [subplots (c)].

In order to assess the quality of the solutions achieved by the three algorithms, we employ the following procedure. Given the limit point (up to the fixed accuracy) $\mathbf{D}^\infty$ of the algorithm under consideration, patients' information is in form of (unlabeled clusters of) data points $\{\mathbf{D}^\infty_{m,:}\}_{m=1}^{56}$, where $\mathbf{D}^\infty_{m,:}$ denotes the m-th row of $\mathbf{D}^\infty$ and represents an individual patient. In order to compare $\mathbf{D}^\infty$ with the labeled ground truth, we need to tag labels to the clustered points of $\mathbf{D}^\infty$. To do so, we run the K-means clustering algorithm on $\{\mathbf{D}^\infty_{m,:}\}_{m=1}^{56}$. Specifically, we first run K-means 100 times and, in each run, we perform a preliminary clustering using 10% of the points (randomly chosen). Then, among the 100 obtained clustering configurations, we picked the one with the smallest "within-cluster sum of point-to-centroid distances".[1] Finally, we assign to each cluster the label associated with the most populated type of cancer in the cluster. Denoting the ground truth classes by $\{\mathcal{C}_i\}_{i=1}^4$ (recall that there are 4 classes/types of cancer), where each $\mathcal{C}_i$ consists of the group of patients with the same type of cancer, and by $\{\widetilde{\mathcal{C}}_i\}_{i=1}^4$ the clustering obtained by the procedure described above applied to the outcome $\mathbf{D}^\infty$ of the simulated algorithms, we measure the quality of the clustering by the *Jaccard index*, defined as

$$J = \frac{\left|\bigcup_i \left(\mathcal{C}_i \cap \widetilde{\mathcal{C}}_i\right)\right|}{\sum_i \left|\mathcal{C}_i \cup \widetilde{\mathcal{C}}_i\right|}.$$

Clearly $0 \leq J \leq 1$, and the higher the index value, the better the quality of the clustering.

In Table 4, we report the average and Maximum Absolute Deviation (MAD) of the Jaccard indices from their average, computed over the aforementioned 5 realizations of the three graph topologies, as in Table 3 (see also Fig. 9). The values in the table clearly show that Plain D^{4}L achieves better results than those produced by Modified ATC or centralized methods. Moreover, the value of the Jaccard index from Plain D^{4}L does not depend on the specific network topology. which is not the case for Modified ATC.

Network #	**Plain D^{4}L**	**Modified ATC**	**Lee et al. (2010)**
N1	0.8983/0	0.7778/0	–
N2	0.8983/0	0.7045/0.3218	–
N3	0.8983/0	0.7892/0.0172	–
Centralized	–	–	0.7231/–

Table 4: Biclustering problem−Average/MAD of Jaccard indices over 5 realizations of digraphs.

5.3. Non-negative Sparse Coding (NNSC) and Sparse PCA (SPCA)

Problem Formulation: We consider the Non-negative Sparse Coding (NSC) formulation (7) (Hoyer, 2004) and the Sparse PCA problem (6) (Mairal et al., 2010). For both formulations, we run experiments using the following two datasets:

- MIT-CBCL face database #1 (Sung, 1996): a pool of $N = 2,429$ vectorized face images of size 19×19 pixels each (i.e. $M = 361$);

1. Given a clustering partition $\{\mathcal{C}_i\}_{i=1}^4$, the "within-cluster sum of point-to-centroid distances" measures the quality of the k-means clustering, and is defined as $\sum_{i=1}^4 \sum_{j\in\mathcal{C}_i} ||\mathbf{D}^\infty_{j,:} - \overline{\mathbf{D}}^\infty_{\mathcal{C}_i}||^2$, where $\overline{\mathbf{D}}_{\mathcal{C}_i} \triangleq \frac{1}{|\mathcal{C}_i|}\sum_{j\in\mathcal{C}_i} \mathbf{D}^\infty_{j,:}$ and $|\mathcal{C}_i|$ denotes the cardinality of the set $\mathcal{C}_i$.

- The VOC 2006 database (Everingham et al., 2010): a pool of $N = 10,000$ vectorized natural image patches of size 16×16 pixels each (i.e. $M = 256$).

Consistently with (Mairal et al., 2010), the free parameters are set as:

- NNSC (7): $K = 49$, $\lambda = \mu = 1/\sqrt{M}$, and $\alpha = 1$;
- Sparse PCA (6): $K = 49$, $\lambda_X = \mu_X = 1/\sqrt{M}$, $\lambda_D = \mu_D = 1/\sqrt{M}$, and $\alpha = 1$.

The total number of variables for the above optimization problems are 136,710 for the MIT-CBCL dataset, and 502,544 for the VOC 2006 dataset.

We simulated the communication network as *static directed* graphs of size I, clustered in n_c groups, where each node has an outgoing arc to another node in the same cluster with probability p_1, while p_2 is the probability of an outgoing arc to a node in a different cluster. We run our tests over 6 different network scenarios, with various size I and probability pair (p_1, p_2), as given in Table 5. Note that if N/I is not an integer, we pad zero columns to the data matrix $\mathbf{S}$ so that all the agents own equal-size partitions $\mathbf{S}_i$'s, thus $n_i = \lceil N/I \rceil$ in both problems (6) and (7).

Network #	I	n_c	p_1	p_2
N4	10	2	0.9	0.3
N5	10	2	0.2	0.1
N6	50	5	0.9	0.3
N7	50	5	0.2	0.1
N8	250	10	0.9	0.3
N9	250	10	0.2	0.1

Table 5: Network setting for the NNSC and Sparse PCA problems.

5.3.1. Non-negative Sparse Coding

Algorithms and tuning: We test the Plain D^{4}L, with $\tilde{f}_i$ and $\tilde{h}_i$ chosen according to (15) and (16), respectively. The other parameters of the algorithm are set to: $\gamma^\nu = \gamma^{\nu-1}(1 - \epsilon\gamma^{\nu-1})$, with $\gamma^0 = 0.2$ and $\epsilon = 10^{-2}$; and $\tau^\nu_{D,i} = 10$ and $\tau^\nu_{X,i} = \max(L_{\nabla X_i}(\mathbf{U}^\nu_{(i)}), 10)$. We compare the proposed scheme with a modified version of ATC, equipped with our new consensus protocol, implementable on directed networks. All the distributed algorithm are initialized setting $\mathbf{X}^0_i = \mathbf{0}$ and $\mathbf{D}^0_{(i)}$ equal to some randomly chosen columns of $\mathbf{S}_i$. Both Plain D^{4}L and Modified ATC call for solving a LASSO problem in updating the private variables (cf. Sec. 3.1); the update of the dictionary has instead a closed form expression, see (22). For both Plain D^{4}L and Modified ATC, the LASSO subproblems at iteration ν are solved using the projected (sub)gradient algorithm with diminishing step-size $\gamma^r = \gamma^{r-1}(1 - \epsilon\gamma^{r-1})$, where $\gamma^0 = 0.9$, $\epsilon = 10^{-3}$, and r denoting the inner iteration index. We terminate the projected subgradient algorithm in the inner loop when $J^r_{X,i} \triangleq \|\widehat{\mathbf{X}}^{\nu,r}_i - \mathbf{X}^{\nu,r}_i\|_{\infty,\infty} \leq 10^{-4}$, where

$$\widehat{\mathbf{X}}^{\nu,r}_i \triangleq \underset{\mathbf{X}_i \in \mathcal{X}_i}{\operatorname{argmin}} \left\langle \nabla_{X_i} f_i(\mathbf{U}^\nu_{(i)}, \mathbf{X}^{\nu,r}_i) + \tau^\nu_{X,i}\left(\mathbf{X}^{\nu,r}_i - \mathbf{X}^\nu_i\right), \mathbf{X}_i - \mathbf{X}^{\nu,r}_i \right\rangle + \frac{1}{2}\left\|\mathbf{X}_i - \mathbf{X}^{\nu,r}_i\right\|^2 + g_i\left(\mathbf{X}_i\right),$$

and $\mathbf{X}_i^{\nu,r}$ denotes the value of $\mathbf{X}_i$ at the ν-th outer and r-th inner iteration. In all our simulations, the above accuracy was reached within 30 (inner) iterations of the projected subgradient algorithm.

Convergence speed and quality of the reconstruction: We run the Plain D^{4}L and Modified ATC algorithms over different network settings, as listed in Table 5, and we terminated them after 1500 message exchanges. We replicated the tests for 5 independent realizations and we reported the *average* of the final values of the objective function, the consensus disagreement, and the distance from stationarity in Table 6 and Table 7, for the MIT-CBCL and VOC 2006 datasets, respectively. In Fig. 10 and Fig. 11 (for MIT-CBCL and VOC 2006 datasets, respectively), we plot the *average* value (over the aforementioned 5 graph realizations) of the objective function [subplot on the left], the consensus disagreement e^ν [subplot in the center], and the distance from stationarity Δ^ν [subplot on the right], versus number of message exchanges, for the two extreme network scenarios N4 [subplot (a)] and N9 [subplot (b)]. These results show that the proposed Plain D^{4}L significantly outperforms the Modified ATC algorithm. They also show a remarkable stability of Plain D^{4}L with respect to the simulated network graphs, which is not observed for the Modified ATC, whose performance deteriorates significantly going from N4 to N9.

Network #	objective value	consensus disagreement	distance from stationarity
N4	169.8/171.9	9.17e-7/2.9e-4	4.7e-4/8.8e-2
N5	169.9/172.2	4.5e-6/6.7e-4	5.3e-4/9.8e-2
N6	169.8/177.2	2.4e-7/1.9e-4	5.1e-4/6.3e-2
N7	169.9/177.3	1.1e-6/6.3e-4	5.5e-4/6.1e-2
N8	169.8/191.0	2.1e-7/1.2e-4	5.1e-4/2.2e-2
N9	169.8/190.9	5.8e-7/3.1e-4	6.6e-4/1.1e-2

Table 6: NNSC problem (MIT-CBCL dataset)–Plain D^{4}L/Modified ATC algorithms: objective value, consensus disagreement, and distance from stationarity obtained after 1500 message exchanges.

Network #	objective value	consensus disagreement	distance from stationarity
N4	845.2/848.8	2.9e-6/1.3e-4	1.1e-3/4.1e-3
N5	845.9/850.1	1.5e-5/5.3e-4	1.5e-3/6.1e-3
N6	844.8/879.5	5.7e-7/2.2e-4	1.3e-3/4.4e-2
N7	844.5/879.3	1.9e-6/1.2e-3	1.3e-3/3.9e-2
N8	844.8/941.2	5.6e-7/1.5e-4	1.6e-3/4.8e-2
N9	845.0/941.8	1.5e-6/3.6e-4	1.1e-3/5.1e-2

Table 7: NNSC problem (VOC 2006 dataset)–Plain D^{4}L/Modified ATC algorithms: objective value, consensus disagreement, and distance from stationarity obtained after 1500 message exchanges.

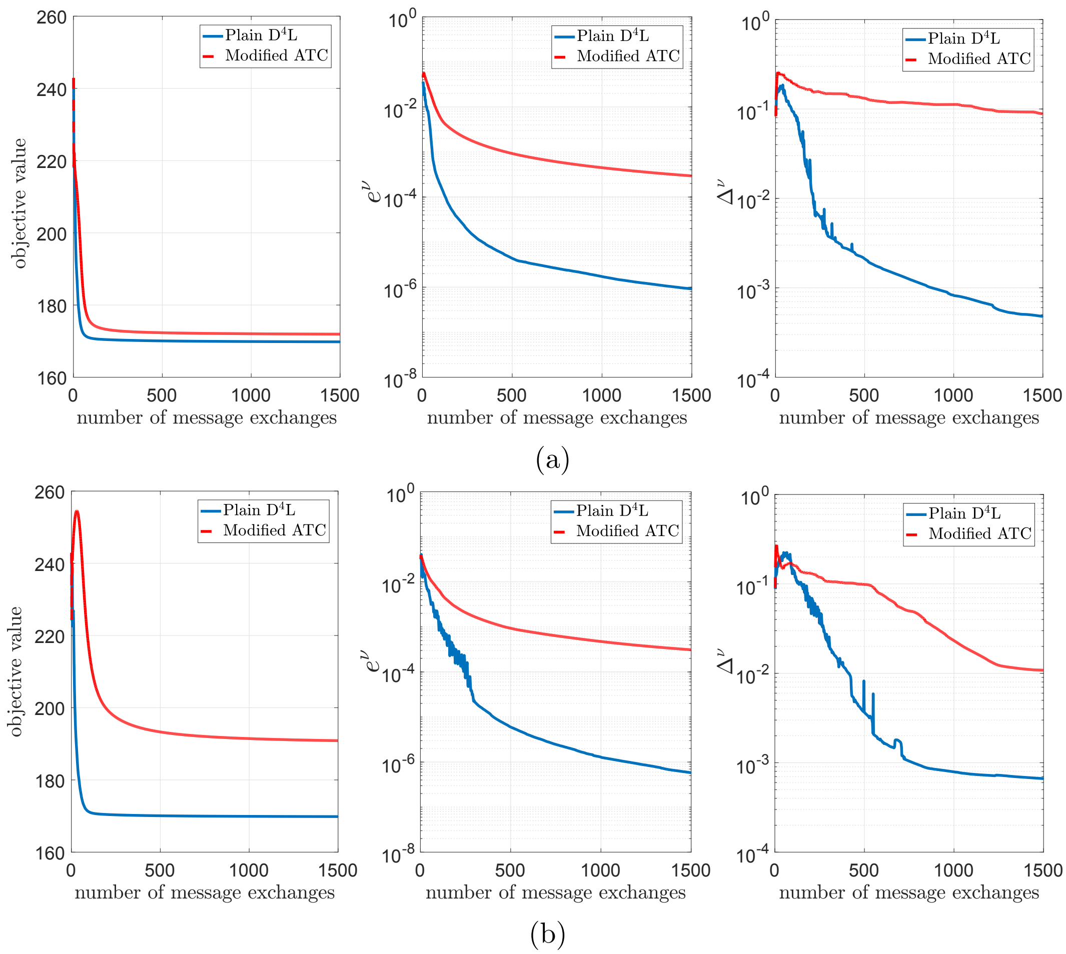
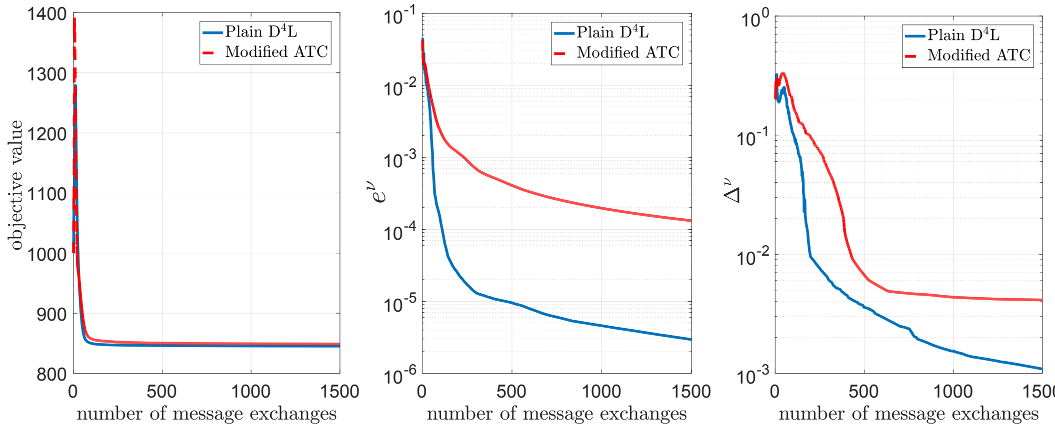


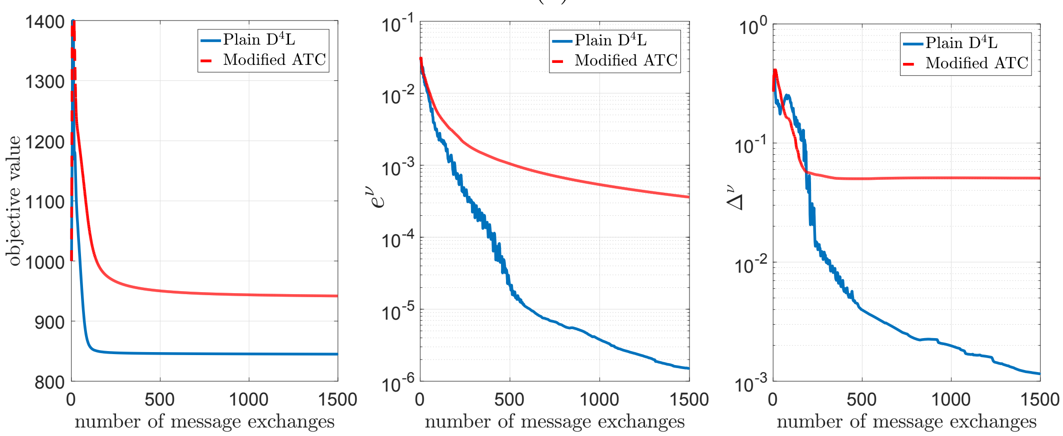
Figure 10: NNSC problem (MIT-CBCL dataset)–Plain D^{4}L and Modified ATC algorithms: objective value [subplots on the left], consensus disagreement [subplots in the center], and distance from stationarity Δ^ν [cf. (26)] [subplots on the right] vs. number of message exchanges. Comparison over the network settings N4 [subplots (a)] and N9 [subplots (b)] (cf. Table 5).

5.3.2. Sparse Principal Component Analysis (SPCA)

Algorithms and tuning: We test the same D^{4}L version as used in the Biclustering experiments (cf. Subsec. 5.2), i.e., $\tilde{f}_i$ and $\tilde{h}_i$ are chosen according to (14) and (16), respectively; we set $\gamma^\nu = \gamma^{\nu-1}(1-\epsilon\gamma^{\nu-1})$, with $\gamma^0 = 0.2$ and $\epsilon = 10^{-2}$; and $\tau^\nu_{D,i} = 10$ and $\tau^\nu_{X,i} = \max(L_{\nabla X_i}(\mathbf{U}^\nu_{(i)}), 10)$. We term it Plain D^{4}L. We compare Plain D^{4}L with a modified version of the ATC algorithm, which has been adapted to solve (6) and equipped with our new consensus protocol to handle directed network topologies. All the distributed algorithm are initialized, setting $\mathbf{X}^0_i = \mathbf{0}$ and $\mathbf{D}^0_{(i)}$ equal to some randomly chosen columns of $\mathbf{S}_i$. The subproblems (8) and (11) at iteration ν are solved using the projected (sub)gradient algorithm; the setting is the same as that used in the Biclustering problem (cf. Subsec. 5.2). We terminate the projected subgradient algorithm in the inner loop when $J^r_{D,i} \leq 10^{-4}$ (in



(a)



(b)

Figure 11: NNSC problem (VOC 2006 dataset)–Plain D^{4}L and Modified ATC algorithms: objective value [subplots on the left], consensus disagreement [subplots in the center], and distance from stationarity Δ^ν [cf. (26)] [subplots on the right] vs. number of message exchanges. Comparison over the network settings N4 [subplots (a)] and N9 [subplots (b)] (cf. Table 5).

solving subproblem (8)) and $J^r_{X,i} \leq 10^{-4}$ (in solving subproblem (11)), where $J^r_{D,i}$ and $J^r_{X,i}$ are defined as those in Subsec. 5.2. In all our simulations, the above accuracy was reached within 30 (inner) iterations of the projected subgradient algorithm.

Convergence speed and quality of the reconstruction: We test the Plain D^{4}L and the Modified ATC in different network settings, as listed in Table 5. The setting of the experiments and the averaging procedure of the reported values is the same of those used for the NNSC problem. The results of our experiments are reported in Table 8 and Figure 12 for the MIT-CBCL dataset; and in Table 9 and Figure 13 for the VOC 2006 dataset. The behaviors or the algorithms are very similar to those described in the NNSC case and confirm all previous observations.

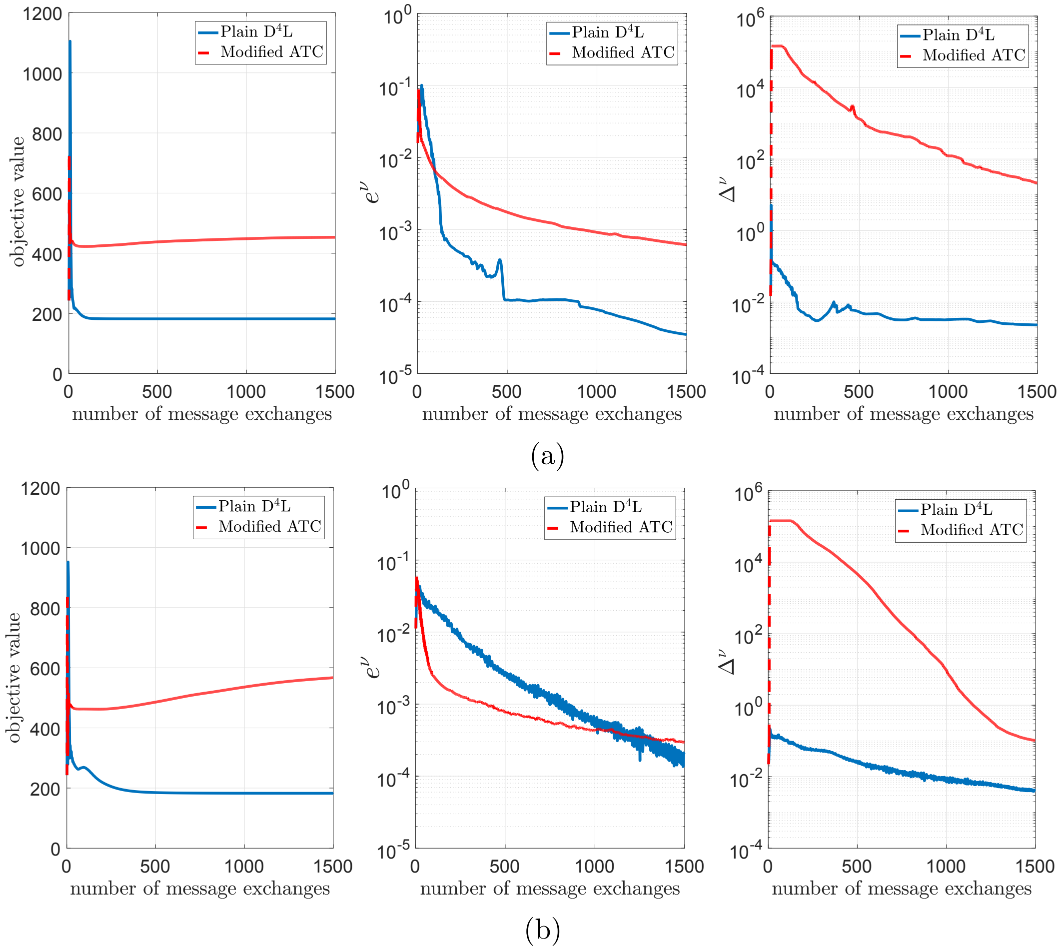


Figure 12: SPCA problem (MIT-CBCL dataset)–Plain D^{4}L and Modified ATC algorithms: objective value [subplots on the left], consensus disagreement [subplots in the center], and distance from stationarity Δ^ν [cf. (26)] [subplots on the right] vs. number of message exchanges. Comparison over the network settings N4 [subplots (a)] and N9 [subplots (b)] (cf. Table 5).

Network #	**objective value**	**consensus disagreement**	**distance from stationarity**
N4	181.9/453.1	3.5e-5/6.1e-4	2.2e-3/21.39
N5	185.1/446.4	2.6e-5/1.5e-3	1.3e-3/136.2
N6	182.7/517.3	5.2e-6/4.3e-4	1.3e-3/1.2e-1
N7	186.3/512.0	2.3e-5/9.2e-4	1.4e-3/1.18
N8	181.9/566.0	4.0e-5/1.7e-4	2.6e-3/1.5e-1
N9	182.6/566.9	1.7e-4/2.9e-4	4.2e-3/1.0e-1

Table 8: SPCA problem (MIT-CBCL dataset)–Plain D^{4}L/Modified ATC algorithms: objective value, consensus disagreement, and distance from stationarity obtained after 1500 message exchanges.

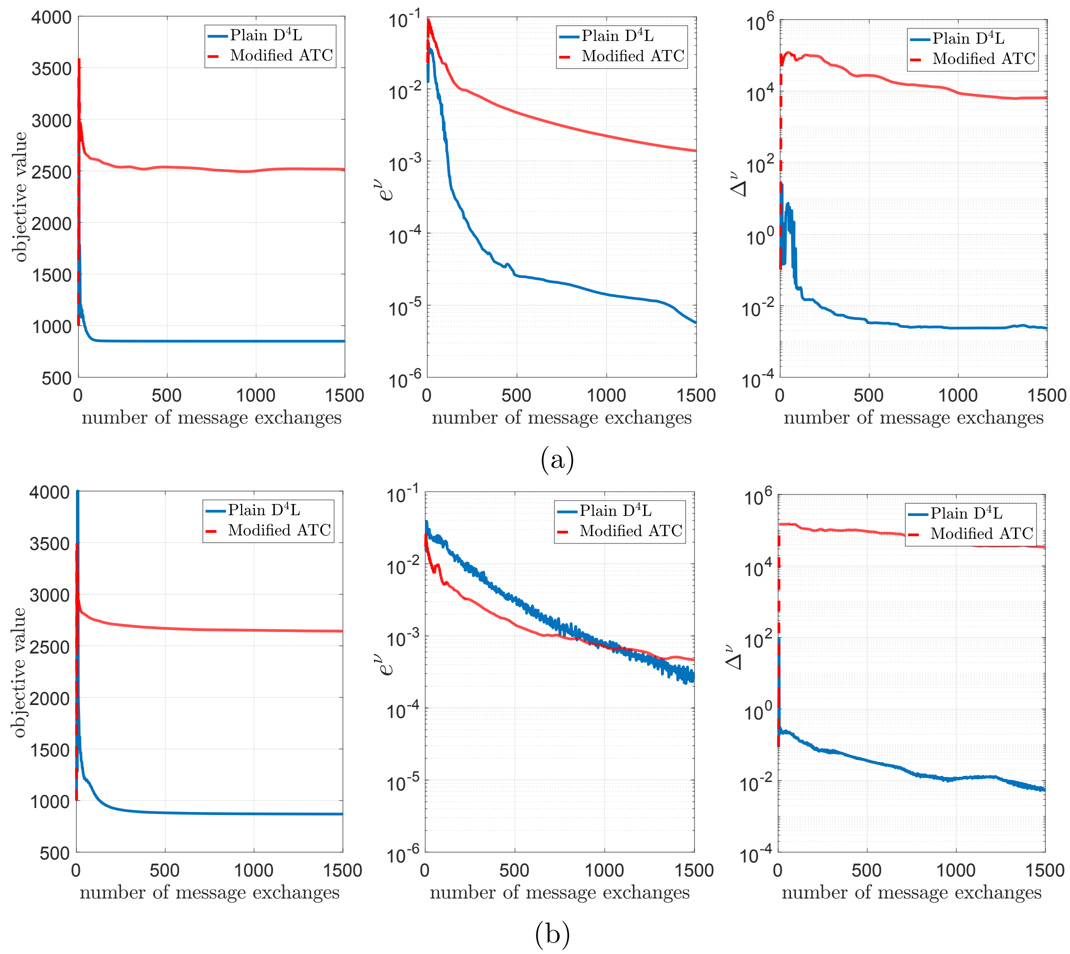


Figure 13: SPCA problem (VOC 2006 dataset)–Plain D^{4}L and Modified ATC algorithms: objective value [subplots on the left], consensus disagreement [subplots in the center], and distance from stationarity Δ^ν [cf. (26)] [subplots on the right] vs. number of message exchanges. Comparison over the network settings N4 [subplots (a)] and N9 [subplots (b)] (cf. Table 5).

Network #	objective value	consensus disagreement	distance from stationarity
N4	849.3/2516.0	5.7e-6/1.3e-3	2.3e-3/6532
N5	866.0/2533.6	1.8e-5/3.5e-3	2.0e-3/1.3e+4
N6	864.8/2621.0	7.2e-6/4.3e-4	3.4e-3/1.7e+4
N7	859.1/2624.1	5.3e-5/1.1e-3	2.4e-3/1.6e+4
N8	872.7/2643.2	6.6e-5/2.3e-4	1.1e-2/2.9e+4
N9	869.1/2642.0	2.3e-4/4.6e-4	5.6e-3/3.4e+4

Table 9: SPCA problem (VOC 2006 dataset)–Plain D^{4}L/Modified ATC algorithms: objective value, consensus disagreement, and distance from stationarity obtained after 1500 message exchanges.

6. Conclusions

This paper studied a fairly general class of distributed dictionary learning problems over time-varying multi-agent networks, with arbitrary topologies. We proposed the first decentralized algorithmic framework−the D^{4}L Algorithm−with provable convergence for this class of problems. Numerical experiments showed promising performance of our scheme with respect to state-of-the-art distributed methods.

Acknowledgments

The work of Daneshmand, Sun, and Scutari has been supported by the the National Science Foundation (NSF grants CIF 1564044 and 1632599) and the Army Research Office (Grant W911NF1810238). The work of Facchinei has been partially supported by the Italian Ministry of Education, Research and University, under the PLATINO (PLATform for INnOvative services in future internet) PON project (Grant Agreement no. PON01_01007).

Appendix A. Appendix

In this section we prove the major results of the paper, Theorems 2 and 3. We begin rewriting the D^{4}L Algorithm in an equivalent vector-matrix form (cf. Sec. A.1), which is more suitable for the analysis. Theorem 2 and Theorem 3 are then proved in Sec. A.2 and Sec. A.3, respectively. Some miscellanea results supporting the main proofs are collected in Appendix A.4. Table 10 below summarizes the symbols appearing in the proofs.

Symbol	**Definition**	**Member of**	**Reference**
$\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}$	$(1/I)\sum_{i=1}^{I}\phi_i^\nu\,\mathbf{D}_{(i)}^\nu$	$\mathcal{D}\subseteq\mathbb{R}^{M\times K}$	(41)
$\overline{\mathbf{U}}_{\boldsymbol{\phi}^\nu}$	$(1/I)\sum_{i=1}^{I}\phi_i^\nu\,\mathbf{U}_{(i)}^\nu$	$\mathcal{D}\subseteq\mathbb{R}^{M\times K}$	(41)
$\widetilde{\mathbf{D}}^\nu$	$[\widetilde{\mathbf{D}}_{(1)}^{\nu\intercal},\widetilde{\mathbf{D}}_{(2)}^{\nu\intercal},\ldots,\widetilde{\mathbf{D}}_{(I)}^{\nu\intercal}]^\intercal$	$\mathbb{R}^{M\times KI}$	(34)
$\mathbf{U}^\nu$	$[\mathbf{U}_{(1)}^{\nu\intercal},\mathbf{U}_{(2)}^{\nu\intercal},\ldots,\mathbf{U}_{(I)}^{\nu\intercal}]^\intercal$	$\mathbb{R}^{M\times KI}$	(34)
$\boldsymbol{\Theta}^\nu$	$[\boldsymbol{\Theta}_{(1)}^{\nu\intercal},\boldsymbol{\Theta}_{(2)}^{\nu\intercal},\ldots,\boldsymbol{\Theta}_{(I)}^{\nu\intercal}]^\intercal$	$\mathbb{R}^{M\times KI}$	(34)
$\mathbf{G}^\nu$	$[\nabla_D f_1(\mathbf{D}_{(1)}^\nu,\mathbf{X}_1^\nu)^\intercal,\ \ldots\ ,\ \nabla_D f_I(\mathbf{D}_{(I)}^\nu,\mathbf{X}_I^\nu)^\intercal]^\intercal$	$\mathbb{R}^{M\times KI}$	(34)
$\mathbf{W}^\nu$	$(w_{ij}^\nu)_{i,j=1}^{I}=(a_{ij}^\nu\,\phi_j^\nu/\phi_i^{\nu+1})_{i,j=1}^{I}=(\boldsymbol{\Phi}^{\nu+1})^{-1}\mathbf{A}^\nu\boldsymbol{\Phi}^\nu$	$\mathbb{R}^{I\times I}$	(35)
$\mathbf{W}^{\nu:l}$	$\mathbf{W}^\nu\cdot\mathbf{W}^{\nu-1}\cdots\mathbf{W}^l,\quad \nu>l$	$\mathbb{R}^{I\times I}$	(44)
$\widehat{\mathbf{W}}^\nu$	$\mathbf{W}^\nu\otimes\mathbf{I}_M$	$\mathbb{R}^{MI\times MI}$	(36)
$\widehat{\mathbf{W}}^{\nu:l}$	$\mathbf{W}^{\nu:l}\otimes\mathbf{I}_M,\quad \nu>l$	$\mathbb{R}^{MI\times MI}$	(44)
$\boldsymbol{\phi}^\nu$	$[\phi_1^\nu,\phi_2^\nu,\ldots,\phi_I^\nu]^\intercal$	$\mathbb{R}^{I}$	(34)
$\boldsymbol{\Phi}^\nu$	$\mathrm{diag}\,(\boldsymbol{\phi}^\nu)$	$\mathbb{R}^{I\times I}$	(34)
$\widehat{\boldsymbol{\Phi}}^\nu$	$\boldsymbol{\Phi}^\nu\otimes\mathbf{I}_M$	$\mathbb{R}^{MI\times MI}$	(36)
$\mathbf{J}_{\boldsymbol{\phi}^\nu}$	$(1/I)\mathbf{1}\boldsymbol{\phi}^{\nu\intercal}$	$\mathbb{R}^{I\times I}$	(45)
$\widehat{\mathbf{J}}_{\boldsymbol{\phi}^\nu}$	$\mathbf{J}_{\boldsymbol{\phi}^\nu}\otimes\mathbf{I}_M$	$\mathbb{R}^{MI\times MI}$	(45)

Table 10: Table of notation (appendix)

A.1. D^{4}L in vector-matrix form

We rewrite Algorithm 1 in a more convenient vector-matrix form. To do so, we introduce the following notation. Recalling the definitions of $\widetilde{\mathbf{D}}^\nu_{(i)}$ [cf. (8)], $\mathbf{U}^\nu_{(i)}$ [cf. (10)], ϕ^ν_i [cf. (12)], and $\mathbf{\Theta}^\nu_{(i)}$ [cf. (13)], define the corresponding aggregate quantities

$$\begin{aligned}
&\widetilde{\mathbf{D}}^\nu \triangleq [\widetilde{\mathbf{D}}^{\nu\intercal}_{(1)}, \widetilde{\mathbf{D}}^{\nu\intercal}_{(2)}, \ldots, \widetilde{\mathbf{D}}^{\nu\intercal}_{(I)}]^\intercal,\\
&\mathbf{U}^\nu \triangleq [\mathbf{U}^{\nu\intercal}_{(1)}, \mathbf{U}^{\nu\intercal}_{(2)}, \ldots, \mathbf{U}^{\nu\intercal}_{(I)}]^\intercal,\\
&\boldsymbol{\phi}^\nu \triangleq [\phi^\nu_1, \phi^\nu_2, \ldots, \phi^\nu_I]^\intercal, \quad \mathbf{\Phi}^\nu \triangleq \operatorname{diag}(\boldsymbol{\phi}^\nu),\\
&\mathbf{\Theta}^\nu \triangleq [\mathbf{\Theta}^{\nu\intercal}_{(1)}, \mathbf{\Theta}^{\nu\intercal}_{(2)}, \ldots, \mathbf{\Theta}^{\nu\intercal}_{(I)}]^\intercal,\\
&\mathbf{G}^\nu \triangleq \left[\nabla_D f_1(\mathbf{D}^\nu_{(1)}, \mathbf{X}^\nu_1)^\intercal,\ \nabla_D f_2(\mathbf{D}^\nu_{(2)}, \mathbf{X}^\nu_2)^\intercal,\ \ldots\ ,\ \nabla_D f_I(\mathbf{D}^\nu_{(I)}, \mathbf{X}^\nu_I)^\intercal\right]^\intercal,
\end{aligned} \tag{34}$$

where $\operatorname{diag}(\mathbf{x})$ is a diagonal matrix whose diagonal entries are the components of the vector $\mathbf{x}$. Combining the weights a^ν_{ij} and the variables ϕ^ν_i in the update of $\mathbf{D}^\nu_{(i)}$ [cf. (12)] in the single coefficient

$$w^\nu_{ij} \triangleq \frac{a^\nu_{ij}\phi^\nu_j}{\sum_k a^\nu_{ik}\phi^\nu_k},$$

we define the weight matrix $\mathbf{W}^\nu$, whose entries are $[\mathbf{W}^\nu]_{i,j} = w^\nu_{ij}$. Note that $\mathbf{W}^\nu$ has the same zero-pattern of $\mathbf{A}^\nu$, and the following properties hold (the latter under Assumption F)

$$\mathbf{W}^\nu = \left(\mathbf{\Phi}^{\nu+1}\right)^{-1}\mathbf{A}^\nu\mathbf{\Phi}^\nu, \quad \text{and} \quad \mathbf{W}^\nu\mathbf{1} = \mathbf{1}. \tag{35}$$

Finally, we define the following "augmented" matrices

$$\widehat{\mathbf{W}}^\nu \triangleq \mathbf{W}^\nu \otimes \mathbf{I}_M \quad \text{and} \quad \widehat{\mathbf{\Phi}}^\nu \triangleq \mathbf{\Phi}^\nu \otimes \mathbf{I}_M, \tag{36}$$

where $\mathbf{I}_M$ is the M-by-M identity matrix.

Using the above notation, the main iterates of the D^{4}L Algorithm, i.e., (10), (12), and (13), can be rewritten in compact form as

$$\mathbf{U}^\nu = \mathbf{D}^\nu + \gamma^\nu(\widetilde{\mathbf{D}}^\nu - \mathbf{D}^\nu), \tag{37}$$

$$\boldsymbol{\phi}^{\nu+1} = \mathbf{A}^\nu\boldsymbol{\phi}^\nu, \tag{38}$$

$$\mathbf{D}^{\nu+1} = \widehat{\mathbf{W}}^\nu\mathbf{U}^\nu, \tag{39}$$

$$\mathbf{\Theta}^{\nu+1} = \widehat{\mathbf{W}}^\nu\mathbf{\Theta}^\nu + \left(\widehat{\mathbf{\Phi}}^{\nu+1}\right)^{-1}\left(\mathbf{G}^{\nu+1} - \mathbf{G}^\nu\right). \tag{40}$$

Instrumental to the analysis of the consensus disagreement are the following weighted average quantities:

$$\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu} \triangleq \frac{1}{I}\sum_{i=1}^I \phi^\nu_i\,\mathbf{D}^\nu_{(i)}, \quad \overline{\mathbf{U}}_{\boldsymbol{\phi}^\nu} \triangleq \frac{1}{I}\sum_{i=1}^I \phi^\nu_i\,\mathbf{U}^\nu_{(i)}. \tag{41}$$

Using (12), (41) and the column stochasticity of $\mathbf{A}^\nu$ (cf. Assumption F3), it is not difficult to check that

$$\overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu+1}} = \overline{\mathbf{U}}_{\boldsymbol{\phi}^\nu}, \tag{42}$$

which, together with (10), leads to the following dynamics for $\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}$:

$$\overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu+1}} = \overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu}} + \frac{\gamma^\nu}{I}\sum_{i=1}^{I}\phi_i^\nu\left(\widetilde{\mathbf{D}}_{(i)}^\nu - \mathbf{D}_{(i)}^\nu\right). \tag{43}$$

A.2. Proof of Theorem 2

We prove Theorem 2 in the following order (consistent with the statements in the theorem):

Step 1–Asymptotic consensus & related properties: We prove that consensus is asymptotically achieved, that is, $\lim_{\nu\to\infty} e^\nu = 0$ [statement (a)], along with some properties on related quantities, which will be used in the other steps–see Sec. A.2.1;

Step 2–Boundedness of the iterates and Lipschitz continuity of ∇f_i: We prove that the sequence $\left\{\left(\mathbf{D}^\nu, \mathbf{X}^\nu\right)\right\}_\nu$ generated by the algorithm is bounded [statement (b-i)], and, as a consequence, ∇f_i, $\nabla_D \tilde{f}_i$, and $\nabla_{X_i}\tilde{h}_i$ are Lipschitz continuous on a (suitably defined) compact set containing $\left\{\left(\mathbf{D}^\nu, \mathbf{X}^\nu\right)\right\}_\nu$ [Remarks 8 and 9]–see Sec. A.2.2;

Step 3–Decrease of $\{U(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu)\}_\nu$: By leveraging the Lipschitz continuity of the gradients as in Step 2, we study the decrease properties of $\{U(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu)\}_\nu$ [statement (b-ii)]–see Sec. A.2.3;

Step 4–Vanishing X-stationarity: We prove that $\lim_{\nu\to\infty}\Delta_X(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu) = 0$ [statement (b-iii)]–see Sec. A.2.4;

Step 5–Vanishing liminf D-stationarity: We prove $\liminf_{\nu\to\infty}\Delta_D(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu) = 0$ [statement (b-iv)]–see Sec. A.2.5;

Step 6–Vanishing D-stationarity: Finally, we prove $\lim_{\nu\to\infty}\Delta_D(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu) = 0$ [statement (b')]–see Sec. A.2.6.

A.2.1. Step 1–Asymptotic consensus and related properties

1) Preliminaries: To analyze the dynamics of the consensus disagreement, we first introduce the following product matrices and their augmented counterparts: Given $\mathbf{W}^\nu$ [cf. (35)], and $\nu, l \in \mathbb{N}_+$, let

$$\mathbf{W}^{\nu:l} \triangleq \begin{cases}\mathbf{W}^\nu\cdot\mathbf{W}^{\nu-1}\cdots\mathbf{W}^l, & \nu > l,\\ \mathbf{W}^\nu, & \nu = l,\\ \mathbf{0}_I, & \nu < l,\end{cases} \quad \text{and} \quad \widehat{\mathbf{W}}^{\nu:l} \triangleq \mathbf{W}^{\nu:l}\otimes\mathbf{I}_M. \tag{44}$$

Define also the weight-average matrices

$$\mathbf{J}_{\boldsymbol{\phi}^\nu} \triangleq \frac{1}{I}\mathbf{1}\boldsymbol{\phi}^{\nu\intercal}, \quad \widehat{\mathbf{J}}_{\boldsymbol{\phi}^\nu} \triangleq \mathbf{J}_{\boldsymbol{\phi}^\nu}\otimes\mathbf{I}_M. \tag{45}$$

For notational simplicity, when $\boldsymbol{\phi}^\nu = \mathbf{1}$, we use $\mathbf{J}$ instead of $\mathbf{J}_{\boldsymbol{\phi}^\nu}$ and $\widehat{\mathbf{J}} \triangleq \mathbf{J}\otimes\mathbf{I}_M$. Using (35) and (41), it is not difficult to check that the following hold:

$$\widehat{\mathbf{J}}_{\boldsymbol{\phi}^\nu}\mathbf{D}^\nu = \mathbf{1}\otimes\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}, \tag{46}$$

$$\widehat{\mathbf{J}}_{\boldsymbol{\phi}^{\nu+1}}\widehat{\mathbf{W}}^{\nu:l} = \widehat{\mathbf{J}}\,\widehat{\boldsymbol{\Phi}}^l = \widehat{\mathbf{J}}_{\boldsymbol{\phi}^l}. \tag{47}$$

The dynamics of the consensus disagreement e^ν boils down to studying the decay of $\|\mathbf{W}^{\nu:l} - \mathbf{J}_{\boldsymbol{\phi}^l}\|_2$ (this will be clear in the proof of Proposition 5 below). The following lemma shows that $\mathbf{W}^{\nu:l}$ converges geometrically to $\mathbf{J}_{\boldsymbol{\phi}^l}$, as $\nu \to \infty$.

Lemma 4 (Scutari and Sun (2018)-Lemma 4.13, Ch. 3.4.2.5) *Let $\{\mathcal{G}^\nu\}_\nu$ be a sequence of digraphs satisfying Assumption B; let $\{\mathbf{A}^\nu\}_\nu$ be a sequence of matrices satisfying Assumption F; and let $\{\mathbf{W}^\nu\}_\nu$ be the sequence of matrices defined in* (35). *Then, there holds*

$$\left\|\mathbf{W}^{\nu:l} - \mathbf{J}_{\boldsymbol{\phi}^l}\right\|_2 \leq c_W\,(\rho)^{\nu-l+1}, \quad \forall \nu \geq l,\ \nu, l \in \mathbb{N}_+,$$

where $c_W > 0$ is a (proper) constant, and $\rho \in (0,1)$ is defined as

$$\rho = \left(1 - \tilde{\kappa}^{-(I-1)B}\right)^{\frac{1}{B(I-1)}} < 1, \tag{48}$$

with $\tilde{\kappa} = \kappa^{IB+1}/I$; and κ is defined in Assumption F.

Furthermore, the sequence $\{\phi_i^\nu\}_\nu$ satisfies

$$\underline{\epsilon}_\phi \triangleq \inf_{\nu \in \mathbb{N}_+}\left(\min_{1\leq i\leq I} \phi_i^\nu\right) \geq \kappa^{IB}, \quad \textit{and} \quad \bar{\epsilon}_\phi \triangleq \sup_{\nu \in \mathbb{N}_+}\left(\max_{1\leq i\leq I} \phi_i^\nu\right) \leq I - (I-1)\kappa^{IB}. \tag{49}$$

If all the matrices $\mathbf{A}^\nu$ are doubly-stochastic, then $\underline{\epsilon}_\phi = \bar{\epsilon}_\phi = 1$.

2) Proof of $\lim_{\nu\to\infty} e^\nu = 0$. Using (30), we can write

$$\begin{aligned} e^\nu &\overset{(a)}{\leq} K\left\|\mathbf{D}^\nu - \mathbf{1}\otimes\overline{\mathbf{D}}^\nu\right\|_F \overset{(b)}{\leq} K\left\|\mathbf{D}^\nu - \mathbf{1}\otimes\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}\right\|_F + K\sqrt{I}\left\|\overline{\mathbf{D}}^\nu - \overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}\right\|_F \\ &\overset{(c)}{\leq} 2K\left\|\mathbf{D}^\nu - \mathbf{1}\otimes\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}\right\|_F, \end{aligned} \tag{50}$$

where (a) follows from the equivalence of norms; (b) is due to the triangle inequality; and in (c) we used $\sum_{i=1}^I a_i \leq \sqrt{I}||\mathbf{a}||$, with $\mathbf{a} = (a_i)_{i=1}^I \in \mathbb{R}^I$.

The following proposition concludes the proof of statement (a), proving that $\|\mathbf{D}^\nu - \mathbf{1}\otimes\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}\|_F$ is square summable, along with some additional properties on related quantities.

Proposition 5 *In the above setting, there hold:*

$$\lim_{\nu\to\infty} ||\mathbf{D}^\nu - \mathbf{1}\otimes\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}||_F = 0; \tag{51}$$

$$\lim_{\nu\to\infty} \sum_{t=1}^{\nu} ||\mathbf{D}^t - \mathbf{1}\otimes\overline{\mathbf{D}}_{\boldsymbol{\phi}^t}||_F^2 < \infty; \tag{52}$$

$$\lim_{\nu\to\infty} \sum_{t=1}^{\nu} ||\mathbf{U}^t - \mathbf{1}\otimes\overline{\mathbf{U}}_{\boldsymbol{\phi}^t}||_F^2 < \infty. \tag{53}$$

Proof To prove (51), let us first expand $\mathbf{D}^\nu - \mathbf{1}\otimes\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}$ as follows: for any $\nu \geq 1$,

$$\begin{aligned} \mathbf{D}^\nu &\overset{(39)}{=} \widehat{\mathbf{W}}^{\nu-1}\mathbf{U}^{\nu-1} = \widehat{\mathbf{W}}^{\nu-1}\mathbf{D}^{\nu-1} + \widehat{\mathbf{W}}^{\nu-1}\left(\mathbf{U}^{\nu-1} - \mathbf{D}^{\nu-1}\right) \\ &= \widehat{\mathbf{W}}^{\nu-1:0}\mathbf{D}^0 + \sum_{t=0}^{\nu-1}\widehat{\mathbf{W}}^{\nu-1:t}\left(\mathbf{U}^t - \mathbf{D}^t\right), \end{aligned} \tag{54}$$

where the last equality follows from induction and the definition of $\widehat{\mathbf{W}}^{\nu:l}$ [cf. (44)]. Similarly, we expand the subtrahend as

$$\begin{aligned}\mathbf{1}\otimes\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu} &\overset{(46)}{=} \widehat{\mathbf{J}}_{\boldsymbol{\phi}^\nu}\mathbf{D}^\nu \overset{(54)}{=} \widehat{\mathbf{J}}_{\boldsymbol{\phi}^\nu}\left(\widehat{\mathbf{W}}^{\nu-1:0}\mathbf{D}^0+\sum_{t=0}^{\nu-1}\widehat{\mathbf{W}}^{\nu-1:t}\left(\mathbf{U}^t-\mathbf{D}^t\right)\right)\\ &\overset{(47)}{=} \widehat{\mathbf{J}}\left(\mathbf{D}^0+\sum_{t=0}^{\nu-1}\widehat{\boldsymbol{\Phi}}^t\left(\mathbf{U}^t-\mathbf{D}^t\right)\right).\end{aligned}\tag{55}$$

Subtracting (55) from (54) and using (37), yields

$$\begin{aligned}\left\|\mathbf{D}^\nu-\mathbf{1}\otimes\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}\right\|_F &\leq \left\|\widehat{\mathbf{W}}^{\nu-1:0}-\widehat{\mathbf{J}}\right\|_2\left\|\mathbf{D}^0\right\|_F+\sum_{t=0}^{\nu-1}\gamma^t\left\|\widehat{\mathbf{W}}^{\nu-1:t}-\widehat{\mathbf{J}}_{\phi^t}\right\|_2\left\|\widetilde{\mathbf{D}}^t-\mathbf{D}^t\right\|_F\\ &\overset{(a)}{\leq} c_1(\rho)^\nu+c_2\sum_{t=0}^{\nu-1}\gamma^t\,(\rho)^{\nu-t}\xrightarrow[\nu\to\infty]{(b)}0,\end{aligned}\tag{56}$$

for some finite constants $c_1, c_2 > 0$, where (a) is due to Lemma 4 and the boundedness of $\{||\widetilde{\mathbf{D}}^\nu-\mathbf{D}^\nu||\}_\nu$; and (b) follows from Lemma 13(a) in Appendix A.4.

Let us now proceed to prove (52). Using (56), we have

$$\begin{aligned}&\lim_{\nu\to\infty}\sum_{t=1}^{\nu}\left\|\mathbf{D}^t-\mathbf{1}\otimes\overline{\mathbf{D}}_{\boldsymbol{\phi}^t}\right\|_F^2 \leq \lim_{\nu\to\infty}\sum_{t=1}^{\nu}\left(c_1(\rho)^t+c_2\sum_{l=0}^{t-1}\gamma^l\,(\rho)^{t-l}\right)^2\\ &\overset{(a)}{\leq}\frac{2c_1^2}{1-(\rho)^2}+2\,c_2^2\lim_{\nu\to\infty}\sum_{t=1}^{\nu}\sum_{l=0}^{t-1}\sum_{k=0}^{t-1}\gamma^l\gamma^k(\rho)^{t-k}(\rho)^{t-l}\\ &\overset{(b)}{\leq}\frac{2c_1^2}{1-(\rho)^2}+c_2^2\lim_{\nu\to\infty}\sum_{t=1}^{\nu}\sum_{l=0}^{t-1}(\gamma^l)^2(\rho)^{t-l}\sum_{k=0}^{t-1}(\rho)^{t-k}+c_2^2\lim_{\nu\to\infty}\sum_{t=1}^{\nu}\sum_{k=0}^{t-1}(\gamma^k)^2(\rho)^{t-k}\sum_{l=0}^{t-1}(\rho)^{t-l}\\ &\leq\frac{2c_1^2}{1-(\rho)^2}+\frac{2c_2^2}{1-\rho}\lim_{\nu\to\infty}\sum_{t=1}^{\nu}\sum_{l=0}^{t-1}(\gamma^l)^2(\rho)^{t-l}\overset{(c)}{<}\infty,\end{aligned}$$

where in (a) and (b) we used $(a+b)^2 \leq 2(a^2+b^2)$ and $ab \leq (a^2+b^2)/2$, respectively, and (c) is due to Lemma 13(b) (cf. Appendix A.4).

We prove now (53). Using (41) and (37), we get

$$\begin{aligned}\left\|\mathbf{U}^t-\mathbf{1}\otimes\overline{\mathbf{U}}_{\boldsymbol{\phi}^t}\right\|_F^2 &= \left\|\gamma^t\left(\widetilde{\mathbf{D}}^t-\mathbf{1}\otimes\frac{1}{I}\sum_{i=1}^{I}\phi_i^t\,\widetilde{\mathbf{D}}_{(i)}^t\right)+\left(1-\gamma^t\right)\left(\mathbf{D}^t-\mathbf{1}\otimes\overline{\mathbf{D}}_{\boldsymbol{\phi}^t}\right)\right\|_F^2\\ &\leq 2\left(\gamma^t\right)^2\left\|\widetilde{\mathbf{D}}^t-\mathbf{1}\otimes\frac{1}{I}\sum_{i=1}^{I}\phi_i^t\,\widetilde{\mathbf{D}}_{(i)}^t\right\|_F^2+2\left\|\mathbf{D}^t-\mathbf{1}\otimes\overline{\mathbf{D}}_{\boldsymbol{\phi}^t}\right\|_F^2,\end{aligned}\tag{57}$$

where in the last inequality we used Jensen's inequality and $(1-\gamma^t) \leq 1$. Therefore,

$$\lim_{\nu\to\infty}\sum_{t=1}^{\nu}\left\|\mathbf{U}^t-\mathbf{1}\otimes\overline{\mathbf{U}}_{\boldsymbol{\phi}^t}\right\|_F^2 \leq \lim_{\nu\to\infty}2\,c_3\sum_{t=1}^{\nu}\left(\gamma^t\right)^2+\lim_{\nu\to\infty}2\sum_{t=1}^{\nu}\left\|\mathbf{D}^t-\mathbf{1}\otimes\overline{\mathbf{D}}_{\boldsymbol{\phi}^t}\right\|_F^2\overset{(a)}{<}\infty,$$

where c_3 is a positive finite constant, and (a) follows from Assumption E and (52). ■

Remark 6 (On $L_{\nabla X_i}(\overline{\mathbf{U}}_{\boldsymbol{\phi}^\nu})$ and $L_{\nabla X_i}(\mathbf{U}_{(i)}^\nu)$) *Recall that $\nabla_{X_i} f_i(\overline{\mathbf{U}}_{\boldsymbol{\phi}^\nu}, \bullet)$ is Lipschitz continuous on $\mathcal{X}_i$, with constant $L_{\nabla X_i}(\overline{\mathbf{U}}_{\boldsymbol{\phi}^\nu})$ (cf. Assumption A2). Since $\|\mathbf{U}_{(i)}^\nu - \overline{\mathbf{U}}_{\boldsymbol{\phi}^\nu}\|_F \underset{\nu\to\infty}{\longrightarrow} 0$ [cf. (53), Proposition 5] and $L_{\nabla X_i}(\mathbf{D})$ is continuous, we have*

$$\left| L_{\nabla X_i}(\overline{\mathbf{U}}_{\boldsymbol{\phi}^\nu}) - L_{\nabla X_i}(\mathbf{U}_{(i)}^\nu) \right| \underset{\nu\to\infty}{\longrightarrow} 0, \quad i = 1, 2, \ldots I. \tag{58}$$

A.2.2. Step 2–Boundedness of the iterates

We show that the sequence $\left\{ \left(\mathbf{D}^\nu, \mathbf{X}^\nu\right) \right\}_\nu$ generated by the D^4L Algorithm is bounded [statement (b-i)]. We prove the result only for $\tilde{h}_i$ given by (17); the proof can be easily tailored to the other choice of $\tilde{h}_i$. If the sets $\mathcal{X}_i$ are bounded [Assumption A5(i)], the result follows readily. Therefore, we consider next the setting under A5(ii).

Throughout the proof, we will use the following properties of $\tilde{f}_i$ and $\tilde{h}_i$.

Remark 7 *The surrogate functions $\tilde{f}_i$ and $\tilde{h}_i$ as in Assumption C have the following properties: for all $i = 1, 2, \ldots, I$,*

(a) $\tilde{f}_i(\bullet; \mathbf{D}, \mathbf{X}_i)$ is strongly convex on $\mathcal{D}$, uniformly with respect to $(\mathbf{D}, \mathbf{X}_i) \in \mathcal{D} \times \mathcal{X}_i$, with constant $\tau_{D,i}^\nu > 0$; and $\nabla_D \tilde{f}_i(\mathbf{D}; \mathbf{D}, \mathbf{X}_i) = \nabla_D f_i(\mathbf{D}, \mathbf{X}_i)$, for all $(\mathbf{D}, \mathbf{X}_i) \in \mathcal{D} \times \mathcal{X}_i$.

(b) $\tilde{h}_i(\bullet; \mathbf{D}, \mathbf{X}_i)$ is strongly convex on $\mathcal{X}_i$, uniformly with respect to $(\mathbf{D}, \mathbf{X}_i) \in \mathcal{D} \times \mathcal{X}_i$, with constant $\tau_{X,i}^\nu > 0$; and $\nabla_{X_i} \tilde{h}_i(\mathbf{X}_i; \mathbf{D}, \mathbf{X}_i) = \nabla_{X_i} f_i(\mathbf{D}, \mathbf{X}_i)$, for all $(\mathbf{D}, \mathbf{X}_i) \in \mathcal{D} \times \mathcal{X}_i$.

By the optimality of $\mathbf{X}_i^{\nu+1}$ in (11), there exist $\boldsymbol{\Xi}_i^0 \in \partial_{X_i} g_i(\mathbf{X}_i^0)$ and $\boldsymbol{\Xi}_i^{\nu+1} \in \partial_{X_i} g_i(\mathbf{X}_i^{\nu+1})$ such that

$$\begin{aligned}
0 &\le \left\langle \nabla_{X_i} \tilde{h}_i(\mathbf{X}_i^{\nu+1}; \mathbf{U}_{(i)}^\nu, \mathbf{X}_i^\nu) + \boldsymbol{\Xi}_i^{\nu+1}, \mathbf{X}_i^0 - \mathbf{X}_i^{\nu+1} \right\rangle \\
&= \left\langle \boldsymbol{\Xi}_i^{\nu+1} - \boldsymbol{\Xi}_i^0 + \tau_{X,i}^\nu (\mathbf{X}_i^{\nu+1} - \mathbf{X}_i^0), \mathbf{X}_i^0 - \mathbf{X}_i^{\nu+1} \right\rangle \\
&\quad + \left\langle \nabla_{X_i} f_i(\mathbf{U}_{(i)}^\nu, \mathbf{X}_i^\nu) - \nabla_{X_i} f_i(\mathbf{U}_{(i)}^\nu, \mathbf{X}_i^0), \mathbf{X}_i^0 - \mathbf{X}_i^{\nu+1} \right\rangle \\
&\quad - \left\langle \tau_{X,i}^\nu (\mathbf{X}_i^\nu - \mathbf{X}_i^0), \mathbf{X}_i^0 - \mathbf{X}_i^{\nu+1} \right\rangle \\
&\quad + \left\langle \nabla_{X_i} f_i(\mathbf{U}_{(i)}^\nu, \mathbf{X}_i^0) + \boldsymbol{\Xi}_i^0, \mathbf{X}_i^0 - \mathbf{X}_i^{\nu+1} \right\rangle.
\end{aligned}$$

Using Remark 7(b) and the μ_i-strongly convexity of g_i's, we obtain

$$\begin{aligned}
(\tau_{X,i}^\nu + \mu_i) \left\| \mathbf{X}_i^{\nu+1} - \mathbf{X}_i^0 \right\|_F^2 \le &\left\langle \tau_{X,i}^\nu \mathbf{X}_i^\nu - \nabla_{X_i} f_i(\mathbf{U}_{(i)}^\nu, \mathbf{X}_i^\nu), \mathbf{X}_i^{\nu+1} - \mathbf{X}_i^0 \right\rangle \\
&- \left\langle \tau_{X,i}^\nu \mathbf{X}_i^0 - \nabla_{X_i} f_i(\mathbf{U}_{(i)}^\nu, \mathbf{X}_i^0), \mathbf{X}_i^{\nu+1} - \mathbf{X}_i^0 \right\rangle \\
&- \left\langle \underbrace{\nabla_{X_i} f_i(\mathbf{U}_{(i)}^\nu, \mathbf{X}_i^0) + \boldsymbol{\Xi}_i^0}_{\triangleq \mathbf{Z}_i^\nu \in \partial_{X_i} U(\mathbf{U}_{(i)}^\nu, \mathbf{X}_i^0)}, \mathbf{X}_i^{\nu+1} - \mathbf{X}_i^0 \right\rangle.
\end{aligned} \tag{59}$$

Define $\Upsilon_i^\nu(\mathbf{X}_i) \triangleq \tau_{X,i}^\nu \mathbf{X}_i - \nabla_{X_i} f_i(\mathbf{U}_{(i)}^\nu, \mathbf{X}_i)$ and rewrite (59) as

$$(\tau_{X,i}^\nu + \mu_i) \|\mathbf{X}_i^{\nu+1} - \mathbf{X}_i^0\|_F \le \left\| \Upsilon_i^\nu(\mathbf{X}_i^\nu) - \Upsilon_i^\nu(\mathbf{X}_i^0) \right\|_F + \|\mathbf{Z}_i^\nu\|_F. \tag{60}$$

Since $\mathcal{D}$ is compact and $\mathbf{X}_i^0$ is given, we have $||\mathbf{Z}_i^\nu||_F \leq B_Z$, for all i, $\nu \geq 1$, and some finite $B_Z > 0$. Let us bound next $||\Upsilon_i^\nu(\mathbf{X}_i^\nu) - \Upsilon_i^\nu(\mathbf{X}_i^0)||_F$. We write

$$\begin{aligned}
\left\|\Upsilon_i^\nu(\mathbf{X}_i^\nu) - \Upsilon_i^\nu(\mathbf{X}_i^0)\right\|_F^2 = (\tau_{X,i}^\nu)^2 \left\|\mathbf{X}_i^\nu - \mathbf{X}_i^0\right\|_F^2 + \left\|\nabla_{X_i} f_i(\mathbf{U}_{(i)}^\nu, \mathbf{X}_i^\nu) - \nabla_{X_i} f_i(\mathbf{U}_{(i)}^\nu, \mathbf{X}_i^0)\right\|_F^2 \\
- 2\tau_{X,i}^\nu \left\langle \nabla_{X_i} f_i(\mathbf{U}_{(i)}^\nu, \mathbf{X}_i^\nu) - \nabla_{X_i} f_i(\mathbf{U}_{(i)}^\nu, \mathbf{X}_i^0), \mathbf{X}_i^\nu - \mathbf{X}_i^0 \right\rangle \\
\stackrel{(a)}{\leq} (\tau_{X,i}^\nu)^2 \left\|\mathbf{X}_i^\nu - \mathbf{X}_i^0\right\|_F^2 \\
+ \left(1 - \frac{2\tau_{X,i}^\nu}{L_{\nabla X_i}(\mathbf{U}_{(i)}^\nu)}\right) \left\|\nabla_{X_i} f_i(\mathbf{U}_{(i)}^\nu, \mathbf{X}_i^\nu) - \nabla_{X_i} f_i(\mathbf{U}_{(i)}^\nu, \mathbf{X}_i^0)\right\|_F^2,
\end{aligned} \tag{61}$$

where in (a) we used the $1/L_{\nabla X_i}(\mathbf{U}_{(i)}^\nu)$-co-coercivity of $\nabla_{X_i} f_i(\mathbf{U}_{(i)}^\nu, \bullet)$ [due to the convexity of $f_i(\mathbf{U}_{(i)}^\nu, \bullet)$ and the $L_{\nabla X_i}(\mathbf{U}_{(i)}^\nu)$-Lipschitianity of $\nabla_{X_i} f_i(\mathbf{U}_{(i)}^\nu, \bullet)$ (Rockafellar and Wets, 1998, Prop.12.60)], i.e.,

$$\begin{aligned}
\left\langle \nabla_{X_i} f_i(\mathbf{U}_{(i)}^\nu, \mathbf{X}_i) - \nabla_{X_i} f_i(\mathbf{U}_{(i)}^\nu, \mathbf{Y}_i), \mathbf{X}_i - \mathbf{Y}_i \right\rangle \geq \\
\frac{1}{L_{\nabla X_i}(\mathbf{U}_{(i)}^\nu)} \cdot \left\|\nabla_{X_i} f_i(\mathbf{U}_{(i)}^\nu, \mathbf{X}_i) - \nabla_{X_i} f_i(\mathbf{U}_{(i)}^\nu, \mathbf{Y}_i)\right\|_F^2, \qquad \forall \mathbf{X}_i, \mathbf{Y}_i \in \mathcal{X}_i.
\end{aligned}$$

Note that $||\Upsilon_i^\nu(\mathbf{X}_i^\nu) - \Upsilon_i^\nu(\mathbf{X}_i^0)||_F \leq \tau_{X,i}^\nu ||\mathbf{X}_i^\nu - \mathbf{X}_i^0||_F$ as long as $\tau_{X,i}^\nu \geq \frac{1}{2} L_{\nabla X_i}(\mathbf{U}_{(i)}^\nu)$, which is satisfied under D1. Therefore, we can bound (60) as

$$(\tau_{X,i}^\nu + \mu_i)||\mathbf{X}_i^{\nu+1} - \mathbf{X}_i^0||_F \leq \tau_{X,i}^\nu \left\|\mathbf{X}_i^\nu - \mathbf{X}_i^0\right\|_F + B_Z. \tag{62}$$

We can now prove that, starting from $\mathbf{X}_i^0$, the iterates $\mathbf{X}_i^\nu$ stays in the ball $\mathcal{B}_i(R_i, \mathbf{X}_i^0) \triangleq \{\mathbf{X}_i \in \mathbb{R}^{K \times n_i} : \|\mathbf{X}_i - \mathbf{X}_i^0\|_F \leq R_i\}$, for all $\nu \geq 1$, where $R_i \geq B_Z/\mu_i$. Let us prove it by induction. Evidently $\mathbf{X}_i^0 \in \mathcal{B}_i(R_i, \mathbf{X}_i^0)$. Let $\mathbf{X}_i^\nu \in \mathcal{B}_i(R_i, \mathbf{X}_i^0)$; by (62), we get

$$||\mathbf{X}_i^{\nu+1} - \mathbf{X}_i^0||_F \leq \frac{\tau_{X,i}^\nu}{\tau_{X,i}^\nu + \mu_i} ||\mathbf{X}_i^\nu - \mathbf{X}_i^0||_F + \frac{B_Z}{\tau_{X,i}^\nu + \mu_i} \leq R_i,$$

where the second inequality is due to $\mathbf{X}_i^\nu \in \mathcal{B}_i(R, \mathbf{X}_i^0)$ and $R \geq B_Z/\mu$. Hence $\mathbf{X}_i^{\nu+1} \in \mathcal{B}_i(R_i, \mathbf{X}_i^0)$. Therefore, $\mathbf{X}_i^\nu \in \mathcal{B}_i(R_i, \mathbf{X}_i^0)$, for all $\nu \geq 0$. Since $\mathcal{D}$ is bounded (cf. Assumption A3), it follows that $(\mathbf{D}_{(i)}^\nu, \mathbf{X}_i^\nu) \in \mathcal{D} \times \mathcal{B}_i(R_i, \mathbf{X}_i^0)$, for all $\nu \geq 0$. □

Remark 8 (On the Lipschitz continuity of ∇f_i's) *Since f_i is $\mathcal{C}^2$, a direct consequence of the boundedness of $\{(\mathbf{D}^\nu, \mathbf{X}^\nu)\}_\nu$ is that ∇f_i [the gradient of f_i with respect to $(\mathbf{D}, \mathbf{X}_i)$] is Lipschitz continuous on $\mathcal{D} \times \mathcal{B}_i(R_i, \mathbf{X}_i^0)$, that is, there exists some positive finite constant $L_{\nabla,i}$ such that*

$$||\nabla f_i(\mathbf{D}, \mathbf{X}_i) - \nabla f_i(\mathbf{D}', \mathbf{X}_i')||_F \leq L_{\nabla,i} ||(\mathbf{D}, \mathbf{X}_i) - (\mathbf{D}', \mathbf{X}_i')||_F, \tag{63}$$

for all $(\mathbf{D}, \mathbf{X}_i), (\mathbf{D}', \mathbf{X}_i') \in \mathcal{D} \times \mathcal{B}_i(R_i, \mathbf{X}_i^0)$, and $i = 1, 2, \ldots, I$. We define $L_\nabla \triangleq \max_i L_{\nabla,i}$.

The above result also implies that $\nabla_D F : \mathcal{D} \times (\mathcal{X}_1 \times \cdots \times \mathcal{X}_I) \to \mathcal{D}$ [cf. (P)] is Lipschitz continuous on $\mathcal{D} \times (\mathcal{B}_1(R_1, \mathbf{X}_1^0) \times \cdots \times \mathcal{B}_I(R_I, \mathbf{X}_I^0))$, with constant $L_{\nabla_D} \triangleq I \cdot L_\nabla$.

Remark 9 (On the Lipschitz continuity of $\nabla_D \tilde{f}_i$ and $\nabla_{X_i}\tilde{h}_i$) *$\nabla_D \tilde{f}_i : \mathcal{D} \times \mathbb{R}^{K\times n_i} \times \mathbb{R}^{K\times n_i} \to \mathbb{R}^{M\times K}$ and $\nabla_{X_i}\tilde{h}_i : \mathbb{R}^{K\times n_i} \times \mathcal{D} \times \mathbb{R}^{K\times n_i} \to \mathbb{R}^{K\times n_i}$ are Lipschitz continuous on $\mathcal{D}\times\mathcal{D}\times\mathcal{B}_i(R_i, \mathbf{X}_i^0)$ and $\mathcal{B}_i(R_i, \mathbf{X}_i^0)\times\mathcal{D}\times\mathcal{B}_i(R_i, \mathbf{X}_i^0)$, respectively, with constants $\tilde{L}^D_{\nabla,i}$ and $\tilde{L}^X_{\nabla,i}$. Let us denote $\tilde{L}^D_{\nabla} \triangleq \max_i \ \tilde{L}^D_{\nabla,i}$ and $\tilde{L}^X_{\nabla} \triangleq \max_i \ \tilde{L}^X_{\nabla,i}$.*

A.2.3. Step 3–Decrease of $\{U(\overline{\mathbf{D}}^{\nu}, \mathbf{X}^{\nu})\}_{\nu}$

We study here the properties of $\{U(\overline{\mathbf{D}}^{\nu}, \mathbf{X}^{\nu})\}_{\nu}$ showing, in particular, that it is convergent [statement (b-ii)].

We begin with the following intermediate result.

Proposition 10 *Consider the setting of Theorem 2(b); there exist positive constants s_X, $\bar{\tau}_D$, c_7, c_8, and a sufficiently large $\bar{\nu}\in\mathbb{N}_+$ such that the following holds: for all $\nu \geq \bar{\nu}$,*

$$U(\overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu+1}}, \mathbf{X}^{\nu+1}) \leq U(\overline{\mathbf{D}}_{\boldsymbol{\phi}^{\bar{\nu}}}, \mathbf{X}^{\bar{\nu}}) - \sum_{l=\bar{\nu}}^{\nu} Y^l + \sum_{l=\bar{\nu}}^{\nu} W^l + E^{\nu,\bar{\nu}}, \tag{64}$$

where

$$Y^l \triangleq s_X \left(||\mathbf{X}^{l+1} - \mathbf{X}^l||_F - \frac{Z^l}{2s_X}\right)^2 + \frac{\bar{\tau}_D}{I}\gamma^l \left(||\widetilde{\mathbf{D}}^l - \mathbf{D}^l||_F - \frac{I\bar{\epsilon}_\phi}{2\,\bar{\tau}_D}T^l\right)^2, \tag{65}$$

$$Z_l \triangleq c_7 \sum_{t=l}^{\infty} \gamma^t (\rho)^{t-l} + L_X \|\mathbf{U}^l - \mathbf{1}\otimes\overline{\mathbf{U}}_{\boldsymbol{\phi}^l}\|_F, \tag{66}$$

$$W^l \triangleq \frac{I\,\bar{\epsilon}_\phi^2}{4\bar{\tau}_D}\gamma^l \left(T^l\right)^2 + \frac{1}{4\,s_X}\left(Z^l\right)^2 + L_G\gamma^l \left\| \mathbf{D}^l - \mathbf{1}\otimes\overline{\mathbf{D}}_{\boldsymbol{\phi}^l}\right\|_F, \tag{67}$$

$$E^{\nu,\bar{\nu}} \triangleq \left[\frac{c_6}{1-\rho}((\rho)^{\bar{\nu}} - (\rho)^{\nu+1}) + \frac{c_7 B_X}{(1-\rho)^2}\right]\cdot\left(\max_{t\geq\bar{\nu}} \gamma^t\right), \tag{68}$$

$$\{T^\nu\}_\nu \ \ \text{is such that} \ \ \sum_{\nu=1}^{\infty} (T^\nu)^2 < \infty, \tag{69}$$

with $\rho\in(0,1)$ and $\bar{\epsilon}_\phi$ defined in (48) and (49), respectively.

Proof See Appendix A.4.3 ■

Note that the sequences $\{Z^l\}_l$, $\{W^l\}_l$, and $\{E^{\nu,\bar{\nu}}\}_{\nu,\bar{\nu}}$ above satisfy

$$\lim_{\nu\to\infty} \sum_{l=\bar{\nu}}^{\nu} (Z^l)^2 < \infty; \tag{70}$$

$$\lim_{\nu\to\infty} \sum_{l=\bar{\nu}}^{\nu} W^l < \infty; \tag{71}$$

$$\lim_{\bar{\nu}\to\infty} \underbrace{\left(\lim_{\nu\to\infty} E^{\nu,\bar{\nu}}\right)}_{\triangleq E^{\infty,\bar{\nu}} < \infty} = 0, \tag{72}$$

where (70) follows from (53) [cf. Prop. 5], Assumption E, and Lemma 13(b) [cf. (107)] in Appendix A.4.1; (71) is a consequence of $\lim_{\nu\to\infty}\sum_{l=\nu_1}^{\nu}\gamma^l(T^l)^2<\infty$ [due to (69)], (52) [cf. Prop. 5], and (70); and eq. (72) is proved by inspection.

It follows from (64), (70)-(72) and the lower-boundedness of U (due to Assumption A1) that $\{U(\overline{\mathbf{D}}^{\nu},\mathbf{X}^{\nu})\}_{\nu}$ is convergent. Indeed, taking the limsup of the LHS of (64) and using (71) and (72), we get

$$-\infty<\limsup_{\nu\to\infty}U(\overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu+1}},\mathbf{X}^{\nu+1})\leq U(\overline{\mathbf{D}}_{\boldsymbol{\phi}^{\bar{\nu}}},\mathbf{X}^{\bar{\nu}})+\sum_{l=\bar{\nu}}^{\infty}W^l+E^{\infty,\bar{\nu}}<\infty.$$

Taking now the liminf of the RHS of the above inequality with respect to $\bar{\nu}$ while using (72) and $\lim_{\bar{\nu}\to\infty}\sum_{l=\bar{\nu}}^{\infty}W^l=0$, yields

$$-\infty<\limsup_{\nu\to\infty}\,U(\overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu+1}},\mathbf{X}^{\nu+1})\leq\liminf_{\bar{\nu}\to\infty}\,U(\overline{\mathbf{D}}_{\boldsymbol{\phi}^{\bar{\nu}}},\mathbf{X}^{\bar{\nu}})<\infty,$$

which implies the convergence of $\{U(\overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu}},\mathbf{X}^{\nu})\}_{\nu}$ to a finite value, and

$$\lim_{\nu\to\infty}\sum_{l=\bar{\nu}}^{\nu}\left(||\mathbf{X}^{l+1}-\mathbf{X}^{l}||_F-\frac{Z^l}{2s_X}\right)^2<\infty, \tag{73}$$

$$\lim_{\nu\to\infty}\sum_{l=\bar{\nu}}^{\nu}\gamma^l\left(||\widetilde{\mathbf{D}}^{l}-\mathbf{D}^{l}||_F-\frac{I\bar{\epsilon}_{\phi}}{2\bar{\tau}_D}T^l\right)^2<\infty. \tag{74}$$

Finally, we deduce that $\{U(\overline{\mathbf{D}}^{\nu},\mathbf{X}^{\nu})\}_{\nu}$ converges to the same limit point of $\{U(\overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu}},\mathbf{X}^{\nu})\}_{\nu}$, due to i) $\sqrt{I}\left\|\overline{\mathbf{D}}^{\nu}-\overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu}}\right\|_F\leq\left\|\mathbf{D}^{\nu}-\mathbf{1}\otimes\overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu}}\right\|_F\xrightarrow[\nu\to\infty]{(51)}0$; ii) the continuity of U; and iii) the boundedness of $\left\{\left(\mathbf{D}^{\nu},\mathbf{X}^{\nu}\right)\right\}_{\nu}$ [cf. Sec. A.2.2]. This concludes the proof. □

A.2.4. Step 4–Vanishing X-stationarity

Building on the results in the previous step, we prove here $\lim_{\nu\to\infty}\Delta_X(\overline{\mathbf{D}}^{\nu},\mathbf{X}^{\nu})=0$ [statement (b-iii)]. For notational simplicity, we will use the shorthand $\widehat{\mathbf{X}}_i^{\nu}\triangleq\widehat{\mathbf{X}}_i(\overline{\mathbf{D}}^{\nu},\mathbf{X}^{\nu})$ and $\widehat{\mathbf{X}}^{\nu}\triangleq\widehat{\mathbf{X}}(\overline{\mathbf{D}}^{\nu},\mathbf{X}^{\nu})$, with $\widehat{\mathbf{X}}_i(\overline{\mathbf{D}}^{\nu},\mathbf{X}^{\nu})$ and $\widehat{\mathbf{X}}(\overline{\mathbf{D}}^{\nu},\mathbf{X}^{\nu})$ defined in (29).

Using the equivalence of norms and the triangle inequality, we have

$$\Delta_X(\overline{\mathbf{D}}^{\nu},\mathbf{X}^{\nu})\leq K_X\|\widehat{\mathbf{X}}^{\nu}-\mathbf{X}^{\nu}\|_F\leq K_X\underbrace{\|\mathbf{X}^{\nu+1}-\mathbf{X}^{\nu}\|_F}_{\texttt{term I}}+K_X\underbrace{\|\mathbf{X}^{\nu+1}-\widehat{\mathbf{X}}^{\nu}\|_F}_{\texttt{term II}},$$

for some $K_X>0$. The rest of the proof consists in showing that `term I` and `term II` above are asymptotically vanishing.

- **On `term I`:** `Term I` can be bounded as

$$\frac{1}{2}||\mathbf{X}^{\nu+1}-\mathbf{X}^{\nu}||_F^2\leq\left(||\mathbf{X}^{\nu+1}-\mathbf{X}^{\nu}||_F-\frac{Z^{\nu}}{2s_X}\right)^2+\left(\frac{Z^{\nu}}{2s_X}\right)^2, \tag{75}$$

for all $\nu\in\mathbb{N}_+$, where the inequality follows from $\frac{1}{2}a^2\leq(a-b)^2+b^2$, with $a,b\in\mathbb{R}$. This, together with (70) and (73), yields

$$\sum_{\nu=0}^{\infty}||\mathbf{X}^{\nu+1}-\mathbf{X}^{\nu}||_F^2<\infty\quad\Longrightarrow\quad\lim_{\nu\to\infty}||\mathbf{X}^{\nu+1}-\mathbf{X}^{\nu}||_F=0. \tag{76}$$

• **On `term II`:** Invoking the optimality of $\widehat{\mathbf{X}}_i^\nu$ [cf. (29)] and $\mathbf{X}_i^{\nu+1}$ [cf. (11)], yields

$$\begin{aligned}\left\langle \nabla_{X_i} f_i(\overline{\mathbf{D}}^\nu, \mathbf{X}_i^\nu) + \hat{\tau}_X(\widehat{\mathbf{X}}_i^\nu - \mathbf{X}_i^\nu), \mathbf{X}_i^{\nu+1} - \widehat{\mathbf{X}}_i^\nu \right\rangle + g_i(\mathbf{X}_i^{\nu+1}) - g_i(\widehat{\mathbf{X}}_i^\nu) \geq 0,\\ \left\langle \nabla_{X_i} \tilde{h}_i(\mathbf{X}_i^{\nu+1}; \mathbf{U}_{(i)}^\nu, \mathbf{X}_i^\nu), \widehat{\mathbf{X}}_i^\nu - \mathbf{X}_i^{\nu+1} \right\rangle + g_i(\widehat{\mathbf{X}}_i^\nu) - g_i(\mathbf{X}_i^{\nu+1}) \geq 0.\end{aligned}$$

Summing the two inequalities above and using Remark 7(b), lead to

$$\begin{aligned}\hat{\tau}_X ||\mathbf{X}_i^{\nu+1} - \widehat{\mathbf{X}}_i^\nu||_F^2 \leq & \left\langle \hat{\tau}_X \left(\mathbf{X}_i^{\nu+1} - \mathbf{X}_i^\nu\right) + \nabla_{X_i} f_i(\overline{\mathbf{D}}^\nu, \mathbf{X}_i^\nu) - \nabla_{X_i} f_i(\mathbf{U}_{(i)}^\nu, \mathbf{X}_i^\nu), \mathbf{X}_i^{\nu+1} - \widehat{\mathbf{X}}_i^\nu \right\rangle \\ & + \left\langle \nabla_{X_i} \tilde{h}_i(\mathbf{X}_i^\nu; \mathbf{U}_{(i)}^\nu, \mathbf{X}_i^\nu) - \nabla_{X_i} \tilde{h}_i(\mathbf{X}_i^{\nu+1}; \mathbf{U}_{(i)}^\nu, \mathbf{X}_i^\nu), \mathbf{X}_i^{\nu+1} - \widehat{\mathbf{X}}_i^\nu \right\rangle.\end{aligned} \tag{77}$$

Using the $\tilde{L}_{\nabla,i}^X$-Lipschitz continuity of $\nabla_{X_i} \tilde{h}_i$ [cf. Remark 9] and the $L_{\nabla,i}$-Lipschitz continuity of ∇f_i, and (10), it is not difficult to show that (77) implies

$$\begin{aligned}& ||\mathbf{X}_i^{\nu+1} - \widehat{\mathbf{X}}_i^\nu||_F \\ & \leq \frac{\tilde{L}_{\nabla,i}^X + \hat{\tau}_X}{\hat{\tau}_X} \left\|\mathbf{X}_i^{\nu+1} - \mathbf{X}_i^\nu\right\|_F + \frac{L_{\nabla,i}}{\hat{\tau}_X} \left\|\overline{\mathbf{D}}^\nu - \mathbf{D}_{(i)}^\nu\right\|_F + \frac{L_{\nabla,i}}{\hat{\tau}_X} \gamma^\nu \left\|\widetilde{\mathbf{D}}_{(i)}^\nu - \mathbf{D}_{(i)}^\nu\right\|_F, \\ & \leq \frac{\tilde{L}_{\nabla,i}^X + \hat{\tau}_X}{\hat{\tau}_X} \left\|\mathbf{X}_i^{\nu+1} - \mathbf{X}_i^\nu\right\|_F + \frac{L_{\nabla,i}}{\hat{\tau}_X} \left\|\overline{\mathbf{D}}^\nu - \mathbf{D}_{(i)}^\nu\right\|_F + \frac{L_{\nabla,i}}{\hat{\tau}_X} \gamma^\nu B_D,\end{aligned} \tag{78}$$

for some $B_D > 0$, where in the last inequality we used the boundedness of $\{||\widetilde{\mathbf{D}}^\nu - \mathbf{D}^\nu||_F\}_\nu$. Eq. (78) together with (76), Theorem 2(a), and $\gamma^\nu \overset{\nu\to\infty}{\longrightarrow} 0$ (cf. Assumption E), yield

$$\lim_{\nu\to\infty} \|\mathbf{X}_i^{\nu+1} - \widehat{\mathbf{X}}_i^\nu\|_F = 0. \tag{79}$$

This concludes the proof of statement (b-iiii). □

A.2.5. Step 5–Vanishing liminf D-stationarity

We prove $\liminf_{\nu\to\infty} \Delta_D(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu) = 0$ [statement (b-iv)] and $\{(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu)\}_\nu$ has at least one limit point which is a stationary solution of P. For notational simplicity, we will use the shorthand $\widehat{\mathbf{D}}^\nu \triangleq \widehat{\mathbf{D}}(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu)$, with $\widehat{\mathbf{D}}(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu)$ defined in (28).

We begin bounding $\Delta_D(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu)$ as

$$\Delta_D(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu) \leq K_D \|\widehat{\mathbf{D}}^\nu - \overline{\mathbf{D}}^\nu\|_F \leq K_D \underbrace{\|\widetilde{\mathbf{D}}_{(i)}^\nu - \mathbf{D}_{(i)}^\nu\|_F}_{\texttt{term I}} + K_D \underbrace{\|\widetilde{\mathbf{D}}_{(i)}^\nu - \widehat{\mathbf{D}}^\nu\|_F}_{\texttt{term II}} + K_D \|\mathbf{D}_{(i)}^\nu - \overline{\mathbf{D}}^\nu\|_F,$$

for some $K_D > 0$. Note that $\|\mathbf{D}_{(i)}^\nu - \overline{\mathbf{D}}^\nu\|_F \to 0$ [Theorem 2(a)]. We show next that $\liminf_{\nu\to\infty}$ $\|\widetilde{\mathbf{D}}_{(i)}^\nu - \mathbf{D}_{(i)}^\nu\|_F = 0$ and $\|\widetilde{\mathbf{D}}_{(i)}^\nu - \widehat{\mathbf{D}}^\nu\|_F \to 0$, which proves $\liminf_{\nu\to\infty} \Delta_D(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu) = 0$.

• **On `term I`:** Similarly to the derivations of (75), there holds

$$\frac{\gamma^\nu}{2} ||\widetilde{\mathbf{D}}^\nu - \mathbf{D}^\nu||_F^2 \leq \gamma^\nu \left(||\widetilde{\mathbf{D}}^\nu - \mathbf{D}^\nu||_F - \frac{I\bar{\epsilon}_\phi}{2\,\bar{\tau}_D} T^\nu\right)^2 + \left(\frac{I\bar{\epsilon}_\phi}{2\,\bar{\tau}_D}\right)^2 \gamma^\nu (T^\nu)^2, \tag{80}$$

for all $\nu \in \mathbb{N}_+$. By eq. (74) and $\sum_{\nu=0}^{\infty} (T^\nu)^2 < \infty$ [cf. (69)], we have

$$\sum_{\nu=0}^{\infty} \gamma^\nu ||\widetilde{\mathbf{D}}^\nu - \mathbf{D}^\nu||_F^2 < \infty \quad \overset{\text{Assumption E}}{\Longrightarrow} \quad \liminf_{\nu\to\infty} ||\widetilde{\mathbf{D}}^\nu - \mathbf{D}^\nu||_F = 0. \tag{81}$$

• **On `term II`:** Using the optimality of $\widehat{\mathbf{D}}^\nu$ [cf. (28)] and $\widetilde{\mathbf{D}}^\nu_{(i)}$ [cf. (8)], yields

$$\left\langle \nabla_D F(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu) + \hat{\tau}_D(\widehat{\mathbf{D}}^\nu - \overline{\mathbf{D}}^\nu), \widetilde{\mathbf{D}}^\nu_{(i)} - \widehat{\mathbf{D}}^\nu \right\rangle + G(\widetilde{\mathbf{D}}^\nu_{(i)}) - G(\widehat{\mathbf{D}}^\nu) \geq 0,$$
$$\left\langle \nabla_D \tilde{f}_i(\widetilde{\mathbf{D}}^\nu_{(i)}; \mathbf{D}^\nu_{(i)}, \mathbf{X}^\nu_i) + I \cdot \mathbf{\Theta}^\nu_i - \nabla_D f_i(\mathbf{D}^\nu_{(i)}, \mathbf{X}^\nu_i), \widehat{\mathbf{D}}^\nu - \widetilde{\mathbf{D}}^\nu_{(i)} \right\rangle + G(\widehat{\mathbf{D}}^\nu) - G(\widetilde{\mathbf{D}}^\nu_{(i)}) \geq 0.$$

Summing the two inequalities above and using Remark 7(a), yields

$$\begin{aligned}
&\hat{\tau}_D ||\widetilde{\mathbf{D}}^\nu_{(i)} - \widehat{\mathbf{D}}^\nu||_F^2 \\
&\leq \left\langle \hat{\tau}_D(\widetilde{\mathbf{D}}^\nu_{(i)} - \mathbf{D}^\nu_{(i)}) + \hat{\tau}_D(\mathbf{D}^\nu_{(i)} - \overline{\mathbf{D}}^\nu) + \nabla_D F(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu) - \nabla_D F(\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}, \mathbf{X}^\nu), \widetilde{\mathbf{D}}^\nu_{(i)} - \widehat{\mathbf{D}}^\nu \right\rangle \\
&\quad + \left\langle \nabla_D F(\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}, \mathbf{X}^\nu) - I \cdot \mathbf{\Theta}^\nu_i, \widetilde{\mathbf{D}}^\nu_{(i)} - \widehat{\mathbf{D}}^\nu \right\rangle \\
&\quad - \left\langle \nabla_D \tilde{f}_i(\widetilde{\mathbf{D}}^\nu_{(i)}; \mathbf{D}^\nu_{(i)}, \mathbf{X}^\nu_i) - \nabla_D \tilde{f}_i(\mathbf{D}^\nu_{(i)}; \mathbf{D}^\nu_{(i)}, \mathbf{X}^\nu_i), \widetilde{\mathbf{D}}^\nu_{(i)} - \widehat{\mathbf{D}}^\nu \right\rangle.
\end{aligned} \tag{82}$$

Using the $\tilde{L}^D_{\nabla,i}$-Lipschitz continuity of $\nabla_D \tilde{f}_i$ [cf. Remark 9] and the L_{∇_D}-Lipschitz continuity of $\nabla_D F$ [cf. Remark 8], it is not difficult to check that (82) implies

$$\begin{aligned}
\left\| \widetilde{\mathbf{D}}^\nu_{(i)} - \widehat{\mathbf{D}}^\nu \right\|_F \leq & \frac{\tilde{L}^D_{\nabla,i} + \hat{\tau}_D}{\hat{\tau}_D} \left\| \widetilde{\mathbf{D}}^\nu_{(i)} - \mathbf{D}^\nu_{(i)} \right\|_F + \left\| \overline{\mathbf{D}}^\nu - \mathbf{D}^\nu_{(i)} \right\|_F + \frac{L_{\nabla_D}}{\hat{\tau}_D \sqrt{I}} \left\| \mathbf{D}^\nu - \mathbf{1} \otimes \overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu} \right\| \\
& + \frac{I}{\hat{\tau}_D} \left\| \mathbf{\Theta}^\nu_i - \frac{1}{I} \nabla_D F(\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}, \mathbf{X}^\nu) \right\|_F,
\end{aligned} \tag{83}$$

for all $i = 1, \ldots, I$. Since $\liminf_{\nu\to\infty} ||\widetilde{\mathbf{D}}^\nu_{(i)} - \mathbf{D}^\nu_{(i)}||_F = 0$ [cf. (81)], $||\overline{\mathbf{D}}^\nu - \mathbf{D}^\nu_{(i)}||_F \to 0$ [Theorem 2(a)], and $\left\| \mathbf{D}^\nu - \mathbf{1} \otimes \overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu} \right\| \to 0$ [cf. (51)], to prove $\liminf_{\nu\to\infty} ||\widetilde{\mathbf{D}}^\nu_{(i)} - \widehat{\mathbf{D}}^\nu||_F = 0$, it is sufficient to show that the last term on the RHS of the above inequality is asymptotically vanishing, which is done in the lemma below.

Lemma 11 (Vanishing gradient-tracking error) *In the setting above, there holds:*

$$\sum_{\nu=0}^{\infty} \left\| \mathbf{\Theta}^\nu - \mathbf{1} \otimes \frac{1}{I} \nabla_D F(\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}, \mathbf{X}^\nu) \right\|_F^2 < \infty. \tag{84}$$

Proof See Sec. A.4.4. ■

By $\liminf_{\nu\to\infty} \Delta_D(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu) = 0$, it follows that there exists an infinte subset $\mathcal{N} \subseteq \mathbb{N}_+$ such that $\lim_{\mathcal{N}\ni\nu\to\infty} \Delta_D(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu) = 0$. Since $\{(\mathbf{D}^\nu, \mathbf{X}^\nu)\}_\nu$ is bounded [cf. Sec. A.2.2], it has a convergent subsequence $\{(\mathbf{D}^\nu, \mathbf{X}^\nu)\}_{\nu\in\mathbb{N}'}$, with $\mathbb{N}' \subseteq \mathbb{N}$; let $(\overline{\mathbf{D}}^\infty, \mathbf{X}^\infty)$ denote its limit point. Then, it must be $\lim_{\mathcal{N}'\ni\nu\to\infty} \Delta_D(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu) = 0$. Combining this result with $\lim_{\nu\to\infty} \Delta_X(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu) = 0$ (cf. Sec. A.2.4), one can conclude $\lim_{\mathcal{N}'\ni\nu\to\infty} \Delta^\nu = 0$; hence, $(\overline{\mathbf{D}}^\infty, \mathbf{X}^\infty)$ is a stationary solution of Problem P.

A.2.6. Step 6–Vanishing D-stationarity

Finally, we prove $\lim_{\nu\to\infty} \Delta_D(\overline{\mathbf{D}}^{\nu}, \mathbf{X}^{\nu}) = 0$ [statement (b')]. In view of the results already proved in Step 5, it is sufficient to show that $\limsup_{\nu\to\infty} ||\widetilde{\mathbf{D}}^{\nu} - \mathbf{D}^{\nu}||_F = 0$.
1) Preliminaries: We begin introducing the following preliminary results.

Proposition 12 *In the setting of Theorem 2(a), the following hold for $\widetilde{\mathbf{D}}^{\nu}$ [cf. (8)] and $\mathbf{X}^{\nu}$ [cf. (11)]:*

(a) There exists some constant $L_D > 0$ and sequence $\{\tilde{T}^{\nu}\}_{\nu}$, with $\lim_{\nu\to\infty} \tilde{T}^{\nu} = 0$, such that, for any $\nu_1, \nu_2 \in \mathbb{N}_+$,

$$||\widetilde{\mathbf{D}}^{\nu_2} - \widetilde{\mathbf{D}}^{\nu_1}||_F \leq L_D\Big(||\overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu_2}} - \overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu_1}}||_F + ||\mathbf{X}^{\nu_2} - \mathbf{X}^{\nu_1}||_F\Big) + \tilde{T}^{\nu_1} + \tilde{T}^{\nu_2}; \tag{85}$$

(b) There exist some constants $0 < p_X < 1$ and $q_X > 0$, and a sufficiently large $\nu_X \in \mathbb{N}_+$ such that, for all $\nu \geq \nu_X$,

$$||\mathbf{X}^{\nu+1} - \mathbf{X}^{\nu}||_F \leq p_X||\mathbf{X}^{\nu} - \mathbf{X}^{\nu-1}||_F + q_X||\mathbf{U}^{\nu} - \mathbf{U}^{\nu-1}||_F. \tag{86}$$

Proof See Appendix A.4.5 ∎

2) Proof of $\limsup_{\nu\to\infty} ||\widetilde{\mathbf{D}}^{\nu} - \mathbf{D}^{\nu}||_F = 0$**.** For notational simplicity, let us define $\Delta\widetilde{\mathbf{D}}^{\nu} \triangleq \widetilde{\mathbf{D}}^{\nu} - \mathbf{D}^{\nu}$. Suppose by contradiction that $\limsup_{\nu\to\infty} ||\Delta\widetilde{\mathbf{D}}^{\nu}||_F > 0$; since $\liminf_{\nu\to\infty} ||\Delta\widetilde{\mathbf{D}}^{\nu}||_F = 0$ [cf. (81)], there exists $\delta > 0$ such that $||\Delta\widetilde{\mathbf{D}}^{\nu}||_F > 2\delta$ and $||\Delta\widetilde{\mathbf{D}}^{\nu'}||_F < \delta$ for infinitely many $\nu, \nu' \in \mathbb{N}_+$. Therefore, one can find an infinite subset of indices, denoted by $\mathcal{K}$, having the following properties: for any $\nu \in \mathcal{K}$, there exists an index $i_\nu > \nu$ such that

$$||\Delta\widetilde{\mathbf{D}}^{\nu}||_F < \delta, \qquad ||\Delta\widetilde{\mathbf{D}}^{i_\nu}||_F > 2\delta, \tag{87}$$

$$\delta \leq ||\Delta\widetilde{\mathbf{D}}^{j}||_F \leq 2\delta, \qquad \nu < j < i_\nu. \tag{88}$$

Let ν_2 be a sufficiently large integer such that (64) holds and $T^{\nu} < \frac{2\bar{\tau}_D\,\delta}{I\bar{\epsilon}_\phi}$, for all $\nu \geq \nu_2$ [such ν_2 exists, due to (69)]. Note that there exists a $\bar{\delta} > 0$ such that $\delta - \frac{I\bar{\epsilon}_\phi}{2\bar{\tau}_D}T^{\nu} \geq \bar{\delta}$, for all $\nu \geq \nu_2$. Choose $\mathcal{K} \ni \nu \geq \nu_2$; using (64), with $\nu = i_\nu$ and $\bar{\nu} = \nu + 1$, yields

$$U(\overline{\mathbf{D}}_{\boldsymbol{\phi}^{i_\nu+1}}, \mathbf{X}^{i_\nu+1}) \leq U(\overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu+1}}, \mathbf{X}^{\nu+1}) - c_8\,\left(\bar{\delta}\right)^2 \sum_{l=\nu+1}^{i_\nu} \gamma^l + \sum_{l=\nu+1}^{i_\nu} W^l + E^{i_\nu,\nu+1}, \tag{89}$$

for some finite constant $c_8 > 0$. Using the convergence of $\{U(\overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu}}, \mathbf{X}^{\nu})\}_{\nu}$, $\sum_{l=1}^{\infty} W^l < \infty$ [cf. (71)], and $\lim_{\mathcal{K}\ni\nu\to\infty} E^{i_\nu,\nu+1} = 0$ [cf. (72)], inequality (89) implies

$$\lim_{\mathcal{K}\ni\nu\to\infty} \sum_{l=\nu+1}^{i_\nu} \gamma^l = 0. \tag{90}$$

We show next that (90) leads to a contradiction.

It follows from (87) and (88) that, for all $\mathcal{K} \ni \nu \geq \nu_2$,

$$\begin{aligned}
\delta <& ||\Delta\widetilde{\mathbf{D}}^{i_\nu}||_F - ||\Delta\widetilde{\mathbf{D}}^{\nu}||_F \\
\overset{(a)}{\leq}& ||\Delta\widetilde{\mathbf{D}}^{i_\nu} - \Delta\widetilde{\mathbf{D}}^{\nu}||_F = ||\widetilde{\mathbf{D}}^{i_\nu} - \mathbf{D}^{i_\nu} - \widetilde{\mathbf{D}}^{\nu} + \mathbf{D}^{\nu}||_F \\
\overset{(b)}{\leq}& \|\widetilde{\mathbf{D}}^{i_\nu} - \widetilde{\mathbf{D}}^{\nu}\| + \|\mathbf{D}^{\nu} \pm \mathbf{1} \otimes \overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu}} \pm \mathbf{1} \otimes \overline{\mathbf{D}}_{\boldsymbol{\phi}^{i_\nu}} - \mathbf{D}^{i_\nu}\|_F \\
\overset{(85)}{\leq}& \left(L_D + \sqrt{I}\right) ||\overline{\mathbf{D}}_{\boldsymbol{\phi}^{i_\nu}} - \overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu}}||_F + \widetilde{E}^{i_\nu,\nu},
\end{aligned} \tag{91}$$

with

$$\widetilde{E}^{i_\nu,\nu} \triangleq L_D||\mathbf{X}^{i_\nu} - \mathbf{X}^{\nu}||_F + ||\mathbf{D}^{\nu} - \mathbf{1} \otimes \overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu}}||_F + ||\mathbf{D}^{i_\nu} - \mathbf{1} \otimes \overline{\mathbf{D}}_{\boldsymbol{\phi}^{i_\nu}}||_F + \tilde{T}^{\nu} + \tilde{T}^{i_\nu},$$

where in (a) we used the reverse triangle inequality (i.e. $||\mathbf{A}||_F - ||\mathbf{B}||_F \leq ||\mathbf{A}-\mathbf{B}||_F, \forall \mathbf{A}, \mathbf{B} \in \mathbb{R}^{MI\times K}$); and in (b) we add/subtracted some dummy terms and used the triangle inequality.

We prove next that $\lim_{\nu\to\infty} \widetilde{E}^{i_\nu,\nu} = 0$. Clearly, if $\lim_{\nu\to\infty} ||\mathbf{X}^{i_\nu} - \mathbf{X}^{\nu}||_F = 0$, then (51) [cf. Proposition 5] and $\tilde{T}^{\nu} \overset{\nu\to\infty}{\longrightarrow} 0$ [cf. Proposition 12(a)] imply $\lim_{\nu\to\infty} \widetilde{E}^{i_\nu,\nu} = 0$. It is then sufficient to show $\lim_{\nu\to\infty} ||\mathbf{X}^{i_\nu} - \mathbf{X}^{\nu}||_F = 0$.

First, we bound $||\mathbf{X}^{i_\nu} - \mathbf{X}^{\nu}||_F$ properly.

Summing (86) from $\nu > \nu_2$ to $i_\nu - 1$, yields

$$\sum_{t=\nu}^{i_\nu-1} ||\mathbf{X}^{t+1} - \mathbf{X}^{t}||_F \leq p_X \sum_{t=\nu}^{i_\nu-1} ||\mathbf{X}^{t} - \mathbf{X}^{t-1}||_F + q_X \sum_{t=\nu}^{i_\nu-1} ||\mathbf{U}^{t} - \mathbf{U}^{t-1}||_F, \tag{92}$$

implying

$$||\mathbf{X}^{i_\nu} - \mathbf{X}^{\nu}||_F \leq \sum_{t=\nu}^{i_\nu-1} ||\mathbf{X}^{t+1} - \mathbf{X}^{t}||_F \leq \frac{p_X}{1-p_X}||\mathbf{X}^{\nu} - \mathbf{X}^{\nu-1}||_F + \frac{q_X}{1-p_X} \sum_{t=\nu}^{i_\nu-1} ||\mathbf{U}^{t} - \mathbf{U}^{t-1}||_F.$$

Since $\lim_{\nu\to\infty} ||\mathbf{X}^{\nu} - \mathbf{X}^{\nu-1}||_F = 0$ [cf. (76)], it follows from the above inequality that $\lim_{\mathcal{K}\ni\nu\to\infty} ||\mathbf{X}^{i_\nu} - \mathbf{X}^{\nu}||_F = 0$, if $\lim_{\mathcal{K}\ni\nu\to\infty} \sum_{t=\nu}^{i_\nu-1} ||\mathbf{U}^{t} - \mathbf{U}^{t-1}||_F = 0$, which is proved next. Rewrite first $||\mathbf{U}^{t} - \mathbf{U}^{t-1}||_F$ as

$$\begin{aligned}
||\mathbf{U}^{t} - \mathbf{U}^{t-1}||_F \leq& ||\mathbf{U}^{t} - \mathbf{1} \otimes \overline{\mathbf{U}}_{\boldsymbol{\phi}^{t}}||_F + ||\mathbf{U}^{t-1} - \mathbf{1} \otimes \overline{\mathbf{U}}_{\boldsymbol{\phi}^{t-1}}||_F + \sqrt{I}||\overline{\mathbf{U}}_{\boldsymbol{\phi}^{t}} - \overline{\mathbf{U}}_{\boldsymbol{\phi}^{t-1}}||_F \\
\overset{(42)-(43)}{=}& ||\mathbf{U}^{t} - \mathbf{1} \otimes \overline{\mathbf{U}}_{\boldsymbol{\phi}^{t}}||_F + ||\mathbf{U}^{t-1} - \mathbf{1} \otimes \overline{\mathbf{U}}_{\boldsymbol{\phi}^{t-1}}||_F + \frac{\gamma^t}{\sqrt{I}} \left\| \sum_{i=1}^{I} \phi_i^t \left(\widetilde{\mathbf{D}}_{(i)}^t - \mathbf{D}_{(i)}^t\right) \right\|_F.
\end{aligned}$$

Since $|| \sum_{i=1}^{I} \phi_i^t (\widetilde{\mathbf{D}}_{(i)}^t - \mathbf{D}_{(i)}^t)||_F$ is bounded (due to $\phi_i^t \leq \bar{\epsilon}_\phi$ [cf. (49)] and compactness of $\mathcal{D}$) and $\lim_{\mathcal{K}\ni\nu\to\infty} \sum_{t=\nu}^{i_\nu} \gamma^t = 0$ [cf. (90)], there holds $\lim_{\mathcal{K}\ni\nu\to\infty} \sum_{t=\nu}^{i_\nu} \gamma^t || \sum_{i=1}^{I} \widetilde{\mathbf{D}}_{(i)}^t - \mathbf{D}_{(i)}^t||_F = 0$. Therefore, to prove $\lim_{\mathcal{K}\ni\nu\to\infty} \sum_{t=\nu}^{i_\nu-1} ||\mathbf{U}^{t} - \mathbf{U}^{t-1}||_F = 0$, it is sufficient to show that $\lim_{\mathcal{K}\ni\nu\to\infty} \sum_{t=\nu}^{i_\nu} ||\mathbf{U}^{t} - \mathbf{1} \otimes \overline{\mathbf{U}}_{\boldsymbol{\phi}^{t}}||_F = 0$ [which implies also $\lim_{\mathcal{K}\ni\nu\to\infty} \sum_{t=\nu}^{i_\nu} ||\mathbf{U}^{t-1} - \mathbf{1} \otimes \overline{\mathbf{U}}_{\boldsymbol{\phi}^{t-1}}||_F = 0$, due to $\lim_{\nu\to\infty} ||\mathbf{U}^{\nu} - \mathbf{1} \otimes \overline{\mathbf{U}}_{\boldsymbol{\phi}^{\nu}}||_F = 0$, see (53)]. By (57), the boundedness of $\{\widetilde{\mathbf{D}}^\nu\}_\nu$, and (90), it is sufficient to show that $\lim_{\mathcal{K}\ni\nu\to\infty}\sum_{t=\nu}^{i_\nu}||\mathbf{D}^t - \mathbf{1}\otimes\overline{\mathbf{D}}_{\boldsymbol{\phi}^t}||_F = 0$. We have

$$\begin{aligned}
&\lim_{\mathcal{K}\ni\nu\to\infty}\sum_{l=\nu}^{i_\nu}||\mathbf{D}^l - \mathbf{1}\otimes\overline{\mathbf{D}}_{\boldsymbol{\phi}^l}||_F \overset{(56)}{\le} \lim_{\mathcal{K}\ni\nu\to\infty}\sum_{l=\nu}^{i_\nu}\left(c_1(\rho)^l + c_2\sum_{t=0}^{l-1}\gamma^t\,(\rho)^{l-t}\right)\\
&= c_2\lim_{\mathcal{K}\ni\nu\to\infty}\sum_{l=\nu}^{i_\nu}\sum_{t=0}^{l-1}\gamma^t\,(\rho)^{l-t} = c_2\lim_{\mathcal{K}\ni\nu\to\infty}\sum_{t=0}^{i_\nu-1}\sum_{l=\max(t+1,\nu)}^{i_\nu}\gamma^t\,(\rho)^{l-t}\\
&= c_2\lim_{\mathcal{K}\ni\nu\to\infty}\sum_{t=0}^{\nu-1}\sum_{l=\nu}^{i_\nu}\gamma^t\,(\rho)^{l-t} + c_2\lim_{\mathcal{K}\ni\nu\to\infty}\sum_{t=\nu}^{i_\nu-1}\sum_{l=t+1}^{i_\nu}\gamma^t\,(\rho)^{l-t}\\
&\le c_2\lim_{\mathcal{K}\ni\nu\to\infty}\sum_{t=0}^{\nu-1}\sum_{l=\nu}^{i_\nu}\gamma^t\,(\rho)^{l-t} + \frac{c_2}{1-\rho}\underbrace{\lim_{\mathcal{K}\ni\nu\to\infty}\sum_{t=\nu}^{i_\nu-1}\gamma^t}_{=0\quad\text{by (90)}}\\
&\le \frac{c_2}{1-\rho}\lim_{\mathcal{K}\ni\nu\to\infty}\sum_{t=0}^{\nu-1}\gamma^t\,(\rho)^{\nu-t} \overset{(104)}{=} 0.
\end{aligned}$$

This proves $\lim_{\nu\to\infty}||\mathbf{X}^{i_\nu} - \mathbf{X}^\nu||_F = 0$ and thus $\lim_{\nu\to\infty}\widetilde{E}^{i_\nu,\nu} = 0$.

We can now prove that (90) leads to a contradiction. Since $\widetilde{E}^{i_\nu,\nu}\overset{\nu\to\infty}{\longrightarrow}0$, there exists a sufficiently large integer $\nu_3\in\mathcal{K}$, such that $\nu_3 > \nu_2$ and $\widetilde{E}^{i_\nu,\nu} < \delta$, for all $\nu > \nu_3$. Define δ' such that $0 < \delta' \le \delta - \widetilde{E}^{i_\nu,\nu}$. Using (43) and $\mathbf{1}^\intercal\boldsymbol{\phi}^\nu = I$, (91) implies

$$\frac{\delta'}{(L_D+1)\sqrt{I}} < \sum_{t=\nu}^{i_\nu-1}\gamma^t\,||\Delta\widetilde{\mathbf{D}}^t||_F \overset{(87)-(88)}{\le} 2\delta\sum_{t=\nu}^{i_\nu-1}\gamma^t, \tag{93}$$

for all $\mathcal{K}\ni\nu>\nu_3$. Equation (93) contradicts (90). Hence, it must be $\limsup_{\nu\to\infty}||\Delta\widetilde{\mathbf{D}}^\nu||_F = 0$, and thus $\lim_{\nu\to\infty}||\widetilde{\mathbf{D}}^\nu - \mathbf{D}^\nu||_F = 0$. □

A.3. Proof of Theorem 3

(a) Rate of consensus error. Fix $\theta\in(0,1)$. Combining (50) and (56), we obtain

$$\begin{aligned}
e^\nu &\le c_8\sum_{l=1}^{\nu}\gamma^{\nu-l}\,(\rho)^l = c_8\sum_{l=1}^{\lfloor(1-\theta)\nu\rfloor}\gamma^{\nu-l}\,(\rho)^l + c_8\sum_{l=\lfloor(1-\theta)\nu\rfloor+1}^{\nu}\gamma^{\nu-l}\,(\rho)^l\\
&\le c_8\,\gamma^{\nu-\lfloor(1-\theta)\nu\rfloor}\sum_{l=0}^{\lfloor(1-\theta)\nu\rfloor}(\rho)^l + c_8\,\gamma^0\sum_{l=\lfloor(1-\theta)\nu\rfloor+1}^{\nu}(\rho)^l,\\
&\overset{(a)}{=} c_8\,\gamma^{\lceil\theta\nu\rceil}\,\frac{1-(\rho)^{\lfloor(1-\theta)\nu\rfloor+1}}{1-\rho} + c_8\,\gamma^0\,(\rho)^{\lfloor(1-\theta)\nu\rfloor+1}\cdot\frac{1-(\rho)^{\lceil\theta\nu\rceil}}{1-\rho}\\
&\le c_9\left(\gamma^{\lceil\theta\nu\rceil} + (\rho)^{(1-\theta)\nu}\right) \overset{(b)}{\le} c_9\left(\gamma^{\lceil\theta\nu\rceil} + \left((\rho)^{\frac{1-\theta}{\theta}}\right)^{\lceil\theta\nu\rceil-1}\right) \overset{(c)}{=} \mathcal{O}\left(\gamma^{\lceil\theta\nu\rceil}\right),
\end{aligned}\tag{94}$$

for some positive constants c_8 and c_9, where in (a) we used $\nu - \lfloor(1-\theta)\nu\rfloor = \lceil\theta\nu\rceil$; (b) follows from $x \ge \lceil x\rceil - 1$, $x\in\mathbb{R}$; and in (c) we used $(\tilde{\rho})^\nu = o(\gamma^\nu)$, for any $\tilde{\rho}\in(0,1)$. This proves statement (a).

(b) Rate of optimization errors. In the following we will use the shorthand: $\widehat{\mathbf{X}}_i^\nu \triangleq \widehat{\mathbf{X}}_i(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu)$ and $\widehat{\mathbf{X}}^\nu \triangleq \widehat{\mathbf{X}}(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu)$, with $\widehat{\mathbf{X}}_i(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu)$ and $\widehat{\mathbf{X}}(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu)$ defined in (29); and $\widehat{\mathbf{D}}^\nu \triangleq \widehat{\mathbf{D}}(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu)$, with $\widehat{\mathbf{D}}(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu)$ defined in (28).

1) Proof of (32)**:** We begin bounding $\Delta_X(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu)$ as

$$\frac{1}{2}(\Delta_X(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu))^2 \overset{(a)}{\leq} \frac{K_1}{2}||\widehat{\mathbf{X}}^\nu - \mathbf{X}^\nu||_F^2 \leq K_1||\mathbf{X}^{\nu+1} - \widehat{\mathbf{X}}^\nu||_F^2 + K_1||\mathbf{X}^{\nu+1} - \mathbf{X}^\nu||_F^2, \tag{95}$$

where (a) holds by equivalence of the norms with K_1 being a proper positive constant.

By squaring both sides of (78) and using $\frac{1}{n}(\sum_{i=1}^n a_i)^2 \leq \sum_{i=1}^n a_i^2, \forall a_i \in \mathbb{R}$ (by Jensen inequality), the first term on the RHS of (95) can be bounded as

$$\frac{1}{3K_2}||\mathbf{X}_i^{\nu+1} - \widehat{\mathbf{X}}_i^\nu||_F^2 \leq \left\|\mathbf{X}_i^{\nu+1} - \mathbf{X}_i^\nu\right\|_F^2 + \left\|\overline{\mathbf{D}}^\nu - \mathbf{D}_{(i)}^\nu\right\|_F^2 + (\gamma^\nu)^2, \tag{96}$$

for some positive constant $K_2 > 0$. Summing (96) over $i = 1, \ldots, I$, yields

$$\begin{aligned}\frac{1}{3K_2}||\mathbf{X}^{\nu+1} - \widehat{\mathbf{X}}^\nu||_F^2 &\leq \left\|\mathbf{X}^{\nu+1} - \mathbf{X}^\nu\right\|_F^2 + \left\|\mathbf{D}^\nu - \mathbf{1}\otimes\overline{\mathbf{D}}^\nu\right\|_F^2 + I(\gamma^\nu)^2 \\ &\leq \left\|\mathbf{X}^{\nu+1} - \mathbf{X}^\nu\right\|_F^2 + 4\left\|\mathbf{D}^\nu - \mathbf{1}\otimes\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}\right\|_F^2 + I(\gamma^\nu)^2.\end{aligned} \tag{97}$$

Finally, combining (95) and (97), yields

$$\frac{1}{2K_1}(\Delta_X(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu))^2 \leq (3K_2+1)\left\|\mathbf{X}^{\nu+1} - \mathbf{X}^\nu\right\|_F^2 + 12K_1\left\|\mathbf{D}^\nu - \mathbf{1}\otimes\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}\right\|_F^2 + 3K_2 I(\gamma^\nu)^2. \tag{98}$$

It follows from (98) together with (76), (52) and Assumption E

$$\sum_{\nu=0}^{\infty}(\Delta_X(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu))^2 < \infty.$$

By the definition of $T_{X,\epsilon}$, there holds

$$T_{X,\epsilon}\epsilon^2 \leq \sum_{\nu=0}^{T_{X,\epsilon}}(\Delta_X(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu))^2 < \infty,$$

which proves (32).

2) Proof of (33)**:** Following the same approach as above, we can bound $\Delta_D(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu)$ as

$$\frac{1}{3}(\Delta_D(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu))^2 \leq \frac{K_3}{3}||\widehat{\mathbf{D}}^\nu - \overline{\mathbf{D}}^\nu||_F^2 \leq K_3||\widehat{\mathbf{D}}^\nu - \widetilde{\mathbf{D}}_{(i)}^\nu||_F^2 + K_3||\widetilde{\mathbf{D}}_{(i)}^\nu - \mathbf{D}_{(i)}^\nu||_F^2 + K_3||\mathbf{D}_{(i)}^\nu - \overline{\mathbf{D}}^\nu||_F^2, \tag{99}$$

for some $K_3 > 0$. Using (83), the first term on the RHS of (99) can be bounded as

$$\begin{aligned}\frac{1}{4K_4}\|\widetilde{\mathbf{D}}_{(i)}^\nu - \widehat{\mathbf{D}}^\nu\|_F^2 \leq \|\widetilde{\mathbf{D}}_{(i)}^\nu - \mathbf{D}_{(i)}^\nu\|_F^2 + \left\|\mathbf{\Theta}_{(i)}^\nu - \frac{1}{I}\nabla_D F(\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}, \mathbf{X}^\nu)\right\|_F^2 + \|\overline{\mathbf{D}}^\nu - \mathbf{D}_{(i)}^\nu\|_F^2 \\ + \left\|\mathbf{D}^\nu - \mathbf{1}\otimes\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}\right\|^2\end{aligned} \tag{100}$$

with some constant $K_4 > 0$. Summing (100) over $i = 1, \ldots, I$, we get

$$\begin{aligned}
&\frac{1}{4K_4}\|\widetilde{\mathbf{D}}^\nu - \mathbf{1}\otimes\widehat{\mathbf{D}}^\nu\|_F^2 \\
\leq &\|\widetilde{\mathbf{D}}^\nu - \mathbf{D}^\nu\|_F^2 + \left\|\boldsymbol{\Theta}^\nu - \mathbf{1}\otimes\frac{1}{I}\nabla_D F(\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}, \mathbf{X}^\nu)\right\|_F^2 + \left\|\mathbf{D}^\nu - \mathbf{1}\otimes\overline{\mathbf{D}}^\nu\right\|_F^2 + I\left\|\mathbf{D}^\nu - \mathbf{1}\otimes\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}\right\|_F^2 \\
\leq &\|\widetilde{\mathbf{D}}^\nu - \mathbf{D}^\nu\|_F^2 + \left\|\boldsymbol{\Theta}^\nu - \mathbf{1}\otimes\frac{1}{I}\nabla_D F(\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}, \mathbf{X}^\nu)\right\|_F^2 + (4+I)\left\|\mathbf{D}^\nu - \mathbf{1}\otimes\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}\right\|_F^2.
\end{aligned} \tag{101}$$

Summing (99) over $i = 1, \ldots, I$ and using (101), yields

$$\begin{aligned}
&\frac{I}{3K_3}(\Delta_D(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu))^2 \\
\leq &(4K_4+1)\|\widetilde{\mathbf{D}}^\nu - \mathbf{D}^\nu\|_F^2 + 4K_4\left\|\boldsymbol{\Theta}^\nu - \mathbf{1}\otimes\frac{1}{I}\nabla_D F(\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}, \mathbf{X}^\nu)\right\|_F^2 \\
&+ 4((4+I)K_4+1)\left\|\mathbf{D}^\nu - \mathbf{1}\otimes\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}\right\|_F^2.
\end{aligned} \tag{102}$$

It follows from (102) together with (81), (84), and (52)

$$\sum_{\nu=0}^{\infty}\gamma^\nu(\Delta_D(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu))^2 < \infty.$$

By definition of $T_{D,\epsilon}$ and non-increasing property of $\{\gamma^\nu\}_\nu$, we get

$$\gamma^{T_{D,\epsilon}}\,T_{D,\epsilon}\,\epsilon^2 \leq \sum_{\nu=0}^{T_{D,\epsilon}}\gamma^\nu(\Delta_D(\overline{\mathbf{D}}^\nu, \mathbf{X}^\nu))^2 < \infty. \tag{103}$$

Using $\gamma^\nu = K/\nu^p$, with some constant $K > 0$ and $p \in (1/2, 1)$, (103) provides the desired result as in (33). □

A.4. Miscellaneous results

This section contains some miscellaneous results used in the proofs of Theorems 2 and 3.

A.4.1. Sequence properties

The following lemma summarizes some summability properties of suitably chosen sequences, which appear in some of the proofs.

Lemma 13 *Given the sequences $\{a^\nu\}_\nu$ and $\{b^\nu\}_\nu$, and a scalar $\lambda \in [0, 1)$, the following hold:*

(a) If $\lim_{\nu\to\infty} a^\nu = 0$, then,

$$\lim_{\nu\to\infty}\sum_{t=1}^{\nu} a^t(\lambda)^{\nu-t} = 0. \tag{104}$$

(b) If $\lim_{\nu\to\infty}\sum_{t=1}^{\nu}\left(a^t\right)^2<\infty$ *and* $\lim_{\nu\to\infty}\sum_{t=1}^{\nu}\left(b^t\right)^2<\infty$*, then*

$$\lim_{\nu\to\infty}\sum_{l=1}^{\nu}\sum_{t=1}^{l}a^t b^l(\lambda)^{l-t}<\infty, \tag{105}$$

$$\lim_{\nu\to\infty}\sum_{l=1}^{\nu}\sum_{t=1}^{l}\left(a^t\right)^2(\lambda)^{l-t}<\infty, \tag{106}$$

$$\lim_{\nu,\nu'\to\infty}\sum_{l=1}^{\nu}\left(\sum_{t=l}^{\nu'}a^t(\lambda)^{t-l}\right)^2<\infty. \tag{107}$$

Proof For the proof of (a) and (105)-(106) in (b), see (Nedić et al., 2010, Lemma 7). We prove next (107). Expand the LHS of (107) as:

$$\begin{aligned}\lim_{\nu,\nu'\to\infty}\sum_{l=1}^{\nu}\left(\sum_{t=l}^{\nu'}a^t(\lambda)^{t-l}\right)^2 &= \lim_{\nu,\nu'\to\infty}\sum_{l=1}^{\nu}\sum_{t=l}^{\nu'}\sum_{k=l}^{\nu'}a^t a^k(\lambda)^{t-l}(\lambda)^{k-l}\\ &\le \lim_{\nu,\nu'\to\infty}\sum_{l=1}^{\nu}\sum_{t=l}^{\nu'}\sum_{k=l}^{\nu'}\frac{(a^t)^2+(a^k)^2}{2}(\lambda)^{t-l}(\lambda)^{k-l}=\lim_{\nu,\nu'\to\infty}\sum_{l=1}^{\nu}\sum_{t=l}^{\nu'}(a^t)^2(\lambda)^{t-l}\sum_{k=l}^{\nu'}(\lambda)^{k-l},\end{aligned}$$

where the inequality is due to $a\cdot b\le(a^2+b^2)/2$. Using the bound on the sum of the geometric series, the above inequality yields

$$\begin{aligned}\lim_{\nu,\nu'\to\infty}\sum_{l=1}^{\nu}\left(\sum_{t=l}^{\nu'}a^t(\lambda)^{t-l}\right)^2 &\le \frac{1}{1-\lambda}\lim_{\nu,\nu'\to\infty}\sum_{l=1}^{\nu}\sum_{t=l}^{\nu'}(a^t)^2(\lambda)^{t-l}=\frac{1}{1-\lambda}\lim_{\nu'\to\infty}\sum_{t=1}^{\nu'}\lim_{\nu\to\infty}\sum_{l=1}^{\min(\nu,t)}(a^t)^2(\lambda)^{t-l}\\ &\le\frac{1}{1-\lambda}\lim_{\nu'\to\infty}\sum_{t=1}^{\nu'}(a^t)^2\sum_{l=1}^{t}(\lambda)^{t-l}\le\frac{1}{(1-\lambda)^2}\lim_{\nu'\to\infty}\sum_{t=1}^{\nu'}(a^t)^2<\infty.\end{aligned}$$

■

A.4.2. On the properties of the best-response map

Some key properties of the best-response maps defined in (8) and (11) are summarized and proved next.

Proposition 14 *Let* $\left\{\left(\mathbf{D}^\nu,\mathbf{X}^\nu\right)\right\}_\nu$ *be the sequence generated by the* D^4L *Algorithm, in the setting of Theorem 2(a). Given the solution maps defined in* (8) *and* (11)*, the following hold:*

(a) There exist some constants $s_D>0$ *and* $\eta>0$*, and a sequence* $\{T^\nu\}_\nu$*, with* $\sum_{\nu=1}^{\infty}(T^\nu)^2<\infty$*, such that: for all* $\nu\ge1$*,*

$$\begin{aligned}&\left\langle\nabla_D F(\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu},\mathbf{X}^\nu),\sum_{i=1}^{I}\phi_i^\nu\left(\widetilde{\mathbf{D}}^\nu_{(i)}-\mathbf{D}^\nu_{(i)}\right)\right\rangle+\sum_{i=1}^{I}\phi_i^\nu\left(G(\widetilde{\mathbf{D}}^\nu_{(i)})-G(\mathbf{D}^\nu_{(i)})\right)\\ &\quad\le -s_D\Big(\|\widetilde{\mathbf{D}}^\nu-\mathbf{D}^\nu\|_F-\frac{I\bar{\epsilon}_\phi}{2\,s_D}T^\nu\Big)^2+\eta\,\|\widetilde{\mathbf{D}}^\nu-\mathbf{D}^\nu\|_F\sum_{t=1}^{\nu}(\rho)^{\nu-t}\|\mathbf{X}^t-\mathbf{X}^{t-1}\|_F+\frac{I^2\,\bar{\epsilon}_\phi^2}{4\,s_D}(T^\nu)^2,\end{aligned}\tag{108}$$

where $\rho \in (0,1)$ and $\bar{\epsilon}_\phi$ are defined in (48) and (49), respectively;
(b) There exist finite constants $s_X > 0$ and $L_X > 0$, such that: for all $\nu \geq 1$,

$$\sum_{i=1}^{I} \left\langle \nabla_{X_i} f_i(\overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu+1}}, \mathbf{X}_i^\nu), \mathbf{X}_i^{\nu+1} - \mathbf{X}_i^\nu \right\rangle + \sum_{i=1}^{I} \left(g_i(\mathbf{X}_i^{\nu+1}) - g_i(\mathbf{X}_i^\nu) \right)$$
$$\leq - \sum_{i=1}^{I} \tau_{X,i}^\nu ||\mathbf{X}_i^{\nu+1} - \mathbf{X}_i^\nu||_F^2 + L_X ||\mathbf{U}^\nu - \mathbf{1} \otimes \overline{\mathbf{U}}_{\boldsymbol{\phi}^\nu}||_F \; ||\mathbf{X}^{\nu+1} - \mathbf{X}^\nu||_F. \tag{109}$$

Proof **(a)** It follows from the optimality of $\widetilde{\mathbf{D}}_{(i)}^\nu$ [cf. (8)] and convexity of G that

$$\left\langle \nabla_D \tilde{f}_i(\widetilde{\mathbf{D}}_{(i)}^\nu; \mathbf{D}_{(i)}^\nu, \mathbf{X}_i^\nu) + I \boldsymbol{\Theta}_{(i)}^\nu - \nabla_D f_i(\mathbf{D}_{(i)}, \mathbf{X}_i), \mathbf{D}_{(i)}^\nu - \widetilde{\mathbf{D}}_{(i)}^\nu \right\rangle + G(\mathbf{D}_{(i)}^\nu) - G(\widetilde{\mathbf{D}}_{(i)}^\nu) \geq 0. \tag{110}$$

Adding and subtracting inside the first term $\sum_j \nabla_D f_j(\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}, \mathbf{X}_j^\nu)$ and using $\nabla_D \tilde{f}_i(\mathbf{D}_{(i)}^\nu; \mathbf{D}_{(i)}^\nu, \mathbf{X}_i^\nu) = \nabla_D f_i(\mathbf{D}_{(i)}^\nu, \mathbf{X}_i^\nu)$ [cf. Remark 7], inequality (110) becomes

$$\begin{aligned} &\left\langle \nabla_D \tilde{f}_i(\widetilde{\mathbf{D}}_{(i)}^\nu; \mathbf{D}_{(i)}^\nu, \mathbf{X}_i^\nu) - \nabla_D \tilde{f}_i(\mathbf{D}_{(i)}^\nu; \mathbf{D}_{(i)}^\nu, \mathbf{X}_i^\nu), \widetilde{\mathbf{D}}_{(i)}^\nu - \mathbf{D}_{(i)}^\nu \right\rangle \\ &+ \left\langle I \cdot \boldsymbol{\Theta}_{(i)}^\nu - \sum_{j=1}^{I} \nabla_D f_j(\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}, \mathbf{X}_j^\nu), \widetilde{\mathbf{D}}_{(i)}^\nu - \mathbf{D}_{(i)}^\nu \right\rangle \\ &+ \left\langle \sum_{j=1}^{I} \nabla_D f_j(\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}, \mathbf{X}_j^\nu), \widetilde{\mathbf{D}}_{(i)}^\nu - \mathbf{D}_{(i)}^\nu \right\rangle + G(\widetilde{\mathbf{D}}_{(i)}^\nu) - G(\mathbf{D}_{(i)}^\nu) \leq 0. \end{aligned}$$

Invoking the uniform strongly convexity of $\tilde{f}_i(\bullet; \mathbf{D}_{(i)}^\nu, \mathbf{X}_i^\nu)$, the definition of $\boldsymbol{\Theta}_{(i)}^\nu$ in (13), and recalling that $\nabla_D F(\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}, \mathbf{X}^\nu) = \sum_j \nabla_D f_j(\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}, \mathbf{X}_j^\nu)$, we get

$$\begin{aligned} &\left\langle \nabla_D F(\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}, \mathbf{X}^\nu), \widetilde{\mathbf{D}}_{(i)}^\nu - \mathbf{D}_{(i)}^\nu \right\rangle + G(\widetilde{\mathbf{D}}_{(i)}^\nu) - G(\mathbf{D}_{(i)}^\nu) \\ &\qquad \leq -\tau_{D,i}^\nu \; \left\| \widetilde{\mathbf{D}}_{(i)}^\nu - \mathbf{D}_{(i)}^\nu \right\|^2 + I \left\| \boldsymbol{\Theta}_{(i)}^\nu - \frac{1}{I} \sum_{j=1}^{I} \nabla_D f_j(\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}, \mathbf{X}_j^\nu) \right\|_F \left\| \widetilde{\mathbf{D}}_{(i)}^\nu - \mathbf{D}_{(i)}^\nu \right\|_F. \end{aligned}$$

Multiplying both side of the above inequality by the positive quantities ϕ_i^ν and summing over $i = 1, 2, \ldots, I$ while using $\phi_i^\nu \leq \bar{\epsilon}_\phi$ [cf. (49)], yields

$$\begin{aligned} &\left\langle \nabla_D F(\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}, \mathbf{X}^\nu), \sum_{i=1}^{I} \phi_i^\nu \left(\widetilde{\mathbf{D}}_{(i)}^\nu - \mathbf{D}_{(i)}^\nu \right) \right\rangle + \sum_{i=1}^{I} \phi_i^\nu \left(G(\widetilde{\mathbf{D}}_{(i)}^\nu) - G(\mathbf{D}_{(i)}^\nu) \right) \\ &\qquad \leq -s_D ||\widetilde{\mathbf{D}}^\nu - \mathbf{D}^\nu||^2 + I \bar{\epsilon}_\phi \underbrace{\left\| \boldsymbol{\Theta}^\nu - \mathbf{1} \otimes \frac{1}{I} \sum_{i=1}^{I} \nabla_D f_i(\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}, \mathbf{X}_i^\nu) \right\|_F}_{\text{gradient tracking error}} \left\| \widetilde{\mathbf{D}}^\nu - \mathbf{D}^\nu \right\|_F, \end{aligned} \tag{111}$$

where s_D is any positive constant such that $s_D \leq \min_{i,\nu} \phi_i^\nu \tau_{D,i}^\nu$ [note that such a constant exists because $\phi_i^\nu \geq \underline{\epsilon}_\phi$, with $\underline{\epsilon}_\phi > 0$ defined in (49), and all $\tau_{D,i}^\nu$ are uniformly bounded away from zero–see Assumption D1].

Now let us bound the *gradient tracking error* term in (111). Using (40) recursively, $\mathbf{\Theta}^\nu$ can be rewritten as

$$\mathbf{\Theta}^\nu = \widehat{\mathbf{W}}^{\nu-1:0}\,\mathbf{\Theta}^0 + \sum_{t=1}^{\nu-1} \widehat{\mathbf{W}}^{\nu-1:t} \left(\widehat{\mathbf{\Phi}}^t\right)^{-1} \left(\mathbf{G}^t - \mathbf{G}^{t-1}\right) + \left(\widehat{\mathbf{\Phi}}^\nu\right)^{-1} \left(\mathbf{G}^\nu - \mathbf{G}^{\nu-1}\right). \tag{112}$$

Using the definition of $\mathbf{G}^\nu$ [cf. (34)] and $\widehat{\mathbf{J}}$ [cf. (45)], write

$$\mathbf{1} \otimes \frac{1}{I} \sum_{i=1}^{I} \nabla_D f_i(\mathbf{D}_{(i)}^\nu, \mathbf{X}_i^\nu) = \widehat{\mathbf{J}}\mathbf{G}^\nu = \widehat{\mathbf{J}}\mathbf{G}^0 + \sum_{t=1}^{\nu} \widehat{\mathbf{J}} \left(\mathbf{G}^t - \mathbf{G}^{t-1}\right),$$

which, using $\mathbf{\Theta}^0 = \mathbf{G}^0$, leads to the following expansion for $\mathbf{1} \otimes \frac{1}{I}\sum_{i=1}^{I} \nabla_D f_i(\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}, \mathbf{X}_i^\nu)$:

$$\begin{aligned} \mathbf{1} \otimes \frac{1}{I} \sum_{i=1}^{I} \nabla_D f_i(\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}, \mathbf{X}_i^\nu) = \widehat{\mathbf{J}}\,\widetilde{\mathbf{\Theta}}^0 &+ \sum_{t=1}^{\nu} \widehat{\mathbf{J}} \left(\mathbf{G}^t - \mathbf{G}^{t-1}\right) \\ &+ \mathbf{1} \otimes \frac{1}{I} \sum_{i=1}^{I} \Big(\nabla_D f_i(\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}, \mathbf{X}_i^\nu) - \nabla_D f_i(\mathbf{D}_{(i)}^\nu, \mathbf{X}_i^\nu)\Big). \end{aligned} \tag{113}$$

Using (112) and (113), the *gradient tracking error* term in (111) can be upper bounded as

$$\begin{aligned} &\left\|\mathbf{\Theta}^\nu - \mathbf{1} \otimes \frac{1}{I} \sum_{i=1}^{I} \nabla_D f_i(\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}, \mathbf{X}_i^\nu)\right\|_F \\ &\overset{(a)}{\leq} \left\|\widehat{\mathbf{W}}^{\nu-1:0} - \widehat{\mathbf{J}}\right\|_2 \left\|\mathbf{\Theta}^0\right\|_F + \frac{1}{\underline{\epsilon}_\phi} \sum_{t=1}^{\nu-1} \left\|\widehat{\mathbf{W}}^{\nu-1:t} - \widehat{\mathbf{J}}_{\phi^t}\right\|_2 \left\|\mathbf{G}^t - \mathbf{G}^{t-1}\right\|_F \\ &\quad + \left\| \left(\widehat{\mathbf{\Phi}}^\nu\right)^{-1} - \widehat{\mathbf{J}}\right\|_2 \left\|\mathbf{G}^\nu - \mathbf{G}^{\nu-1}\right\|_F + \frac{1}{\sqrt{I}} \sum_{i=1}^{I} \left\|\nabla_D f_i(\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}, \mathbf{X}_i^\nu) - \nabla_D f_i(\mathbf{D}_{(i)}^\nu, \mathbf{X}_i^\nu)\right\|_F \\ &\overset{(b)}{\leq} c_4\,(\rho)^\nu + c_5\,L_\nabla \sum_{t=1}^{\nu} (\rho)^{\nu-t} \left(\left\|\mathbf{D}^t - \mathbf{D}^{t-1}\right\|_F + \left\|\mathbf{X}^t - \mathbf{X}^{t-1}\right\|_F\right) + L_\nabla \left\|\mathbf{D}^\nu - \mathbf{1} \otimes \overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}\right\|_F \\ &\overset{(c)}{=} T^\nu + c_5\,L_\nabla \sum_{t=1}^{\nu} (\rho)^{\nu-t} \left\|\mathbf{X}^t - \mathbf{X}^{t-1}\right\|_F, \end{aligned} \tag{114}$$

for some positive finite constants c_4 and c_5, where in (a) we used the lower bound $\phi_i^\nu \geq \underline{\epsilon}_\phi$ [cf. (49)] and $\widehat{\mathbf{J}}_{\phi^t} = \widehat{\mathbf{J}}\,\widehat{\mathbf{\Phi}}^t$ [cf. (47)]; and in (b) we used (34), (63) (cf. Remark 8), and Lemma 4; and in (c) we defined T^ν as

$$T^\nu \triangleq c_4\,(\rho)^\nu + c_5\,L_\nabla \sum_{t=1}^{\nu} (\rho)^{\nu-t} \left\|\mathbf{D}^t - \mathbf{D}^{t-1}\right\|_F + L_\nabla \left\|\mathbf{D}^\nu - \mathbf{1} \otimes \overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}\right\|_F. \tag{115}$$

Substituting (114) into (111) yields

$$\begin{aligned}
&\left\langle \nabla_D F(\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu},\mathbf{X}^\nu), \sum_{i=1}^I \phi_i^\nu \left(\widetilde{\mathbf{D}}^\nu_{(i)} - \mathbf{D}^\nu_{(i)}\right) \right\rangle + \sum_{i=1}^I \phi_i^\nu \left(G(\widetilde{\mathbf{D}}^\nu_{(i)}) - G(\mathbf{D}^\nu_{(i)})\right)\\
&\quad\le -s_D||\widetilde{\mathbf{D}}^\nu - \mathbf{D}^\nu||_F^2 + I\bar{\epsilon}_\phi T^\nu ||\widetilde{\mathbf{D}}^\nu - \mathbf{D}^\nu||_F\\
&\qquad + I\,\bar{\epsilon}_\phi\, c_5\, L_\nabla\, ||\widetilde{\mathbf{D}}^\nu - \mathbf{D}^\nu||_F \sum_{t=1}^{\nu} (\rho)^{\nu-t} \left\|\mathbf{X}^t - \mathbf{X}^{t-1}\right\|_F\\
&\quad= -s_D\left(||\widetilde{\mathbf{D}}^\nu - \mathbf{D}^\nu||_F - \frac{I\,\bar{\epsilon}_\phi}{2\,s_D} T^\nu\right)^2 + \frac{I^2\,\bar{\epsilon}_\phi^2}{4\,s_D}\,(T^\nu)^2\\
&\qquad + \eta\, ||\widetilde{\mathbf{D}}^\nu - \mathbf{D}^\nu||_F \sum_{t=1}^{\nu} (\rho)^{\nu-t} \left\|\mathbf{X}^t - \mathbf{X}^{t-1}\right\|_F,
\end{aligned}$$

with $\eta = I\bar{\epsilon}_\phi c_5 L_\nabla$.

To complete the proof, we need to show that $\sum_{\nu=1}^\infty (T^\nu)^2 < \infty$. Note that the first term on the RHS of (115) is square summable, and so is the third one, due to Proposition 5 [cf. (52)]. Invoking Jensen's inequality, it is sufficient to show that the second term on RHS of (115) is square summable. Following the same approach used to prove (52), we have

$$\begin{aligned}
&\lim_{\nu\to\infty} \sum_{t=1}^{\nu} \left(\sum_{l=1}^{t} (\rho)^{t-l} \left\|\mathbf{D}^l - \mathbf{D}^{l-1}\right\|_F \right)^2\\
&\le \lim_{\nu\to\infty} \sum_{t=1}^{\nu} \sum_{l=1}^{t} \sum_{k=1}^{t} (\rho)^{t-l} (\rho)^{t-k} \left\|\mathbf{D}^l - \mathbf{D}^{l-1}\right\|_F \left\|\mathbf{D}^k - \mathbf{D}^{k-1}\right\|_F\\
&\overset{(a)}{\le} \frac{1}{1-\rho} \lim_{\nu\to\infty} \sum_{t=1}^{\nu} \sum_{l=1}^{t} (\rho)^{t-l} (\gamma^l)^2 \left\|\widetilde{\mathbf{D}}^{l-1} - \mathbf{D}^{l-1}\right\|_F^2 \overset{(b)}{<} \infty,
\end{aligned}$$

where (a) follows from $ab \le (a^2 + b^2)/2$, and (b) is due to Lemma 13 and Assumption A3. Hence $\sum_{\nu=1}^\infty (T^\nu)^2 < \infty$.

(b) We prove this statement using the definition (15) of $\tilde{f}_i$; the same conclusion holds also using the alternative choice (14) of $\tilde{f}_i$; the proof is thus omitted. Invoking the optimality of $\mathbf{X}^{\nu+1}$ [cf. (11)] together with the convexity of g_i, yield

$$\begin{aligned}
&\left\langle \nabla_{X_i} \tilde{h}_i(\mathbf{X}_i^{\nu+1}; \mathbf{U}^\nu_{(i)}, \mathbf{X}_i^\nu) - \nabla_{X_i} \tilde{h}_i(\mathbf{X}_i^{\nu}; \mathbf{U}^\nu_{(i)}, \mathbf{X}_i^\nu), \mathbf{X}_i^\nu - \mathbf{X}_i^{\nu+1} \right\rangle\\
&+ \left\langle \nabla_{X_i} \tilde{h}_i(\mathbf{X}_i^{\nu}; \mathbf{U}^\nu_{(i)}, \mathbf{X}_i^\nu) - \nabla_{X_i} \tilde{h}_i(\mathbf{X}_i^{\nu}; \overline{\mathbf{U}}_{\boldsymbol{\phi}^\nu}, \mathbf{X}_i^\nu), \mathbf{X}_i^\nu - \mathbf{X}_i^{\nu+1} \right\rangle\\
&+ \left\langle \nabla_{X_i} \tilde{h}_i(\mathbf{X}_i^{\nu}; \overline{\mathbf{U}}_{\boldsymbol{\phi}^\nu}, \mathbf{X}_i^\nu), \mathbf{X}_i^\nu - \mathbf{X}_i^{\nu+1} \right\rangle + g_i\left(\mathbf{X}_i^\nu\right) - g_i(\mathbf{X}_i^{\nu+1}) \ge 0.
\end{aligned}$$

Using Remark 7 and $\overline{\mathbf{U}}_{\boldsymbol{\phi}^\nu} = \overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu+1}}$ [cf. (42)], we obtain

$$\begin{aligned}
&\left\langle \nabla_{X_i} f_i(\overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu+1}}, \mathbf{X}_i^\nu), \mathbf{X}_i^{\nu+1} - \mathbf{X}_i^\nu \right\rangle + g_i(\mathbf{X}_i^{\nu+1}) - g_i\left(\mathbf{X}_i^\nu\right)\\
&\qquad\le -\tau^\nu_{X,i} \left\|\mathbf{X}_i^{\nu+1} - \mathbf{X}_i^\nu\right\|_F^2 + \left\langle \nabla_{X_i} f_i(\overline{\mathbf{U}}_{\boldsymbol{\phi}^\nu}, \mathbf{X}_i^\nu) - \nabla_{X_i} f_i(\mathbf{U}^\nu_{(i)}, \mathbf{X}_i^\nu), \mathbf{X}_i^{\nu+1} - \mathbf{X}_i^\nu \right\rangle,
\end{aligned}$$

which, together with (63), leads to the desired result (109), with $L_X = \sqrt{I} L_\nabla$. ■

A.4.3. Proof of Proposition 10

We begin observing that

$$G(\overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu+1}}) \leq G(\overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu}}) + \frac{\gamma^\nu}{I}\sum_{i=1}^{I}\phi_i^\nu\left(G(\widetilde{\mathbf{D}}_{(i)}^\nu) - G(\mathbf{D}_{(i)}^\nu)\right) + \frac{\gamma^\nu}{I}\sum_{i=1}^{I}\phi_i^\nu\left(G(\mathbf{D}_{(i)}^\nu) - G(\overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu}})\right), \tag{116}$$

due to (43) and the convexity of G [together with $\mathbf{1}^\intercal\boldsymbol{\phi}^\nu = 1$].

Invoking the descent lemma for $f_i(\overline{\mathbf{U}}_{\boldsymbol{\phi}^\nu}, \bullet)$ and using $\overline{\mathbf{U}}_{\boldsymbol{\phi}^\nu} = \overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu+1}}$ [cf. (42)], we get: for sufficiently large ν, say $\nu \geq \nu_0$,

$$\begin{aligned}
&U(\overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu+1}}, \mathbf{X}^{\nu+1}) \\
&\leq \sum_{i=1}^{I}\left\{f_i(\overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu+1}}, \mathbf{X}_i^\nu) + g_i(\mathbf{X}_i^\nu)\right\} + G(\overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu+1}}) + \sum_{i=1}^{I}\left\{g_i(\mathbf{X}_i^{\nu+1}) - g_i(\mathbf{X}_i^\nu)\right\} \\
&\quad + \sum_{i=1}^{I}\left\langle \nabla_{X_i} f_i(\overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu+1}}, \mathbf{X}_i^\nu), \mathbf{X}_i^{\nu+1} - \mathbf{X}_i^\nu\right\rangle + \frac{1}{2}\sum_{i=1}^{I} L_{\nabla X_i}(\overline{\mathbf{U}}_{\boldsymbol{\phi}^\nu})\left\|\mathbf{X}_i^{\nu+1} - \mathbf{X}_i^\nu\right\|_F^2 \\
&\overset{(a)}{\leq} U(\overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu+1}}, \mathbf{X}^\nu) - \sum_{i=1}^{I}\left\{\left(\tau_{X,i}^\nu - \frac{1}{2}L_{\nabla X_i}(\overline{\mathbf{U}}_{\boldsymbol{\phi}^\nu})\right)\left\|\mathbf{X}_i^{\nu+1} - \mathbf{X}_i^\nu\right\|_F^2\right\} \\
&\quad + L_X||\mathbf{U}^\nu - \mathbf{1}\otimes\overline{\mathbf{U}}_{\boldsymbol{\phi}^\nu}||_F\ ||\mathbf{X}^{\nu+1} - \mathbf{X}^\nu||_F \\
&\overset{(b)}{\leq} U(\overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu+1}}, \mathbf{X}^\nu) - s_X||\mathbf{X}^{\nu+1} - \mathbf{X}^\nu||_F^2 + L_X||\mathbf{U}^\nu - \mathbf{1}\otimes\overline{\mathbf{U}}_{\boldsymbol{\phi}^\nu}||_F\ ||\mathbf{X}^{\nu+1} - \mathbf{X}^\nu||_F,
\end{aligned} \tag{117}$$

where in (a) we used Proposition 14(b); and in (b) $s_X > 0$ is a constant such that $\inf_{\nu\geq\nu_0}(\tau_{X,i}^\nu - \frac{1}{2}L_{\nabla X_i}(\overline{\mathbf{U}}_{\boldsymbol{\phi}^\nu})) \geq s_X$, for all $i = 1, \ldots, I$. Note that such a constant exists because of (58) and Assumption D1.

To upper bound $U(\overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu+1}}, \mathbf{X}^\nu)$, we apply the descent lemma to $F(\bullet, \mathbf{X}^\nu)$. Recalling that $\nabla_D F(\bullet, \mathbf{X}^\nu)$ is Lipschitz continuous with constant L_{∇_D} and using (43), we get

$$\begin{aligned}
U(\overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu+1}}, \mathbf{X}^\nu) &\leq F(\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}, \mathbf{X}^\nu) + \frac{\gamma^\nu}{I}\left\langle \nabla_D F(\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}, \mathbf{X}^\nu), \sum_{i=1}^{I}\phi_i^\nu\left(\widetilde{\mathbf{D}}_{(i)}^\nu - \mathbf{D}_{(i)}^\nu\right)\right\rangle \\
&\quad + \frac{L_{\nabla_D}}{2}\left(\frac{\gamma^\nu}{I}\right)^2\left\|\sum_{i=1}^{I}\phi_i^\nu\left(\widetilde{\mathbf{D}}_{(i)}^\nu - \mathbf{D}_{(i)}^\nu\right)\right\|_F^2 + \sum_{i=1}^{I} g_i(\mathbf{X}_i^\nu) + G(\overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu+1}}), \\
&\overset{(a)}{\leq}\ U(\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}, \mathbf{X}^\nu) - \frac{s_D\cdot\gamma^\nu}{I}\left(||\widetilde{\mathbf{D}}^\nu - \mathbf{D}^\nu||_F - \frac{I\bar{\epsilon}_\phi}{2s_D}T^\nu\right)^2 \\
&\quad + \frac{\eta\cdot\gamma^\nu}{I}||\widetilde{\mathbf{D}}^\nu - \mathbf{D}^\nu||_F\sum_{t=1}^{\nu}(\rho)^{\nu-t}||\mathbf{X}^t - \mathbf{X}^{t-1}||_F + \frac{I\bar{\epsilon}_\phi^2\gamma^\nu}{4s_D}(T^\nu)^2 \\
&\quad + \frac{L_{\nabla_D}}{2}(\gamma^\nu)^2\,||\widetilde{\mathbf{D}}^\nu - \mathbf{D}^\nu||_F^2 + \underbrace{\frac{\gamma^\nu}{I}\sum_{i=1}^{I}\phi_i^\nu\left(G(\mathbf{D}_{(i)}^\nu) - G(\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu})\right)}_{\leq L_G\gamma^\nu\left\|\mathbf{D}^\nu - \mathbf{1}\otimes\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}\right\|_F},
\end{aligned} \tag{118}$$

where in (a) we used (116) and Proposition 14(a), and the Lipschitz continuity of G, due to the convexity of G (G is thus locally Lipschitz continuous) and the compactness of $\mathcal{D}$; we denoted by $L_G > 0$ the Lipschitz constant.

Combining (117) with (118) and defining $\bar{\tau}_D^\nu \triangleq s_D - \gamma^\nu \frac{IL_{\nabla_D}}{2}$, we get: for $\nu \geq \nu_0$,

$$\begin{aligned} U(\overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu+1}}, \mathbf{X}^{\nu+1}) \leq\ & U(\overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu}, \mathbf{X}^\nu) - s_X ||\mathbf{X}^{\nu+1} - \mathbf{X}^\nu||_F^2 + L_X ||\mathbf{U}^\nu - \mathbf{1} \otimes \overline{\mathbf{U}}_{\boldsymbol{\phi}^\nu}||_F \; ||\mathbf{X}^{\nu+1} - \mathbf{X}^\nu||_F \\ & - \frac{\bar{\tau}_D^\nu \gamma^\nu}{I} \left(||\widetilde{\mathbf{D}}^\nu - \mathbf{D}^\nu||_F - \frac{I\,\bar{\epsilon}_\phi}{2\,\bar{\tau}_D^\nu} T^\nu \right)^2 + \frac{I\,\bar{\epsilon}_\phi^2\,\gamma^\nu}{4\,\bar{\tau}_D^\nu} \, (T^\nu)^2 \\ & + \frac{\eta\,\gamma^\nu}{I} ||\widetilde{\mathbf{D}}^\nu - \mathbf{D}^\nu||_F \underbrace{\sum_{t=1}^{\nu} (\rho)^{\nu-t} ||\mathbf{X}^t - \mathbf{X}^{t-1}||_F}_{\leq c_5\,(\rho)^\nu + \rho^{-1} \sum_{t=1}^{\nu} (\rho)^{\nu-t} ||\mathbf{X}^{t+1} - \mathbf{X}^t||_F} + L_G\,\gamma^\nu \left\| \mathbf{D}^\nu - \mathbf{1} \otimes \overline{\mathbf{D}}_{\boldsymbol{\phi}^\nu} \right\|_F , \end{aligned} \tag{119}$$

where $c_5 \triangleq (\rho)^{-1} ||\mathbf{X}^1 - \mathbf{X}^0||_F$. Since $\gamma^\nu \downarrow 0$, there exists an integer $\nu_1 \geq \nu_0$ and some $\bar{\tau}_D$ such that $\bar{\tau}_D^\nu \geq \bar{\tau}_D > 0$, for all $\nu \geq \nu_1$. Let $\bar{\nu}$ be any integer $\bar{\nu} \geq \nu_1$. Then, applying (119) recursively on $\nu, \nu - 1, \ldots, \bar{\nu} + 1, \bar{\nu}$, and using the boundedness of $\{||\widetilde{\mathbf{D}}^\nu - \mathbf{D}^\nu||_F\}_\nu$, we obtain

$$\begin{aligned} & U(\overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu+1}}, \mathbf{X}^{\nu+1}) \\ & \leq U(\overline{\mathbf{D}}_{\boldsymbol{\phi}^{\bar{\nu}}}, \mathbf{X}^{\bar{\nu}}) - s_X \sum_{l=\bar{\nu}}^{\nu} ||\mathbf{X}^{l+1} - \mathbf{X}^l||_F^2 + L_X \sum_{l=\bar{\nu}}^{\nu} ||\mathbf{U}^l - \mathbf{1} \otimes \overline{\mathbf{U}}_{\boldsymbol{\phi}^l}||_F \; ||\mathbf{X}^{l+1} - \mathbf{X}^l||_F \\ & - \frac{\bar{\tau}_D}{I} \sum_{l=\bar{\nu}}^{\nu} \gamma^l \left(||\widetilde{\mathbf{D}}^l - \mathbf{D}^l||_F - \frac{I\bar{\epsilon}_\phi}{2\,\bar{\tau}_D} T^l \right)^2 + \frac{I\bar{\epsilon}_\phi^2}{4\,\bar{\tau}_D} \sum_{l=\bar{\nu}}^{\nu} \gamma^l \left(T^l \right)^2 \\ & + c_6 \sum_{l=\bar{\nu}}^{\nu} \gamma^l (\rho)^l + c_7 \sum_{l=\bar{\nu}}^{\nu} \sum_{t=1}^{l} \gamma^l (\rho)^{l-t} \left\| \mathbf{X}^{t+1} - \mathbf{X}^t \right\|_F + L_G \sum_{l=\bar{\nu}}^{\nu} \gamma^l \left\| \mathbf{D}^l - \mathbf{1} \otimes \overline{\mathbf{D}}_{\boldsymbol{\phi}^l} \right\|_F , \end{aligned} \tag{120}$$

for some finite constants $c_6, c_7 > 0$. Using the boundedness of $\{||\mathbf{X}^{\nu+1} - \mathbf{X}^\nu||_F\}_\nu$ (cf. Step 2), i.e., $||\mathbf{X}^{\nu+1} - \mathbf{X}^\nu||_F \leq B_X$, for all ν and some $B_X > 0$, we can bound the double-sum term on the RHS of (120) as

$$\begin{aligned} & \sum_{l=\bar{\nu}}^{\nu} \sum_{t=1}^{l} \gamma^l (\rho)^{l-t} \left\| \mathbf{X}^{t+1} - \mathbf{X}^t \right\|_F = \sum_{l=1}^{\nu} \sum_{t=\max(\bar{\nu},l)}^{\nu} \gamma^t (\rho)^{t-l} \left\| \mathbf{X}^{l+1} - \mathbf{X}^l \right\|_F \\ & = \sum_{l=1}^{\bar{\nu}-1} \left\| \mathbf{X}^{l+1} - \mathbf{X}^l \right\|_F \sum_{t=\bar{\nu}}^{\nu} \gamma^t (\rho)^{t-l} + \sum_{l=\bar{\nu}}^{\nu} \left\| \mathbf{X}^{l+1} - \mathbf{X}^l \right\|_F \sum_{t=l}^{\nu} \gamma^t (\rho)^{t-l} \\ & \overset{(a)}{\leq} B_X \cdot \left(\max_{\bar{\nu} \leq t \leq \nu} \gamma^t \right) \cdot \frac{\left(1 - (\rho)^{\bar{\nu}}\right)\left(1 - (\rho)^{\nu - \bar{\nu} + 1}\right)}{(1 - \rho)^2} + \sum_{l=\bar{\nu}}^{\nu} \left\| \mathbf{X}^{l+1} - \mathbf{X}^l \right\|_F \sum_{t=l}^{\infty} \gamma^t (\rho)^{t-l} \\ & \leq \frac{B_X}{(1-\rho)^2} \left(\max_{t \geq \bar{\nu}} \gamma^t \right) + \sum_{l=\bar{\nu}}^{\nu} \left\| \mathbf{X}^{l+1} - \mathbf{X}^l \right\|_F \sum_{t=l}^{\infty} \gamma^t (\rho)^{t-l}, \end{aligned} \tag{121}$$

where in (a) we used the summability of $\sum_{t=l}^{\nu} \gamma^t (\rho)^{t-l}$, and the following bound

$$\sum_{l=1}^{\bar{\nu}-1} \sum_{t=\bar{\nu}}^{\nu} \gamma^t (\rho)^{t-l} \leq \left(\max_{t \geq \bar{\nu}} \gamma^t \right) \frac{\left(1 - (\rho)^{\bar{\nu}}\right)\left(1 - (\rho)^{\nu - \bar{\nu} + 1}\right)}{(1 - \rho)^2}.$$

Substituting (121) in (120), yields

$$
\begin{aligned}
U(\overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu+1}},\mathbf{X}^{\nu+1}) \leq\; & U(\overline{\mathbf{D}}_{\boldsymbol{\phi}^{\bar\nu}},\mathbf{X}^{\bar\nu}) - s_X \sum_{l=\bar\nu}^{\nu}\left\|\mathbf{X}^{l+1}-\mathbf{X}^{l}\right\|_F^2 \\
&+ \sum_{l=\bar\nu}^{\nu}\underbrace{\left(c_7\sum_{t=l}^{\infty}\gamma^t(\rho)^{t-l} + L_X\|\mathbf{U}^l - \mathbf{1}\otimes\overline{\mathbf{U}}_{\boldsymbol{\phi}^l}\|_F\right)}_{\triangleq Z^l}\left\|\mathbf{X}^{l+1}-\mathbf{X}^{l}\right\|_F \\
&- \frac{\bar\tau_D}{I}\sum_{l=\bar\nu}^{\nu}\gamma^l\left(\|\widetilde{\mathbf{D}}^l-\mathbf{D}^l\|_F - \frac{I\bar\epsilon_\phi}{2\,\bar\tau_D}T^l\right)^2 + \frac{I\bar\epsilon_\phi^2}{4\,\bar\tau_D}\sum_{l=\bar\nu}^{\nu}\gamma^l\left(T^l\right)^2 \\
&+ \underbrace{\left[\frac{c_6}{1-\rho}\left((\rho)^{\bar\nu}-(\rho)^{\nu+1}\right)+\frac{c_7B_X}{(1-\rho)^2}\right]\cdot\left(\max_{t\geq\bar\nu}\gamma^t\right)}_{\triangleq E^{\nu,\bar\nu}} + L_G\sum_{l=\bar\nu}^{\nu}\gamma^l\left\|\mathbf{D}^l-\mathbf{1}\otimes\overline{\mathbf{D}}_{\boldsymbol{\phi}^l}\right\|_F
\end{aligned}
$$

which complete the proof. □

A.4.4. Proof of Lemma 11

The result follows by squaring both sides of eq. (114) [in the proof of Proposition 14 (a)], using $\sum_{\nu=1}^{\infty}(T^\nu)^2<\infty$ [cf. Proposition 14 (a)], and

$$
\begin{aligned}
&\lim_{\nu\to\infty}\sum_{t=1}^{\nu}\left(\sum_{l=1}^{t}(\rho)^{t-l}\left\|\mathbf{X}^l-\mathbf{X}^{l-1}\right\|_F\right)^2 \\
&\leq \lim_{\nu\to\infty}\sum_{t=1}^{\nu}\sum_{l=1}^{t}\sum_{k=1}^{t}(\rho)^{t-l}(\rho)^{t-k}\left\|\mathbf{X}^l-\mathbf{X}^{l-1}\right\|_F\left\|\mathbf{X}^k-\mathbf{X}^{k-1}\right\|_F \\
&\overset{(a)}{\leq} \frac{1}{1-\rho}\lim_{\nu\to\infty}\sum_{t=1}^{\nu}\sum_{l=1}^{t}(\rho)^{t-l}\left\|\mathbf{X}^l-\mathbf{X}^{l-1}\right\|_F^2 \overset{(b)}{<}\infty,
\end{aligned}
$$

where (a) follows from $ab\leq(a^2+b^2)/2,\forall a,b\in\mathbb{R}$, and (b) is due to (76) and Lemma 13. □

A.4.5. Proof of Proposition 12

(a) Using the optimality of $\widetilde{\mathbf{D}}_{(i)}^{\nu}$ defined in (8) together with convexity of G, yields

$$
\begin{aligned}
&\left\langle\nabla_D\tilde f_i(\widetilde{\mathbf{D}}_{(i)}^{\nu_1};\mathbf{D}_{(i)}^{\nu_1},\mathbf{X}_i^{\nu_1}) + I\cdot\boldsymbol{\Theta}_{(i)}^{\nu_1} - \nabla_D f_i(\mathbf{D}_{(i)}^{\nu_1},\mathbf{X}_i^{\nu_1}),\widetilde{\mathbf{D}}_{(i)}^{\nu_2}-\widetilde{\mathbf{D}}_{(i)}^{\nu_1}\right\rangle + G(\widetilde{\mathbf{D}}_{(i)}^{\nu_2}) - G(\widetilde{\mathbf{D}}_{(i)}^{\nu_1})\geq 0,\\
&\left\langle\nabla_D\tilde f_i(\widetilde{\mathbf{D}}_{(i)}^{\nu_2};\mathbf{D}_{(i)}^{\nu_2},\mathbf{X}_i^{\nu_2}) + I\cdot\boldsymbol{\Theta}_{(i)}^{\nu_2} - \nabla_D f_i(\mathbf{D}_{(i)}^{\nu_2},\mathbf{X}_i^{\nu_2}),\widetilde{\mathbf{D}}_{(i)}^{\nu_1}-\widetilde{\mathbf{D}}_{(i)}^{\nu_2}\right\rangle + G(\widetilde{\mathbf{D}}_{(i)}^{\nu_1}) - G(\widetilde{\mathbf{D}}_{(i)}^{\nu_2})\geq 0.
\end{aligned}
$$

Summing the two inequalities above while adding/subtracting inside the inner product $\nabla_D\tilde f_i(\widetilde{\mathbf{D}}_{(i)}^{\nu_2};\mathbf{D}_{(i)}^{\nu_1},\mathbf{X}_i^{\nu_1})$ and using (13), yield

$$
\begin{aligned}
&\left\langle\nabla_D\tilde f_i(\widetilde{\mathbf{D}}_{(i)}^{\nu_2};\mathbf{D}_{(i)}^{\nu_1},\mathbf{X}_i^{\nu_1}) - \nabla_D\tilde f_i(\widetilde{\mathbf{D}}_{(i)}^{\nu_2};\mathbf{D}_{(i)}^{\nu_2},\mathbf{X}_i^{\nu_2}),\widetilde{\mathbf{D}}_{(i)}^{\nu_2}-\widetilde{\mathbf{D}}_{(i)}^{\nu_1}\right\rangle \\
&-\left\langle\nabla_D f_i(\mathbf{D}_{(i)}^{\nu_1},\mathbf{X}_i^{\nu_1}) - \nabla_D f_i(\mathbf{D}_{(i)}^{\nu_2},\mathbf{X}_i^{\nu_2}),\widetilde{\mathbf{D}}_{(i)}^{\nu_2}-\widetilde{\mathbf{D}}_{(i)}^{\nu_1}\right\rangle + I\left\langle\boldsymbol{\Theta}_{(i)}^{\nu_1}-\boldsymbol{\Theta}_{(i)}^{\nu_2},\widetilde{\mathbf{D}}_{(i)}^{\nu_2}-\widetilde{\mathbf{D}}_{(i)}^{\nu_1}\right\rangle \\
&\geq\left\langle\nabla_D\tilde f_i(\widetilde{\mathbf{D}}_{(i)}^{\nu_1};\mathbf{D}_{(i)}^{\nu_1},\mathbf{X}_i^{\nu_1}) - \nabla_D\tilde f_i(\widetilde{\mathbf{D}}_{(i)}^{\nu_2};\mathbf{D}_{(i)}^{\nu_1},\mathbf{X}_i^{\nu_1}),\widetilde{\mathbf{D}}_{(i)}^{\nu_1}-\widetilde{\mathbf{D}}_{(i)}^{\nu_2}\right\rangle \geq \tau_{D,i}^{\nu_1}\,\|\widetilde{\mathbf{D}}_{(i)}^{\nu_2}-\widetilde{\mathbf{D}}_{(i)}^{\nu_1}\|_F^2,
\end{aligned}
\tag{122}
$$

where the second inequality follows from the $\tau_{D,i}^{\nu_1}$-strong convexity of $\tilde{f}_i(\bullet; \mathbf{D}_{(i)}^{\nu_1}, \mathbf{X}_i^{\nu_1})$ [cf. Remark 7]. To bound the first term on the LHS of the above inequality, let us use the expression (15) of $\tilde{f}_i$, and write

$$\begin{aligned}
&\nabla_D \tilde{f}_i(\widetilde{\mathbf{D}}_{(i)}^{\nu_2}; \mathbf{D}_{(i)}^{\nu_1}, \mathbf{X}_i^{\nu_1}) - \nabla_D \tilde{f}_i(\widetilde{\mathbf{D}}_{(i)}^{\nu_2}; \mathbf{D}_{(i)}^{\nu_2}, \mathbf{X}_i^{\nu_2}) \\
&\quad = \nabla_D f_i(\mathbf{D}_{(i)}^{\nu_1}, \mathbf{X}_i^{\nu_1}) - \nabla_D f_i(\mathbf{D}_{(i)}^{\nu_2}, \mathbf{X}_i^{\nu_2}) + \tau_{D,i}^{\nu_1}\left(\mathbf{D}_{(i)}^{\nu_2} - \mathbf{D}_{(i)}^{\nu_1}\right) + (\tau_{D,i}^{\nu_1} - \tau_{D,i}^{\nu_2})\left(\widetilde{\mathbf{D}}_{(i)}^{\nu_2} - \mathbf{D}_{(i)}^{\nu_2}\right).
\end{aligned} \tag{123}$$

Substituting (123) in (122), we get

$$\begin{aligned}
&\left\|\widetilde{\mathbf{D}}_{(i)}^{\nu_2} - \widetilde{\mathbf{D}}_{(i)}^{\nu_1}\right\|_F \\
\leq & \frac{|\tau_{D,i}^{\nu_1} - \tau_{D,i}^{\nu_2}|}{\tau_{D,i}^{\nu_1}} \left\|\widetilde{\mathbf{D}}_{(i)}^{\nu_2} - \mathbf{D}_{(i)}^{\nu_2}\right\|_F + \left\|\mathbf{D}_{(i)}^{\nu_1} - \mathbf{D}_{(i)}^{\nu_2}\right\|_F + \frac{I}{\tau_{D,i}^{\nu_1}} \left\|\mathbf{\Theta}_{(i)}^{\nu_1} - \mathbf{\Theta}_{(i)}^{\nu_2}\right\|_F \\
\stackrel{(a)}{\leq} & \frac{|\tau_{D,i}^{\nu_1} - \tau_{D,i}^{\nu_2}|}{\tau_{D,i}^{\nu_1}} \left\|\widetilde{\mathbf{D}}_{(i)}^{\nu_2} - \mathbf{D}_{(i)}^{\nu_2}\right\|_F + \left\|\overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu_2}} - \overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu_1}}\right\|_F + \left\|\mathbf{D}_{(i)}^{\nu_1} - \overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu_1}}\right\|_F + \left\|\mathbf{D}_{(i)}^{\nu_2} - \overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu_2}}\right\|_F \\
&+ \frac{I}{\tau_{D,i}^{\nu_1}} \left\|\mathbf{\Theta}_{(i)}^{\nu_1} - \frac{1}{I}\sum_{i=1}^{I} \nabla_D f_i(\overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu_1}}, \mathbf{X}_i^{\nu_1})\right\|_F \\
&+ \frac{I}{\tau_{D,i}^{\nu_1}} \left\|\mathbf{\Theta}_{(i)}^{\nu_2} - \frac{1}{I}\sum_{i=1}^{I} \nabla_D f_i(\overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu_2}}, \mathbf{X}_i^{\nu_2})\right\|_F \\
&+ \frac{L_{\nabla_D}}{\tau_{D,i}^{\nu_1}} \left\|(\overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu_1}}, \mathbf{X}^{\nu_1}) - (\overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu_2}}, \mathbf{X}^{\nu_2})\right\|,
\end{aligned} \tag{124}$$

where (a) holds by add/subtracting average quantities $\frac{1}{I}\sum_{i=1}^{I} \nabla_D f_i(\overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu}}, \mathbf{X}_i^{\nu})$ and $\overline{\mathbf{D}}_{(i)}^{\nu}$, triangle inequality, and invoking the Lipschitz continuity bound (63). By the compactness of $\mathcal{D}$, we have $||\widetilde{\mathbf{D}}_{(i)}^{\nu_2} - \mathbf{D}_{(i)}^{\nu_2}||_F \leq B_D$, for some finite $B_D > 0$. Furthermore, by Assumption D, $\tau_{D,i}^{\nu}$ is convergent to some $\tau_{D,i}^{\infty} > 0$ and there exists a sufficiently small $\tilde{s}_D > 0$ such that $\tilde{s}_D \leq \tau_{D,i}^{\nu} \leq \tilde{s}_D^{-1}$, for all ν and i. Thus (124) gives

$$\left\|\widetilde{\mathbf{D}}_{(i)}^{\nu_2} - \widetilde{\mathbf{D}}_{(i)}^{\nu_1}\right\|_F \leq \left(\frac{L_{\nabla_D}}{\tilde{s}_D} + 1\right) \left\|(\overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu_1}}, \mathbf{X}^{\nu_1}) - (\overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu_2}}, \mathbf{X}^{\nu_2})\right\| + \tilde{T}_i^{\nu_1} + \tilde{T}_i^{\nu_2} \tag{125}$$

with $\tilde{T}_i^{\nu} \triangleq \frac{I}{\tilde{s}_D}||\widetilde{\mathbf{\Theta}}_i^{\nu} - \frac{1}{I}\sum_{i=1}^{I} \nabla_D f_i(\overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu}}, \mathbf{X}_i^{\nu})||_F + ||\mathbf{D}_{(i)}^{\nu} - \overline{\mathbf{D}}_{\boldsymbol{\phi}^{\nu}}||_F + \frac{B_D}{\tilde{s}_D}|\tau_{D,i}^{\nu} - \tau_{D,i}^{\infty}|$. Convergence of $\tau_{D,i}^{\nu}$ (Assumption D2), Lemma 11 and Proposition 5 yield $\tilde{T}_i^{\nu} \to 0$ as $\nu \to \infty$. Summing (125) leads to the desired result, with $L_D = I(\frac{L_{\nabla_D}}{\tilde{s}_D} + 1)$ and $\tilde{T}^{\nu} \triangleq \sum_{i=1}^{I} \tilde{T}_i^{\nu}$.

It is clear that the claim also holds when (14) is chosen for $\tilde{f}_i$ [specifically, trivial extension is to modify the RHS of (123) and following a similar steps as in the rest of the proof]; we omit further details.

(b) We prove (86) when $\tilde{h}_i$ is given by (17); we leave the proof under (16) to the reader, since it is almost identical to that under (17).

Invoking optimality of each $\mathbf{X}_i^\nu$ and $\mathbf{X}_i^{\nu+1}$ defined in (11) while using the strong convexity of $\tilde{h}_i(\bullet;\mathbf{U}_{(i)}^\nu,\mathbf{X}_i^\nu)$ and g_i's, it is not difficult to show that the following holds:

$$\begin{aligned}(\tau_{X,i}^\nu+\mu_i)\left\|\mathbf{X}_i^{\nu+1}-\mathbf{X}_i^\nu\right\|_F^2 \le \\ \left\langle \nabla_{X_i}\tilde{h}_i(\mathbf{X}_i^\nu;\mathbf{U}_{(i)}^\nu,\mathbf{X}_i^\nu)-\nabla_{X_i}\tilde{h}_i(\mathbf{X}_i^\nu;\mathbf{U}_{(i)}^{\nu-1},\mathbf{X}_i^{\nu-1}),\mathbf{X}_i^\nu-\mathbf{X}_i^{\nu+1}\right\rangle.\end{aligned}$$

Using (17), the definition $\Upsilon_i^\nu(\mathbf{X}_i)\triangleq\tau_{X,i}^\nu\mathbf{X}_i-\nabla_{X_i}f_i(\mathbf{U}_{(i)}^\nu,\mathbf{X}_i)$, and the Cauchy-Schwarz inequality, the above inequality yields

$$\begin{aligned}(\tau_{X,i}^\nu+\mu_i)\left\|\mathbf{X}_i^{\nu+1}-\mathbf{X}_i^\nu\right\|_F^2 &\le \left\|\Upsilon_i^\nu(\mathbf{X}_i^\nu)-\Upsilon_i^\nu(\mathbf{X}_i^{\nu-1})\right\|_F\left\|\mathbf{X}_i^{\nu+1}-\mathbf{X}_i^\nu\right\|_F\\ &+\left\|\nabla_{X_i}f_i(\mathbf{U}_{(i)}^\nu,\mathbf{X}_i^{\nu-1})-\nabla_{X_i}f_i(\mathbf{U}_{(i)}^{\nu-1},\mathbf{X}_i^{\nu-1})\right\|_F\left\|\mathbf{X}_i^{\nu+1}-\mathbf{X}_i^\nu\right\|_F\\ &+|\tau_{X,i}^\nu-\tau_{X,i}^{\nu-1}|\left\|\mathbf{X}_i^\nu-\mathbf{X}_i^{\nu-1}\right\|_F\left\|\mathbf{X}_i^{\nu+1}-\mathbf{X}_i^\nu\right\|_F.\end{aligned}\tag{126}$$

Following the same steps used to prove (61), it is not difficult to check that, under Assumption D1, $||\Upsilon_i^\nu(\mathbf{X}_i^\nu)-\Upsilon_i^\nu(\mathbf{X}_i^0)||_F\le\tau_{X,i}^\nu||\mathbf{X}_i^\nu-\mathbf{X}_i^0||_F$. Using in (126) this bound together with the Lipschitz continuity of $\nabla_{X_i}f_i(\bullet,\mathbf{X}_i^{\nu-1})$ [cf. Remark 8] and summing over i, yield

$$\begin{aligned}||\mathbf{X}^{\nu+1}-\mathbf{X}^\nu||_F^2&=\sum_{i=1}^I||\mathbf{X}_i^{\nu+1}-\mathbf{X}_i^\nu||_F^2\\ &\le\sum_{i=1}^I\frac{\tau_{X,i}^\nu+|\tau_{X,i}^\nu-\tau_{X,i}^{\nu-1}|}{\tau_{X,i}^\nu+\mu_i}||\mathbf{X}_i^\nu-\mathbf{X}_i^{\nu-1}||_F\;||\mathbf{X}_i^{\nu+1}-\mathbf{X}_i^\nu||_F\\ &+\sum_{i=1}^I\frac{L_\nabla}{\tau_{X,i}^\nu+\mu_i}||\mathbf{U}_{(i)}^\nu-\mathbf{U}_{(i)}^{\nu-1}||_F\;||\mathbf{X}_i^{\nu+1}-\mathbf{X}_i^\nu||_F.\end{aligned}\tag{127}$$

Define

$$p_X^\nu\triangleq\max_i\frac{\tau_{X,i}^\nu+|\tau_{X,i}^\nu-\tau_{X,i}^{\nu-1}|}{\tau_{X,i}^\nu+\mu_i},\quad\text{and}\quad q_X^\nu\triangleq\max_i\frac{L_\nabla}{\tau_{X,i}^\nu+\mu_i}.\tag{128}$$

Note that $0<p_X^\nu,q_X^\nu<\infty$. Then, (127) becomes

$$\begin{aligned}||\mathbf{X}^{\nu+1}-\mathbf{X}^\nu||_F^2&\le p_X^\nu\sum_{i=1}^I||\mathbf{X}_i^\nu-\mathbf{X}_i^{\nu-1}||_F\;||\mathbf{X}_i^{\nu+1}-\mathbf{X}_i^\nu||_F+q_X^\nu\sum_{i=1}^I||\mathbf{U}_{(i)}^\nu-\mathbf{U}_{(i)}^{\nu-1}||_F\;||\mathbf{X}_i^{\nu+1}-\mathbf{X}_i^\nu||_F\\ &\le p_X^\nu||\mathbf{X}^\nu-\mathbf{X}^{\nu-1}||_F\;||\mathbf{X}^{\nu+1}-\mathbf{X}^\nu||_F+q_X^\nu||\mathbf{U}^\nu-\mathbf{U}^{\nu-1}||_F\;||\mathbf{X}^{\nu+1}-\mathbf{X}^\nu||_F,\end{aligned}$$

where in the last inequality we used $\sum_i a_ib_i\le||\mathbf{a}||\cdot||\mathbf{b}||$. Therefore,

$$||\mathbf{X}^{\nu+1}-\mathbf{X}^\nu||_F\le p_X^\nu||\mathbf{X}^\nu-\mathbf{X}^{\nu-1}||_F+q_X^\nu||\mathbf{U}^\nu-\mathbf{U}^{\nu-1}||_F.\tag{129}$$

If, in addition, Assumption D2 holds, then it follows from (128) that there exists a sufficiently large $\nu_X>0$ such that $p_X^\nu\in(0,\delta_0)$, for all $\nu\ge\nu_X$ and some $\delta_0\in(0,1)$. $\square$

References

University of Southern California, Signal and image processing institute. Volume 3: Miscellaneous image database, 1997. Available online: `http://sipi.usc.edu/database/database.php?volume=misc`.

M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11): 4311–4322, November 2006.

A. Aravkin, S. Becker, V. Cevher, and P. Olsen. A variational approach to stable principal component pursuit. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, UAI'14, pages 32–41, Arlington, Virginia, United States, 2014. AUAI Press. ISBN 978-0-9749039-1-0.

D. P. Bertsekas and J. N. Tsitsiklis. *Parallel and distributed computation: numerical methods*. Athena Scientific, 1997.

P. Bianchi and J. Jakubowicz. Convergence of a multi-agent projected stochastic gradient algorithm for non-convex optimization. *IEEE Transactions on Automatic Control*, 58(2): 391–405, February 2013.

T. Bouwmans, A. Sobral, S. Javed, S. K. Jung, and E. Zahzah. Decomposition into low-rank plus additive matrices for background/foreground separation: A review for a comparative evaluation with a large-scale dataset. *Computer Science Review*, 23:1–71, February 2017.

E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):1–37, June 2011.

P. Chainais and C. Richard. Learning a common dictionary over a sensor network. In *Proceedings of the 2013 5th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 133–136, Saint Martin, French West Indies, France, December 2013.

J. Chen, Z. J. Towfic, and A. H. Sayed. Dictionary learning over distributed models. *IEEE Transactions on Signal Processing*, 63(4):1001–1016, February 2015.

S. Chouvardas, Y. Kopsinis, and S. Theodoridis. An online algorithm for distributed dictionary learning. In *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3292–3296, Brisbane, Queensland, Australia, April 2015.

A. Daneshmand, F. Facchinei, V. Kungurtsev, and G. Scutari. Hybrid random/deterministic parallel algorithms for convex and nonconvex big data optimization. *IEEE Transactions on Signal Processing*, 63(13):3914–3929, August 2015.

A. Daneshmand, G. Scutari, and F. Facchinei. Distributed dictionary learning. In *Proceedings of the 50th Asilomar Conference on Signals, Systems and Computers*, pages 1001–1005, November 2016.

P. Di Lorenzo and G. Scutari. NEXT: In-network nonconvex optimization. *IEEE Transactions on Signal and Information Processing over Networks*, 2(2):120–136, June 2016.

M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 15(12):3736–3745, December 2006.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2): 303–338, June 2010.

F. Facchinei, G. Scutari, and Simone Sagratella. Parallel selective algorithms for nonconvex big data optimization. *IEEE Transactions on Signal Processing*, 63(7):1874–1889, April 2015.

T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity.* CRC Press, Taylor & Francis Group, 2015.

M. Hong, D. Hajinezhad, and M. M. Zhao. Prox-PDA: The proximal primal-dual algorithm for fast distributed nonconvex optimization and learning over networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70, pages 1529–1538, Sydney, Australia, August 2017.

P. O. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469, December 2004.

Z. Jiang, Z. Lin, and L. S. Davis. Learning a discriminative dictionary for sparse coding via label consistent k-svd. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 1697–1704, Colorado Springs, CO, USA, June 2011.

D. Kempe, A. Dobra, and J. Gehrke. Gossip-based computation of aggregate information. In *Proceedins of the 44th Annual IEEE Symposium on Foundations of Computer Science*, pages 482–491, Cambridge, MA, USA, October 2003.

J. Kim. *High-radix Interconnection Networks.* PhD thesis, Stanford, CA, USA, 2008.

D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, October 1999.

M. Lee, H. Shen, J. Z. Huang, and J. S. Marron. Biclustering via sparse singular value decomposition. *Biometrics*, 66(4):1087–1095, December 2010.

J. Liang, M. Zhang, X. Zeng, and G. Yu. Distributed dictionary learning for sparse representation in sensor networks. *IEEE Transactions on Image Processing*, 23(6):2528–2541, June 2014.

S. C. Madeira and A. L. Oliveira. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(1): 24–45, January 2004.

J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Supervised dictionary learning. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pages 1033–1040, Lake Tahoe, NV, December 2008.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11:19–60, January 2010.

A. Nedić, A. Ozdaglar, and P. A. Parrilo. Constrained consensus and optimization in multi-agent networks. *IEEE Transactions on Automatic Control*, 55(4):922–938, April 2010.

H. Raja and W. U. Bajwa. Cloud K-SVD: Computing data-adaptive representations in the cloud. In *Proceedings of 51st Annual Allerton Conference*, pages 1474–1481, Allerton House, UIUC, Illinois, USA, October 2013.

M. Razaviyayn, M. Hong, Z. Luo, and J. Pang. Parallel successive convex approximation for nonsmooth nonconvex optimization. In *NIPS*, 2014a.

M. Razaviyayn, H. W. Tseng, and Z. Q. Luo. Dictionary learning for sparse representation: Complexity and algorithms. In *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5247–5251, Florence, Italy, May 2014b.

B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, August 2010.

R.T. Rockafellar and J.B. Wets. *Variational Analysis*. Springer, 1998.

G. Scutari and Y. Sun. Parallel and distributed successive convex approximation methods for big-data optimization. In F. Facchinei and J.-S. Pang, editors, *Multi-agent Optimization*, chapter 3, pages 141–308. Lecture Notes in Mathematics 2224, Springer, 2018.

G. Scutari and Y. Sun. Distributed nonconvex constrained optimization over time-varying digraphs. *Mathematical Programming*, to appear 2019.

G. Scutari, F. Facchinei, P. Song, D. P. Palomar, and J.-S. Pang. Decomposition by partial linearization: Parallel optimization of multiuser systems. *IEEE Transaction on Signal Processing*, 62:641–656, February 2014.

H. Shen and J. Z. Huang. Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, 6(99):1015–1034, July 2008.

N. Srebro and A. Shraibman. Rank, trace-norm and max-norm. In *Proceedings of the Learning Theory: 18th Annual Conference on Learning Theory (COLT)*, pages 545–560, Bertinoro, Italy, June 2005.

Y. Sun, G. Scutari, and D. Palomar. Distributed nonconvex multiagent optimization over time-varying networks. In *Proceedings of the 2016 50th Asilomar Conference on Signals, Systems and Computers*, pages 788–794, November 2016.

K. K. Sung. *Learning and Example Selection for Object and Pattern Recognition.* PhD thesis, Artificial Intelligence Laboratory and Center for Biological and Computational Learning, MIT, Cambridge, MA, 1996.

T. Tatarenko and B. Touri. Non-convex distributed optimization. *IEEE Trans. on Automatic Control*, 62:3744–3757, August 2017.

I. Tosic and P. Frossard. Dictionary learning. *IEEE Signal Processing Magazine*, 28(2): 27–38, March 2011.

M. Udell, C. Horn, R. Zadeh, and S. Boyd. Generalized low rank models. *Foundations and Trends in Machine Learning*, 9(1):1–118, June 2016.

H. T. Wai, T. H. Chang, and A. Scaglione. A consensus-based decentralized algorithm for non-convex optimization with application to dictionary learning. In *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3546–3550, Brisbane, Queensland, Australia, April 2015.

H. T. Wai, J. Lafond, A. Scaglione, and E. Moulines. Decentralized frankwolfe algorithm for convex and nonconvex problems. *IEEE Transaction on Automatic Control*, 62:5522–5537, November 2017.

L. Xiao, S. Boyd, and S. Lall. A scheme for robust distributed sensor fusion based on average consensus. In *Proceedings of the 4th international symposium on Information processing in sensor networks*, pages 63–70, Los Angeles, CA, April 2005.

L. Xiao, S. Boyd, and S. J. Kim. Distributed average consensus with least-mean-square deviation. *Journal of Parallel and Distributed Computing*, 67(1):33–46, January 2007.

M. Zhao, Q. Shi, and M. Hong. A distributed algorithm for dictionary learning over networks. In *Proceedings of 2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 505–509, December 2016.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, March 2005.