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, and we establish its asymptotic convergence to stationary solutions. 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 $D \in \mathbb{R}^{M \times K}$ (a.k.a. the dictionary), by which the data matrix $S \in \mathbb{R}^{M \times N}$ can be represented through a matrix $X \in \mathbb{R}^{K \times N}$, with a favorable structure on $D$ and $X$ (e.g., sparsity). We target scenarios where computational resources and data are not centrally available, but distributed over a group of $I$ agents,
 Each agent $i \in \{1, 2, \ldots, I\}$ owns one block $S_i \in \mathbb{R}^{M \times n_i}$ of the data $S \triangleq [S_1, \ldots, S_I]$, where $\sum_{i=1}^I n_i = N$. Partitioning the representation matrix $X \triangleq [X_1, \ldots, X_I]$ according to $S$, with $X_i \in \mathbb{R}^{K \times n_i}$, the class of distributed DL problems we aim at studying reads

$$
\min_{D, (X_i)_{i=1}^I} \sum_{i=1}^I \left[ f_i(D, X_i) + g_i(X_i) \right] + G(D)
$$

(P)

where $f_i : D \times \mathcal{X}_i \rightarrow \mathbb{R}$ is the fidelity function of agent $i$, which measures the mismatch between the data $S_i$ and the (local) model; this function is assumed to be smooth and biconvex (i.e., convex in $D$ for fixed $X_i$, and vice versa); $g_i : \mathcal{X}_i \rightarrow \mathbb{R}$ are (possibly non-smooth) convex functions, which are generally used to impose extra structure on the solution (e.g., low-rank or sparsity); and $D \subseteq \mathbb{R}^{M \times K}$ and $\mathcal{X}_i \subseteq \mathbb{R}^{K \times n_i}$ are some closed convex sets. To avoid scaling ambiguity in the model, $D$ is assumed to be bounded, without loss of generality. Since all $f_i$’s share the common variable $D$, we call it a shared variable and, by the same token, $X_i$’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.
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 $S_i$ and thus can only compute its own function $f_i$;

(iii) Each $f_i$ depends on a common set of variables—the dictionary $D$—shared among all the agents, as well as the private variables $X_i$. 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 and the Hessian of each $f_i$ are in general unbounded on the feasible region. This represents a challenge in the design of provably convergent distributed algorithms, as boundedness of the gradient is a standard assumption in the analysis of current distributed schemes for nonconvex problems;

(v) $G$ and $g_i$’s are nonsmooth;

(vi) The graph is directed and time-varying, with no specific structure.

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., 2014). However, we are not aware of any distributed algorithm that can address even just interesting subsets of challenges i)-vi), as documented next.

Ad-hoc heuristics: Several attempts have been made to extend centralized approaches 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).

**Distributed nonconvex optimization:** Since the DL problem $P$ is an instance of nonconvex optimization problems, one could hope to leverage some of the few recent distributed methods for general non-convex optimization (Bianchi and Jakubowicz, 2013; Tatarenko and Touri, 2016; Di Lorenzo and Scutari, 2016; Sun et al., 2016; Hong, 2016). However, none of these schemes can deal with (even a subset of) challenges i)-vi). In fact, (Bianchi and Jakubowicz, 2013) can handle only smooth objective functions over undirected static graphs; the push-sum gradient algorithm with diminishing step-size, studied in (Tatarenko and Touri, 2016), can be applied to nonconvex smooth unconstrained problems and time-varying networks; and the dual-based methods in (Hong, 2016) can handle only a very special class of nonconvex smooth functions $f_i$, it cannot deal with constraints, it requires detailed spectral information about the network to set the step-size (proximal constants), and it can be applied only to undirected static graphs. The first provably convergent distributed scheme for a special instances of $P$, with $G \neq 0$, constraints $\mathcal{D}$, and over time-varying digraphs, is NEXT, proposed in our previous work (Di Lorenzo and Scutari, 2016). However, the consensus protocol employed therein uses doubly-stochastic weight matrices, and thus is limited to very special digraphs. This assumption has been removed in (Daneshmand et al., 2017) and (Sun et al., 2016). In addition to the limitation mentioned above, all the aforementioned algorithms proposed for general non-convex optimization require that the (sub)gradient of the objective function is bounded on the feasible set of the problem while the gradient of the smooth part must be globally Lipschitz continuous. Furthermore, they cannot adequately deal with private and shared variables, which are fundamentally important in our context.

**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 novel 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 arbitrary time-varying undirected or directed graphs; in the latter case only column-stochasticity of the weight matrices is required. We remark that the consensus mechanism employed in this paper differs from push-sum protocols proposed in the literature to deal with directed graphs (Kempe et al., 2003), because problem $P$ is constrained and current push-sum-based updates do not preserve feasibility of the iterates. Convergence to stationary solutions of Problem $P$ is established, under mild assumptions on the step-size employed by the algorithm. On the technical side, we contribute to the literature of distributed algorithms for nonconvex nonsmooth constrained optimization by putting forth a new non-trivial analysis that, for the first time, avoids any boundedness assumption on the gradients or Hessians of the $f_i$’s. Extensive numerical results 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. 4 whereas Sec. 5 draws some conclusions.

1.4 Notation

Throughout the paper we use the following notation. Vectors are denoted by bold lowercase letters (e.g., $\mathbf{x}$) whereas matrices are denoted by bold capital letters (e.g., $\mathbf{X}$). The $k$-th canonical vector is denoted by $\mathbf{e}_k$. The inner product between two real matrices, $\mathbf{X}$ and $\mathbf{Y}$, is denoted by $\langle \mathbf{X}, \mathbf{Y} \rangle \triangleq \text{tr}(\mathbf{X}^\top \mathbf{Y})$, where $\text{tr}()$ is the trace operator. Given the real matrix $\mathbf{X}$, with $ij$-entries denoted by $X_{ij}$, we will use the following matrix norms: the Frobenius norm $||X||_F \triangleq \sqrt{\sum_{i,j} |X_{ij}|^2}$; the $L_{1,1}$ norm $||X||_{1,1} \triangleq \sum_{i,j} |X_{ij}|$; the $L_{2,\infty}$ norm $||X||_{2,\infty} \triangleq \max_i \sqrt{\sum_j X_{ij}^2}$; the $L_{\infty,\infty}$ norm $||X||_{\infty,\infty} = \max_{i,j} |X_{ij}|$; and the spectral norm $||X||_2 \triangleq \sigma_{\max}(X)$, where $\sigma_{\max}(X)$ denotes the maximum singular value of $X$. The matrix quantities $\nabla_D f_i(D, X_i)$ and $\nabla_{X_i} f_i(D, X_i)$ are the gradients of $f_i$ with respect to $D$ and $X_i$, respectively, with the partial derivatives arranged according to the patterns of $D$ and $X_i$, respectively. The same convention is adopted for subgradients of $g_i$ and $G$, that are therefore written as matrices of the same dimensions of $X_i$ and $D$, respectively. With this convention, it is easy to see that a feasible tuple $(\tilde{D}, \tilde{X})$, with $\tilde{X} \triangleq [\tilde{X}_1, \ldots, \tilde{X}_I]$ is a stationarity solution of Problem P if the following holds:

$$\left\langle \nabla_D F(\tilde{D}, \tilde{X}) + \Omega, D - \tilde{D} \right\rangle \geq 0, \quad \forall D \in \mathcal{D},$$

$$\left\langle \nabla_{X_i} f_i(\tilde{D}, \tilde{X}_i) + \Xi_i, X_i - \tilde{X}_i \right\rangle \geq 0, \quad \forall X_i \in \mathcal{X}_i, \ i = 1, \ldots, I,$$

for some $\Omega \in \partial G(\tilde{D})$ and $\Xi_i \in \partial g_i(\tilde{X}_i)$, and $i = 1, \ldots, I$.

2. Problem Setup and Motivating Examples

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

2.1 Problem Assumptions

We consider Problem P under the following assumptions.

Assumption A (On Problem P)

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

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

(A3) $\mathcal{D}$ is compact and convex; and each $\mathcal{X}_i$ is closed and convex (not necessarily bounded);
Figure 2: Illustration of in-neighborhood set of agent $i$ at time $\nu$.

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

(A5) Either i) $\mathcal{X}_i$ is bounded and $g_i : \mathcal{O}_i \to \mathbb{R}$ is convex; or ii) $g_i$ is $\mu_i$-strongly convex.

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

Network topology

We study Problem P under the following network setting. Time is slotted and in each time-slot $\nu$ the network of the $I$ agents is modeled as a digraph $G^\nu = (\mathcal{V}, \mathcal{E}^\nu)$, where the set of vertices $\mathcal{V} = \{1, \ldots, I\}$ represents the set of agents, and the set of edges $\mathcal{E}^\nu \triangleq \{(i, j) : \text{agent } j \text{ can receive information from agent } i \text{ at time slot } \nu\}$ represents the (possibly) time-varying directed communication links. The in-neighborhood of agent $i \in \mathcal{V}$ at time $\nu$ 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 $d_{i}^{\nu} \triangleq |\mathcal{N}_{i}^{\text{out}}[\nu]|$, where $|\bullet|$ denotes the cardinality of a set. To let information propagate over the network, we assume that the sequence $\{G^\nu\}_\nu$ possesses some “long-term” connectivity property, as stated next.

Assumption B (B-strong connectivity) The graph sequence $\{G^\nu\}_\nu$ 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 $\bigcup_{t=kB}^{(k+1)B-1} \mathcal{E}^t$ is strongly connected, for all $k \geq 0$.

To the best of our knowledge, this is the weakest connectivity condition used in the literature to analyze distributed algorithms on time-varying networks; it simply states that information has the possibility to propagate from every node $i$ to every other node in the network. Assumption B is quite mild and 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 therefore 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

\[
\min_{D \in \mathcal{D}, X_i \in \mathbb{R}^{K \times n_i}} \sum_{i=1}^{I} \left\{ \frac{1}{2} \| S_i - DX_i \|_F^2 + \lambda \| X_i \|_{1,1} + \frac{\mu}{2} \| X_i \|_F^2 \right\} \\
\text{s.t.} \quad D \in \mathcal{D}, \quad X_i \in \mathbb{R}^{K \times n_i}, \quad i = 1, 2, \ldots, I,
\]

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

Supervised DL (Mairal et al., 2008)

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

\[
\min_{\substack{D^{(1)}, D^{(2)}}} \sum_{i=1}^{I} \sum_{n \in S_i} \left[ \ell \left( y_n \zeta_n \left( D^{(2)}, x_n, s_n \right) \right) + \frac{1}{2} \left\| s_n - D^{(1)} x_n \right\|_2^2 + g_n(x_n) \right] \\
\text{s.t.} \quad D^{(1)} \in \mathcal{D}^{(1)}, \quad D^{(2)} \in \mathcal{D}^{(2)}, \quad x_n \in \mathbb{R}^K, \quad n = 1, 2, \ldots, N,
\]

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

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 \( S \approx L + HQU \): the data matrix \( S \) is decomposed as the superposition of a low-rank matrix \( L \) (capturing the correlations among data)
and HQU, where \( Q \in \mathbb{R}^{M \times K} \) is an over-complete dictionary (capturing the representative modes of the data), \( U \in \mathbb{R}^{K \times N} \) is a sparse matrix (representing the data parsimoniously), and \( H \) is a given degradation matrix, which accounts for tasks such as denoising, superresolution, and deblurring. To enforce \( L \) to be low-rank, we employ the nuclear norm \( \|L\|_* \) regularizer, which can be equivalently rewritten as \( \|L\|_* = \inf \{ \frac{1}{2} \|P\|_F^2 + \frac{\zeta}{2} \|V\|_F^2 : L = PV \} \), where \( P \in \mathbb{R}^{M \times L}, V \in \mathbb{R}^{L \times N} \), and \( L \ll \min(M, N) \) (Srebro and Shraibman, 2005; Recht et al., 2010). Partitioning \( V \) and \( U \) according to \( S \), i.e., \( S_i = PV_i + HQU_i \), the problem reads

\[
\min_{P, Q, (V_i, U_i)_{i=1}^I} \sum_{i=1}^I \left[ \frac{1}{2} \|S_i - [P\; HQ] [V_i\; U_i]\|_F^2 + \frac{\zeta}{2} \|P\|_F^2 + I \cdot \|V_i\|_F^2 \right] + \lambda \|X_i\|_{1,1} + \frac{\mu}{2} \|X_i\|_F^2 \tag{3}
\]

subject to \( D \in D, \quad X_i \in \mathbb{R}^{(L+K) \times n_i}, \quad i = 1, 2, \ldots, I \),

where \( D \) is some compact set; \( \zeta > 0 \) is a constant used to promote the low-rank structure on \( L \) while sparsity on \( X \) is enforced by the elastic net regularization, with constants \( \lambda, \mu > 0 \). Problem (3) is clearly an instance of Problem P where \( f_i \) is the quadratic loss, and \([P, Q]\) and \([V_i^T, U_i^T]\) are the shared and private variables (\( K = L+K \)), respectively. Assumption A is satisfied, and the Lipschitz constant in A2 is given by \( L_{\nabla X_i}(D) = (\sigma_{\text{max}}(PHQ))^2 \).

A variant of this problem, which still is a particular case of Problem (3), is obtained by replacing the quadratic loss function with the smoothed Huber function to achieve robustness against outliers.

**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 \( D \) and \( X \) of \( S \). More specifically, we have

\[
\min_{D, (X_i)_{i=1}^I} \sum_{i=1}^I \left\{ \frac{1}{2} \|S_i - DX_i\|_F^2 + \lambda_X \|X_i\|_{1,1} + \frac{\mu_X}{2} \|X_i\|_F^2 \right\} + \lambda_D \|D\|_{1,1} + \frac{\mu_D}{2} \|D\|_F^2 \tag{4}
\]

subject to \( D \in D \triangleq \{ D \in \mathbb{R}^{M \times K} : \|D\|_{2,\infty} \leq \alpha \}, \quad X_i \in \mathbb{R}^{K \times n_i}, \quad i = 1, 2, \ldots, I \),

where \( \lambda_D, \lambda_X, \mu_D, \mu_X, \alpha > 0 \) are given constants. Problem (4) is an instance of P, with \( f_i(D, X_i) = (1/2) \cdot \|S_i - DX_i\|_F^2; \quad G(D) = \lambda_D \|D\|_{1,1} + (\mu_D/2) \cdot \|D\|_F^2, \quad \text{and} \quad g_i(X_i) = \lambda_X \|X_i\|_{1,1} + (\mu_X/2) \cdot \|X_i\|_F^2 \). 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 \( L_{\nabla X_i}(D) = (\sigma_{\text{max}}(D))^2 \).

**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 $D$ and $X$. 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 $X$. The problem reads

$$\min_{D,(X_i)_{i=1}^I} \sum_{i=1}^I \left\{ \frac{1}{2} \|S_i - DX_i\|_F^2 + \lambda \|X_i\|_{1,1} + \frac{\mu}{2} \|X_i\|_F^2 \right\}$$

s.t. $D \in \mathcal{D} \triangleq \{D \in \mathbb{R}_+^{M \times K} \mid \|D\|_{2,\infty} \leq \alpha \}$, $X_i \in \mathbb{R}_+^{K \times n_i}$, $i = 1, 2, \ldots, I$,

for some $\lambda, \mu, \alpha > 0$. Problem (5) is another instance of $P$, with $f_i(D, X_i) = (1/2) \cdot \|S_i - DX_i\|_F^2$, $g_i(X_i) = \lambda \|X_i\|_{1,1} + (\mu/2) \cdot \|X_i\|_F^2$, $G(D) = 0$, $\mathcal{D} = \{D \in \mathbb{R}_+^{M \times K} \mid \|D\|_{2,\infty} \leq \alpha \}$, and $X_i = \mathbb{R}_+^{K \times n_i}$. Assumption A is satisfied, and the Lipschitz constant in A2 is given by $L_{\nabla X_i}(D) = (\sigma_{\text{max}}(D))^2$.

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 $i$ maintains a local copy of the shared variables $D$, denoted by $D_{(i)}$, and controls its private variable $X_i$. The optimization variables $(D_{(i)}, X_i)$ need to be updated so that asymptotically: i) all $D_{(i)}$’s reach a consensus, i.e., $D_{(i)} = D_{(j)}, \forall i \neq j$; ii) and each tuple $(D_{(i)}, (X_j)_{j=1}^I)$ is a stationary solutions of Problem P. To achieve these goals, agents face two main challenges: the non-convexity of $P$ and the lack of global knowledge of the $f_i$’s. We deal with these issues by leveraging Successive Convex Approximation techniques (Step 1 below) and a novel consensus/tracking protocol (Step 2), as described next.

Step 1: Local Optimization

To update $(D_{(i)}, X_i)$, agent $i$ should iteratively solve Problem $P$. However $f_i$ is not convex in $(D_{(i)}, X_i)$, and $\sum_{j \neq i} f_j$ is unknown. The idea is then to somehow “locally approximate” $P$ so that each agent can compute the new updates, locally and efficiently.

Since $f_i$ is bi-convex in $(D_{(i)}, X_i)$, a natural approach is to update $D_{(i)}$ and $X_i$ in an alternating fashion. Specifically, at iteration $\nu$, agent $i$ fixes $X_i = X_i^\nu$ and computes $D_{(i)}$ by solving the following strongly convex problem:

$$\hat{D}_{(i)}^{\nu} \triangleq \arg\min_{D_{(i)} \in \mathcal{D}} \tilde{f}_i\left(D_{(i)}; D_{(i)}^{\nu}, X_i^\nu\right) + \sum_{j \neq i} \nabla_{D_j} \nu_j(D_{(i)}, X_j^\nu, D_{(i)} - D_{(i)}^{\nu}) + G(D_{(i)})$$

(6)

where $\tilde{f}_i(\bullet; D_{(i)}^{\nu}, X_i^\nu)$ is a “suitably” chosen strongly convex approximation of $f_i(\bullet, X_i^\nu)$ at the current iterate $(D_{(i)}^{\nu}, X_i^\nu)$. Since $f_i$ is convex in $D_{(i)}$, a natural choice for the surrogate $\tilde{f}_i$ is

$$\tilde{f}_i(D_{(i)}; D_{(i)}^{\nu}, X_i^\nu) = f_i(D_{(i)}, X_i^\nu) + \frac{\tau_{D_{(i)}}^{\nu}}{2} \|D_{(i)} - D_{(i)}^{\nu}\|_F^2$$

(7)
where the quadratic term, with $\tau_{D,i}^\nu > 0$, serves the purpose of making $\tilde{f}_i$ strongly convex. However, the computation of $\tilde{D}^\nu_{(i)}$ has a severe drawback: the evaluation of $\Pi^\nu_i$ in (6) would require the knowledge of $\nabla_D f_j(D^\nu_{(i)}, X_j^\nu)$ for all $j \neq i$, which is not available to agent $i$. To cope with this issue, we replace $\Pi^\nu_i$ in (6) with a “local estimate”, denoted by $\tilde{\Pi}^\nu_i$, and solve instead,

$$
\tilde{D}^\nu_{(i)} \triangleq \arg\min_{D_{(i)} \in D} \tilde{f}_i(D_{(i)}; D^\nu_{(i)}, X^\nu_i) + \left\langle \tilde{\Pi}^\nu_i, D_{(i)} - D^\nu_{(i)} \right\rangle + G(D_{(i)}).
$$

(8)

We will show how to update $\tilde{\Pi}^\nu_i$ using only local information (Step 2 below), so that $\|\tilde{\Pi}^\nu_i - \Pi^\nu_i\|_F \rightarrow 0$.

Given $\tilde{D}^\nu_{(i)}$, a step-size is introduced in the update of $D_{(i)}$ from agent $i$ and an intermediate step is taken by defining a matrix $U^\nu_{(i)}$:

$$
U^\nu_{(i)} = D^\nu_{(i)} + \gamma^\nu(\tilde{D}^\nu_{(i)} - D^\nu_{(i)}),
$$

(9)

where $\gamma^\nu$ is a positive scalar to be properly chosen (see Assumption D1). As it will be clear later, the step-size is instrumental to reach convergence and enforce asymptotic consensus among the local copies $\tilde{D}^\nu_{(i)}$.

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

$$
X^\nu_{i+1} \triangleq \arg\min_{X_i \in X_i} \tilde{h}_i(X_i; U^\nu_{(i)}, X^\nu_i) + g_i(X_i),
$$

(10)

where

$$
\tilde{h}_i(X_i; U^\nu_{(i)}, X^\nu_i) \triangleq f_i(U^\nu_{(i)}, X_i) + \frac{\tau^\nu_{X,i}}{2} \|X_i - X^\nu_i\|_F^2,
$$

(11)

and $\tau^\nu_{X,i}$ is a positive scalar (to be properly chosen together with $\tau^\nu_{D,i}$, see Assumption D).

**Step 2: Local Communications**

Let us design now a local communication mechanism ensuring asymptotic consensus over the local copies $D_{(i)}$’s and tracking of each $\Pi^\nu_i$ by $\tilde{\Pi}^\nu_i$. A common approach used in the literature—see, e.g., (Dimakis et al., 2010; Sayed, 2014)—is to let the agents broadcast their local copies $U_{(i)}$ and then combine the variables received from their neighbors according to the following consensus-based rule:

$$
D^\nu_{(i)} = \sum_{j \in N^0_{(i)}[\nu]} \omega^\nu_{ij} U^\nu_{(j)},
$$

(12)

where $\omega^\nu_{ij}$s are some weights “matching” the graph $G^\nu$, in the sense that $\omega^\nu_{ij} > 0$ if $(j, i) \in E^\nu$; and $\omega^\nu_{ij} = 0$ otherwise. We denote by $W^\nu$ the matrix of all these weights, with $[W^\nu]_{i,j} = \omega^\nu_{ij}$. To ensure convergence and consensus, a widely used property on $W^\nu$ is that it is doubly-stochastic (Nedić et al., 2010; Di Lorenzo and Scutari, 2016), i.e., $1^T W^\nu = 1^T$ and $W^\nu 1 = 1$, for all $\nu$. Furthermore, under row stochasticity, (12) becomes a convex combination of
the $U^{\nu}_{(j)}$'s, and thus the update (12) preserves feasibility of the new iterate. However, a directed graph may not admit a doubly-stochastic weight matrix; some form of balancedness of the graph is needed (Gharesifard and Cortés, 2010), which limits the class of network topologies. Second, necessary and sufficient conditions for a digraph to admit a doubly-stochastic weight matrix $W^{\nu}$ cannot be easily checked. Third, constructing a doubly-stochastic weight matrix matching the graph, even when possible, calls for computationally intense, generally centralized, algorithms. Therefore, the goal is to modify (12) in such a way so as to guarantee both asymptotic consensus and feasibility of the iterates without requiring $W^{\nu}$ to be doubly-stochastic.

For this purpose, we build on the communication protocol introduced in (Sun et al., 2016): each agent $i$ updates its own local estimate $D^{\nu}_{(i)}$ in (12) with

$$w^{\nu}_{ij} = \frac{a^{\nu}_{ij} \phi^{\nu}_j}{\phi^{\nu+1}_i},$$

where $a^{\nu}_{ij}$'s are some weights (to be properly chosen) matching the graph $G^{\nu}$, and $\phi^{\nu}_i$ is an extra scalar variable that is updated according to

$$\phi^{\nu+1}_i = \sum_{j \in \mathcal{N}_i^{in}[\nu]} a^{\nu}_{ij} \phi^{\nu}_j.$$  

The purpose of the $\phi$--variables is to dynamically build row-stochasticity in the update (12) using only local information, which is needed to preserve feasibility of the $D^{\nu}_{(i)}$'s, and to further guarantee asymptotic agreement among the agents. Substituting (13) into (12), leads to the final update of $D^{\nu}_{(i)}$:

$$D^{\nu+1}_{(i)} = \frac{1}{\phi^{\nu+1}_i} \sum_{j \in \mathcal{N}_i^{in}[\nu]} a^{\nu}_{ij} \phi^{\nu}_j U^{\nu}_{(j)}.$$ 

where $\phi^0_i = 1$ for all $i$. We denote by $A^{\nu}$ the matrix whose entries are the weights $a^{\nu}_{ij}$'s, i.e., $[A^{\nu}]_{i,j} = a^{\nu}_{ij}$. This matrix is chosen so that the following conditions are satisfied.

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

1. **(C1)** $a^{\nu}_{ii} \geq \kappa > 0$ for all $i = 1, \ldots, I$;
2. **(C2)** $a^{\nu}_{ij} \geq \kappa > 0$, if $(j, i) \in \mathcal{E}^{\nu}$; and $a^{\nu}_{ij} = 0$ otherwise;
3. **(C3)** $A^{\nu}$ is column stochastic, i.e., $1^T A^{\nu} = 1^T$.

Note that $A^{\nu}$ is not required to be doubly-stochastic, but only column-stochastic and this can easily be guaranteed by each agent independently. We discuss some practical rules satisfying Assumption C in Sec. 3.1.

The updates in (14) and (15) are simple to implement: all agents only need to (i) send their local variable $U^{\nu}_{(j)}$ and the scalar weight $a^{\nu}_{ij} \phi^{\nu}_j$ to their neighbors; and (ii) collect locally the information coming from the neighbors.
A similar scheme can be put forth to update $\tilde{\Pi}^\nu_i$'s in (8). Let us first rewrite $\Pi^\nu_i$ [defined in (6)] as

$$
\Pi^\nu_i = I \cdot \left( \frac{1}{T} \sum_{j=1}^{T} \nabla_D f_j(D^\nu_i, X^\nu_j) \right) - \nabla_D f_i(D^\nu_i, X^\nu_i). 
$$

(16)

Building on the gradient tracking mechanism (Di Lorenzo and Scutari, 2016), and leveraging the introduced communication protocol in (14)-(15), we propose to compute $\tilde{\Pi}^\nu_i$ mimicking (16), i.e.,

$$
\tilde{\Pi}^\nu_i + 1 = I \cdot \tilde{\Theta}^\nu_{(i)} - \nabla_D f_i(D^\nu_i + 1, X^\nu_i + 1), 
$$

(17)

where $\tilde{\Theta}^\nu_{(i)}$ is an extra (matrix) variable maintained by agent $i$, and instrumental to track $\Theta^\nu_i$ in (16). To this end, we update $\tilde{\Theta}^\nu_{(i)}$ according to the following dynamic consensus averaging rule:

$$
\tilde{\Theta}^\nu_{(i)} = \frac{1}{\phi^\nu_{(i)} + 1} \sum_{j \in \mathcal{N}_a[u]} a^\nu_{ij} \phi^\nu_{(j)} \tilde{\Theta}^\nu_{(j)} + \frac{1}{\phi^\nu_{(i)} + 1} \left( \nabla_D f_i(D^\nu_{(i)} + 1, X^\nu_{(i)} + 1) - \nabla_D f_i(D^\nu_{(i)}, X^\nu_{(i)}) \right), 
$$

(18)

with $\tilde{\Theta}_0 = \nabla_D f_i(D_0, X_0)$. The update (18) follows similar logic as the update of $D^\nu_i$ in (15) to ensure asymptotic consensus, but with the difference that it also includes a correction term [the second term in RHS of (18)], which helps $\tilde{\Theta}^\nu_i$ to asymptotically track the average value $\Theta^\nu_i$. Note that the update of $\tilde{\Theta}_i$ and $\tilde{\Pi}_i$ can be performed locally by agent $i$, following the same procedure as described for (14)-(15).

Combining the above steps, we can now formally introduce the proposed distributed algorithm for the DL problems $P$, as described in Algorithm 1; we call this the $D^4L$ (Decentralized Dictionary Learning over Dynamic Digraphs) Algorithm. We discuss next the key properties of the $D^4L$ Algorithm along with its convergence properties.

### 3.1 Discussion

Before stating the main convergence result for the $D^4L$ Algorithm, we discuss how to choose the free parameters of the algorithm, namely: the surrogate functions $\tilde{f}_i$ and $\tilde{h}_i$, the consensus weights $(a_{ij}^\nu)_{i,j=1}^{I}$, the step-size $\gamma^\nu$, and the coefficients $(\tau_{i}^\nu X_{i})_{i=1}^{I}$ and $(\tau_{D, i}^\nu)_{i=1}^{I}$.

**On the choice of $\tilde{f}_i$ and $\tilde{h}_i$.** The non-smooth strongly convex subproblems (8) and (10) can be solved using standard solvers, e.g., projected subgradient methods. When dealing with large-scale instances of such problems, effective methods are also (Facchinei et al., 2015; Daneshmand et al., 2015). However, when specific loss functions $f_i$, and penalty functions $G$ and $g_i$ are considered, appropriate choices for $\tilde{f}_i$ and $\tilde{h}_i$ can be employed to abate the computational complexity of solving the resulting subproblems (8) and (10).

As a first observation, note that instead of using the surrogate $\tilde{f}_i$ and $\tilde{h}_i$ given in (7) and (11), respectively, one can always choose them as the linearization of $f_i$ (plus a proximal regularization), that is,

$$
\tilde{f}_i(D_{(i)}, \tilde{D}^\nu_{(i)}, X^\nu_{(i)}) = \left( \nabla_D f_i(D^\nu_{(i)}, X^\nu_{(i)}), D_{(i)} - \tilde{D}^\nu_{(i)} \right) + \frac{\tau_{D,i}^\nu}{2} \| D_{(i)} - \tilde{D}^\nu_{(i)} \|_F^2.
$$

(19)
Algorithm 1: Decentralized Dictionary Learning over Dynamic Digraphs (D4L)

Initialization: \( \phi_i^0 = 1, \ D_i^0 \in \mathcal{D}, \ X_i^0 \in X_i, \ \Theta_i^0 = \nabla_D f_i(D_i^0, X_i^0), \)
\( \Pi_i^0 = I \cdot \Theta_i^0 - \nabla_D f_i(D_i^0, X_i^0), \) for all \( i = 1, 2, \ldots, I; \) and set \( \nu = 0; \)

S1. If \( (D_i^\nu, X_i^\nu) \) satisfies stopping criterion: STOP;

S2. Local Optimization: Each agent \( i \) computes:

(a) \( \widetilde{D}_i^\nu = \arg\min_{D_i \in \mathcal{D}} \hat{f}_i(D_i; D_i^\nu, X_i^\nu) + \langle \Pi_i, D_i - D_i^\nu \rangle + G(D_i); \)

(b) \( U_i^\nu = D_i^\nu + \gamma^\nu(\widetilde{D}_i - D_i^\nu); \)

(c) \( X_i^{\nu+1} = \arg\min_{X_i \in X_i} \tilde{h}_i(X_i; U_i^\nu, X_i^\nu) + g_i(X_i); \)

S3. Local Communications: Each agent \( i \) collects data from the neighbors and updates:

(a) \( \phi_i^{\nu+1} = \sum_{j \in \mathcal{N}_i^\nu} a_{ij}^\nu \phi_j^\nu; \)

(b) \( D_i^{\nu+1} = \frac{1}{\phi_i^\nu+1} \sum_{j \in \mathcal{N}_i^\nu} a_{ij}^\nu \phi_j^\nu U_j^\nu; \)

(c) \( \Theta_i^{\nu+1} = \frac{1}{\phi_i^\nu+1} \sum_{j \in \mathcal{N}_i^\nu} a_{ij}^\nu \phi_j^\nu \Theta_j^\nu + \frac{1}{\phi_i^\nu+1} \left( \nabla_D f_i(D_i^{\nu+1}, X_i^{\nu+1}) - \nabla_D f_i(D_i^\nu, X_i^\nu) \right); \)

(d) \( \Pi_i^{\nu+1} = I \cdot \Theta_i^{\nu+1} - \nabla_D f_i(D_i^{\nu+1}, X_i^{\nu+1}); \)

S4. Set \( \nu + 1 \rightarrow \nu, \) and go to S1.

and
\[
\tilde{h}_i(X_i; U_i^\nu, X_i^\nu) = \left\langle \nabla X_i f_i(U_i^\nu, X_i^\nu), X_i - X_i^\nu \right\rangle + \frac{\tau_X^i}{2} \|X_i - X_i^\nu\|^2_F. \tag{20}
\]

This usually leads to simpler subproblems. We elaborate on this point by considering as an example the elastic net sparse DL problem (1) in Sec. 2.2, where \( f_i(D, X_i) = \frac{1}{2} \|S_i - DX_i\|_F^2; \) \( G(D) = 0; \) and \( g_i(X_i) = \lambda \|X_i\|_1 + \frac{\mu}{2} \|X_i\|_F^2, \) with \( \lambda, \mu > 0. \) By using (19), the resulting subproblem (8) admits the following closed form solution:

\[
\bar{D}_i^\nu = P_D \left[ D_i^\nu - \frac{1}{\tau_{D,i}} \left( \nabla_D f_i(D_i^\nu, X_i^\nu) + \Pi_i^\nu \right) \right]. \tag{21}
\]

Referring to the sparse coding subproblem (10), if \( \tilde{h}_i \) is chosen according to (11), computing the update \( X_i^{\nu+1} \) results in solving a LASSO problem. If instead one uses the surrogate in (20), the solution of (10) can be computed in closed form as

\[
X_i^{\nu+1} = \frac{\tau_X^i}{\mu + \tau_X^i} T_{X,i} \left( X_i^\nu - \frac{1}{\tau_X^i} \nabla X_i f_i(U_i^\nu, X_i^\nu) \right), \tag{22}
\]
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 componentwise.

**On the choice of $\gamma^\nu$, $\tau^\nu_{X,i}$ and $\tau^\nu_{D,i}$.** While several choices are possible for the aforementioned quantities, the following minimal conditions need to be satisfied for convergence.

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

1. $\{\gamma^\nu\}_\nu$ satisfies: $\gamma^\nu \in (0, 1]$ for all $\nu$; $\sum_{\nu=0}^\infty \gamma^\nu = \infty$; and $\sum_{\nu=0}^\infty (\gamma^\nu)^2 < \infty$;
2. $(\tau^\nu_{D,i})_\nu$ and $(\tau^\nu_{X,i})_\nu$ satisfy

$$
0 < \inf_{\nu} \tau^\nu_{D,i} \leq \sup_{\nu} \tau^\nu_{D,i} < +\infty, \quad (23)
$$

and

$$
\sup_{\nu} \tau^\nu_{X,i} < +\infty, \quad \tau^\nu_{X,i} \geq \frac{1}{2} L_{\nabla X_i} (U^\nu_{(i)}) + \epsilon, \quad \forall \nu \geq 1, \quad (24)
$$

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.

3. Stronger convergence results [cf. Theorem 2] can be obtained if the sequences $(\tau^\nu_{D,i})_\nu$ and $(\tau^\nu_{X,i})_\nu$, in addition to D2, also satisfy

$$
\sum_{t=0}^\infty \left| \tau^\nu_{D,i}^{t+1} - \tau^\nu_{D,i}^t \right| < \infty, \quad (25)
$$

and

$$
\limsup_{\nu} \left| \tau^\nu_{X,i} - \tau^{\nu-1}_{X,i} \right| < \mu, \quad (26)
$$

where $\mu \triangleq \min_i \mu_i$ and $\mu_i$ is the strongly convexity constant of $f_i$ [cf. Assumption A5].

Roughly speaking, D2 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 D3 guarantees that these parameters are asymptotically “stable”. A trivial choice for $\tau^\nu_{D,i}$ satisfying both (23) and (25) is $\tau^\nu_{D,i} = c$, for some $c > 0$ while some practical rules for $\tau^\nu_{X,i}$ satisfying both (24) and (26) are the following:

(a) The most straightforward choice is to use a constant $\tau^\nu_{X,i}$, that is,

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

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

(b) A less conservative choice is to satisfy (24) iteratively, while guaranteeing that $\tau^\nu_{X,i}$ is uniformly positive:

$$
\tau^\nu_{X,i} = \max (L_{\nabla X_i} (U^\nu_{(i)}), \tilde{\epsilon}), \quad (27)
$$

where $\tilde{\epsilon}$ is any positive (possibly small) constant;
Decentralized Dictionary Learning Over Time-Varying Digraphs

(c) A generalization of (b) is
\[ \tau_{X,i} \in \left[ \max(L_{\nabla X_i}(U_{(i)}^\nu), \tilde{\epsilon}), L_{\nabla X_i}(U_{(i)}^\nu) + \tilde{\mu} \right]. \]

for some \( \tilde{\epsilon} \) and \( \tilde{\mu} \), such that \( 0 < \tilde{\epsilon} \leq \tilde{\mu} < \mu \).

Remark 1 The cases (a)-(c) above clearly satisfy (24), but also (26) is satisfied by continuity of \( L_{\nabla X_i}(\cdot) \) and Proposition 9 (cf. Sec. A.2).

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.

Several choices are possible for the step-size sequence \( \{\gamma^\nu\}_{\nu} \) satisfying the standard diminishing-rule D1; see, e.g., (Bertsekas and Tsitsiklis, 1997). Here, we only recall one rule that we found very effective in our experiments, namely (Facchinei et al., 2015):
\[ \gamma^\nu = \gamma^\nu - 1 \left( 1 - \epsilon_0 \gamma^\nu - 1 \right) \text{ with } \gamma_0 \in (0, 1] \text{ and } \epsilon_0 \in (0, 1 / \gamma_0). \]

On the choice of matrix \( A^\nu \). A valid choice of \( A^\nu \) satisfying Assumption B when the graph \( G^\nu \) is directed, is the following:
\[ a^\nu_{ij} = 1 / d^\nu_j \text{ if } j \in N^\nu_i, \text{ and } a^\nu_{ij} = 0 \text{ otherwise, where } d^\nu_j \text{ is the out-degree of agent } j \text{ at time } \nu. \]
The resulting communication protocol (14), (15) and (18) 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.

Note that the above choice of \( A^\nu \) is the same as that used in the push-sum protocol (Kempe et al., 2003). However, there is a fundamental difference between (Kempe et al., 2003) and the proposed consensus scheme: push-sum-like protocols do not preserve feasibility of the iterates, and thus cannot be readily used to solve distributed constrained optimization problems. In fact, the few works in the literature using push-sum in the context of distributed optimization [see (Nedić and Olshevsky, 2015) and references therein] all consider only unconstrained problems.

3.2 Convergence of Algorithm 1

We can now provide the main convergence result for the \( D^4L \) Algorithm. We will use the following notation: \( D^\nu \triangleq [D^\nu(1)^T, D^\nu(2)^T, \ldots, D^\nu(I)^T]^T \) and \( X^\nu \triangleq [X^\nu_1, X^\nu_2, \ldots, X^\nu_I] \).

Theorem 2 Let \( \{(D^\nu, X^\nu)\}_\nu \) be the sequence generated by the \( D^4L \) Algorithm for a given initial point \( (D^0, X^0) \), with either choice of \( \tilde{f}_i \)'s [see (7) and (19)] and \( \tilde{h}_i \)'s [see (11) and (20)]; and let \( \overline{D}^\nu \triangleq \frac{1}{I} \sum_{i=1}^I D^\nu_{(i)} \). Suppose that Assumptions A1-A4, A5 (i), B, C, D1-D2 are satisfied. Then,

(a) [Convergence]: \( \{(\overline{D}^\nu, X^\nu)\}_\nu \) is bounded and at least one of its limit points is a stationary solution of Problem P;

(b) [Consensus]: All \( D^\nu_{(i)} \)'s asymptotically reach consensus, i.e., \( \lim_{\nu \to \infty} \|D^\nu_{(i)} - \overline{D}^\nu\|_F = 0 \), for all \( i = 1, 2, \ldots, I \).

If instead Assumptions A1-A4, A5(ii), B, C, and D are satisfied, then convergence in (a) can be strengthened as follows:

15
(a') Every limit point of \( \{(D^\nu, X^\nu)\}_\nu \) is a stationary solution of Problem \( P \).

**Proof** The proof is quite involved and is given in the appendix, see Sec. A.3.

The above theorem states two main convergence results: i) asymptotic subsequence convergence of \( (D^\nu, X^\nu) \) to a stationary solution of Problem \( P \); and ii) asymptotic consensus of all \( D^\nu_i \) to a common value \( \bar{D}^\nu \). Convergence is guaranteed under two different set of assumptions; the first one (stated first) leads to weaker convergence results [subsequence convergence, as stated in (a)], while under the second set of assumptions, the stronger results in (a') can be proven [every limit point is a stationary solution].

4. Numerical Results

In this section, we numerically test 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, besides the \( D^4L \) Algorithm, there exists no other provably convergent distributed method for these classes of problems. Therefore, we compare our scheme with what we found in our experiments to be the best available heuristic: the Adapt-Then-Combine (ATC) Algorithm (Chainais and Richard, 2013). Note that ATC was not designed to be applied to directed graphs; hence, when dealing with directed graphs, we extended ATC in a natural way, partly using the ideas of this paper. Still, no formal proof of convergence is available for any simulated instance of ATC. When available, as benchmarks, we also simulate **centralized** algorithms tailored to the specific problem under consideration.

In our tests, we monitor the progress of the algorithms in terms of (i) the value of the objective function, (ii) the consensus disagreement, and (iii) a suitable stationarity measure. Specifically, given the tuple \( (D^\nu, X^\nu) \) generated by the algorithm under consideration at iteration \( \nu \), i) the objective function is evaluated at \( (D^\nu, X^\nu) \); ii) the consensus disagreement is defined as \( e^\nu = \max_i ||D^\nu_i - \bar{D}^\nu||_{\infty, \infty} \); and iii) the distance from stationarity is defined as \( \Delta^\nu = ||\text{vec}(\Delta_D^\nu, \Delta_X^\nu)||_{\infty} \) (Facchinei et al., 2015, Prop. 8(b)), with

\[
\Delta_D^\nu \triangleq \bar{D}^\nu - \tilde{D}^\nu, \quad \Delta_X^\nu \triangleq X^\nu - \tilde{X}^\nu,
\]

\[
\tilde{D}^\nu \triangleq \arg\min_{D \in \mathcal{D}} \sum_{i=1}^{I} \tilde{f}_i(D; D^\nu, X_i^\nu) + G(D),
\]

\[
\tilde{X}^\nu \triangleq \arg\min_{X: X_i \in \mathcal{X}_i, \forall i} \sum_{i=1}^{I} \tilde{h}_i(X_i; D^\nu, X_i^\nu) + g_i(X_i),
\]

where \( \tilde{f}_i \) and \( \tilde{h}_i \) are defined as in (19) and (20), respectively, with \( \tau_D^{i, \nu} = 1/I \) and \( \tau_X^{i, \nu} = 1/I \), for all \( i \) and \( \nu \). It is not difficult to check that \( \Delta^\nu \) is a valid distance from stationarity: \( \Delta^\nu \) is continuous and zero if and only if its argument is a stationary solution of Problem \( P \).

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.

4.1 Image Denoising

**Problem formulations:** We consider denoising a 512 \( \times \) 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 \( \text{SNR} \triangleq 10 \log_2 \left( \frac{\| \text{vec}(\mathbf{F}_0) \|_2}{\sqrt{\text{MSE}}} \right) \) while the Peak SNR (in dB) is defined as \( \text{PSNR} \triangleq 20 \log_2 (\max_j (\text{vec}(\mathbf{F}_0)_j \|) / \sqrt{\text{MSE}}) \), 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 (1). 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 \( 64 \times 1701 \) \((K = 64 \) and \( n_i = 1701)\). The total number of optimization variables is then 16,333,696. The free parameters \( \lambda \) and \( \mu \) in (1) are set to \( \lambda = 1/s \) and \( \mu = \lambda \), 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 ATC algorithm (Chaintais and Richard, 2013); and iii) the centralized K-SVD algorithm (Elad and Aharon, 2006) (KSVDB-Box v13 package), used as a benchmark to compare the final denoised images with those obtained by the aforementioned distributed algorithms. The two instances of the D\(^4\)L Algorithm are:

- **Plain D\(^4\)L:** \( \tilde{h}_i \) is chosen as in (11) (the original function) and \( \tilde{f}_i \) as in (19);
- **Linearized D\(^4\)L:** \( \tilde{h}_i \) is given in (20) (first-order approximation) and \( \tilde{f}_i \) is given in (19).

All the algorithms are initialized to the same value: \( \mathbf{D}^0 \)’s coincide with randomly (uniformly) chosen columns of \( \mathbf{S}_{(i)} \)’s whereas the sparse coding matrices \( \mathbf{X}_i^0 \)’s are set to the zero matrix. In both versions of D\(^4\)L schemes: \( \{\gamma^\nu\}_\nu \) is generated according to \( \gamma^\nu = \gamma^{\nu-1}(1 - \epsilon \gamma^{\nu-1}) \), with \( \gamma^0 = 0.5 \) and \( \epsilon = 10^{-2} \); and the proximal weights are set to \( \tau^\nu_{X,i} = 10 \) and \( \tau^\nu_{D,i} = \max(L\nabla \mathbf{X}_i (\mathbf{U}^\nu_{(i)}), 1) \) [cf. (27)].

While the subproblems solved at each iteration \( \nu \) in Linearized D\(^4\)L admit a closed-form—see (22) and (21)—in both Plain D\(^4\)L and ATC, the update of the dictionary has the closed form expression (21), but the update of the private variables calls for the solution of a LASSO problem (cf. Sec. 3.1). For both Plain D\(^3\)L and ATC, the LASSO subproblems at iteration \( \nu \) are solved using the (sub)gradient algorithm, with the following tuning. A diminishing step-size is used, set to \( \gamma = 0.9, \epsilon = 10^{-3}, \) and \( r \) denotes the inner iteration index. A warm start is used for the subgradient algorithm: the initial points are set to be equal to \( \mathbf{X}_i^\nu \); where \( \nu \) is the iteration index of the outer loop. We terminate the subgradient algorithm in the inner loop when \( J_i^\nu \leq 10^{-6} \), with

\[
J_i^\nu \triangleq \left\| \mathbf{X}_i^\nu - \frac{s}{1+s} \nabla^\frac{1}{2} \left( \mathbf{X}_i^\nu - \left( \nabla \mathbf{X}_i f_i (\mathbf{U}^\nu_{(i)}, \mathbf{X}_i^\nu) + \tau^\nu_{X,i} (\mathbf{X}_i^\nu - \mathbf{X}_i^\nu) \right) \right) \right\|_{\infty, \infty},
\]

where \( \mathbf{X}_i^\nu \) denotes the value of \( \mathbf{X}_i \) at the \( r \)-th inner iteration and outer iteration \( \nu \); and \( \nabla f (x) \triangleq \max(|x| - \theta, 0) \cdot \text{sign}(x) \) 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.

We simulated both undirected and directed static graphs. In the former case, all \( \phi^\nu \) in (14) are set to one, for all \( \nu \), which leads to doubly-stochastic weights \( w^\nu_{ij} = a^\nu_{ij} \). These
weights are set to be time-invariant and satisfying the Metropolis-Hasting rule (Xiao et al., 2007). When the graph is directed, we use the weights $w^{\nu}_{ij}$ given by (13), with $a^{\nu}_{ij}$ chosen according to the push-sum protocol (Kempe et al., 2003) (cf. Sec. 3.1).

**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 arcs 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 $e^{\nu}$ [subplot in the center], and the distance from stationarity $\Delta^{\nu}$ [subplot on the right] versus the number of message exchanges, achieved by Plain D$^4$L, Linearized D$^4$L, and ATC. Note that at each iteration ATC requires one message exchange while the D$^4$L schemes need two message exchanges per iteration. Therefore, in order to make a fair comparison, we report results by number of message exchanges and not by iteration number. Hence, 1000 message exchanges means 1000 iterations for ATC and 500 iterations for D$^4$L schemes. The figures clearly show that both versions of D$^4$L are much faster than ATC (or, equivalently, they require fewer information exchanges than ATC). Moreover, ATC does not seem to reach a consensus on the local copies of the dictionary, while D$^4$L schemes reach an agreement quite soon. The faster behavior of the proposed schemes seems mainly due to the gradient tracking mechanism. 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 ATC. 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 D$^4$L schemes and ATC after 200 message exchanges (we also plot the benchmark obtained by K-SVD that was run untill optimality). The figure shows that both versions of D$^4$L attain high quality solutions even if terminated after few message exchanges while ATC lags behind. This means that, in practice, there is no need to run D$^4$L till higher accuracy on $e^{\nu}$ or $\Delta^{\nu}$ is achieved.

Since the algorithms do not have the same cost-per-iteration—both Plain D$^4$L and ATC call for the solution at each iteration of a LASSO problem whereas the subproblems solved by Linearized D$^4$L admit a closed form solution [cf. (22) and (21)]—to get further insights
Decentralized Dictionary Learning Over Time-Varying Digraphs

Figure 4: Denoising problem – D⁴L and ATC algorithms: objective value [subplot on the left], consensus disagreement [subplot in the center], and distance from stationarity $\Delta^\nu$ [cf. (28)] [subplot on the right] vs. number of message exchanges.

Figure 5: Denoising outcome. (a): original image; (b): corrupted image; (c)-(e): denoising achieved by D⁴L and ATC terminated after 1000 message exchanges; and (f): denoising achieved by centralized K-SVD (KSVD-Box v13).

Into the performance of these algorithms, we also compare them in terms of running time. In Table 1, 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
D4L) 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 D4L scheme significantly outperforms ATC. Also, Linearized D4L performs considerably better than Plain D4L, when terminated early; the explanation is in Table 1 which shows that the time of one iteration of the former algorithm is much shorter than that of Plain D4L.

| Algorithm       | Average Time per Iteration (sec) |
|-----------------|----------------------------------|
| Linearized D4L  | 2.862                            |
| Plain D4L       | 11.328                           |
| ATC             | 9.838                            |

Table 1: D4L vs. 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
Figure 7: Denoising outcome. Denoising achieved by the Linearized D^4L [subplot on the left], the Plain D^4L [subplot in the center], and ATC [subplot on the right], all terminated after after 273 seconds run-time (corresponding to 200 message exchanges of the Linearized D^4L).

Table 2; 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 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^4L 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.

| 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 2: Network setting

In Fig. 8 we plot the average value of the objective function [subplot on the left], the consensus disagreement $e^\nu$ [subplot in the center], and the distance from stationarity $\Delta^\nu$ [subplot on the right], achieved by Plain D^4L, Linearized D^4L, 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 the aforementioned 5 digraph realizations. While also this batch of tests confirms the better behavior of D^4L 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^4L. The only aspect for which a reasonable influence can be seen is on consensus. In fact, with respect to consensus, Linearized D^4L seems to improve over Plain D^4L, when connectivity decreases. This has a natural interpretation. Plain D^4L solves much more accurate subproblems at each iteration and this, 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^4L is better than Plain D^4L, but soon, as information spreads, Plain D^4L becomes, even if slightly, better than Linearized D^4L.
Figure 8: Denoising problem–D⁴L and Modified ATC algorithms: objective value [subplots on the left], consensus disagreement [subplots in the center], and distance from stationarity $\Delta^n$ [cf. (28)] [subplots on the right] vs. number of message exchanges. Comparison over three network settings [cf. Table 2]: N1 [subplots (a)], N2 [subplots (b)], and N3 [subplots (c)].
The same behavior can be observed for scenario N2, but this time the initial advantage of Linearized D^4L 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.

4.2 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 (4), applied to genetic information. We solved the problem simulating a networked computer cluster composed of 500 nodes (see Table 2). The genetic data is borrowed from (Lee et al., 2010) (centered and normalized): the data matrix S of size 56 × 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 X_i, and the data matrix 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^4L where \( \tilde{f}_i \) and \( \tilde{h}_i \) are chosen according to (7) and (11), 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 \( \tau_D,i = 100 \) and \( \tau_X,i = \max(\nabla_X(U_\nu(i)), 100) \). We term such an instance of D^4L Plain D^4L. We compared Plain D^4L with the following algorithms: i) (a modified version of) the distributed ATC algorithm (Chainais and Richard, 2013), where the optimization of D is adjusted to solve (4) (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 \( X^0_i = 0, \) and each \( D^0_{(i)} \) equal to some randomly chosen columns of \( S_i. \)

In D^4L, the subproblems (8) and (10) at iteration \( \nu \) do not have a closed form solution; they are solved using the projected (sub)gradient algorithm, with diminishing step-size \( \gamma^\nu = \gamma^{\nu-1}(1 - \epsilon\gamma^{\nu-1}) \), 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 \( D^0_{(i)} \) and \( X^0_{(i)} \) in problems (8) and (10), respectively, where \( \nu \) is the iteration index of the outer loop. We terminate the projected subgradient algorithm solving (8) and (10) when \( J^\nu_{D,i} \triangleq \|\tilde{D}^{\nu,r}_{(i)} - D^{\nu,r}_{(i)}\|_{\infty,\infty} \leq 10^{-6} \) and \( J^\nu_{X,i} \triangleq \|\tilde{X}^{\nu,r}_{i} - X^{\nu,r}_{i}\|_{\infty,\infty} \leq 10^{-6}, \) respectively, where
measure the quality of the clustering by the Jaccard index. Jaccard indices from their average, computed over the aforementioned 5 realizations of the procedure described above applied to the outcome of the algorithm under consideration, patients’ information is in form of (unlabeled clusters of) data points \( \{D^m_{i,m}\}_{m=1}^{56} \), where \( D^m_{i,m} \) denotes the \( m \)-th row of \( D^\infty \) and represents an individual patient. In order to compare \( D^\infty \) with the labeled ground truth, we need to tag labels to the clustered points of \( D^\infty \). To do so, we run the K-means clustering algorithm on \( \{D^\infty_{m,i}\}_{m=1}^{56} \). 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”. 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 \( \{C_i\}_{i=1}^{4} \) (recall that there are 4 classes/types of cancer), where each \( C_i \) consists of the group of patients with the same type of cancer, and by \( \{\tilde{C}_i\}_{i=1}^{4} \) the clustering obtained by the procedure described above applied to the outcome \( D^\infty \) of the simulated algorithms, we measure the quality of the clustering by the Jaccard index, defined as

\[
J = \frac{\left| \bigcup_i (C_i \cap \tilde{C}_i) \right|}{\sum_i \left| C_i \cup \tilde{C}_i \right|}.
\]

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

In Table 3, we report the average and Maximum Absolute Deviation (MAD) of the Jaccard indices from their average, computed over the aforementioned 5 realizations of the

\[
\hat{D}^{\nu,r}_{(i)} = \arg\min_{D^\nu(i) \in \mathcal{D}} \left\{ \nabla_D f_i(D^\nu(i), X_i^{\nu,r}) + \Pi^\nu_{i} + \tau^\nu_{D,i} (D^\nu(i) - D^\nu(i)), D^\nu(i) - D^\nu_{(i)} \right\} + \frac{100}{2} \left\| D^\nu(i) - D^\nu_{(i)} \right\|^2 + G(D^\nu(i)),
\]

\[
\bar{X}_{i}^{\nu,r} = \arg\min_{X_i \in \mathbb{R}^{K \times n_i}} \left\{ \nabla_X f_i(U^\nu(i), X_i^{\nu,r}) + \tau^\nu_{X,i} (X_i^{\nu,r} - X_{i}), X_{i} - X_{i}^{\nu,r} \right\} + \frac{100}{2} \left\| X_{i} - X_{i}^{\nu,r} \right\|^2 + g_i(X_{i}),
\]

with \( D^\nu_{(i)} \) and \( X_{i}^{\nu,r} \) denoting the value of \( D^\nu(i) \) and \( X_{i} \) at the \( \nu \)-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 2. In Fig. 9 we plot the average value of the objective function [subplot on the left], the consensus disagreement \( e^\nu \) [subplot in the center], and the distance from stationarity \( \Delta^\nu \) [subplot on the right], achieved by Plain D^4L 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^4L 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.

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) \( D^\infty \) of the algorithm under consideration, patients’ information is in form of (unlabeled clusters of) data points \( \{D^\infty_{m,i}\}_{m=1}^{56} \), where \( D^\infty_{m,i} \) denotes the \( m \)-th row of \( D^\infty \) and represents an individual patient. In order to compare \( D^\infty \) with the labeled ground truth, we need to tag labels to the clustered points of \( D^\infty \). To do so, we run the K-means clustering algorithm on \( \{D^\infty_{m,i}\}_{m=1}^{56} \). 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”. 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 \( \{C_i\}_{i=1}^{4} \) (recall that there are 4 classes/types of cancer), where each \( C_i \) consists of the group of patients with the same type of cancer, and by \( \{\tilde{C}_i\}_{i=1}^{4} \) the clustering obtained by the procedure described above applied to the outcome \( D^\infty \) of the simulated algorithms, we measure the quality of the clustering by the Jaccard index, defined as

\[
J = \frac{\left| \bigcup_i (C_i \cap \tilde{C}_i) \right|}{\sum_i \left| C_i \cup \tilde{C}_i \right|}.
\]

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

In Table 3, we report the average and Maximum Absolute Deviation (MAD) of the Jaccard indices from their average, computed over the aforementioned 5 realizations of the

---

1. Given a clustering partition \( \{C_i\}_{i=1}^{4} \), the “within-cluster sum of point-to-centroid distances” measures the quality of the k-means clustering, and is defined as \( \sum_{i=1}^{4} \sum_{j \in C_i} ||D^\infty_{j,i} - D^\infty_{C_i}||^2 \), where \( D^\infty_{C_i} \equiv \frac{1}{|C_i|} \sum_{j \in C_i} D^\infty_{j,i} \) and \( |C_i| \) denotes the cardinality of the set \( C_i \).
Figure 9: Biclustering problem—Plain D^2L and Modified ATC algorithms: objective value [subplots on the left], consensus disagreement [subplots in the center], and distance from stationarity $\Delta^\nu$ [cf. (28)] [subplots on the right] vs. number of message exchanges. Comparison over three network settings [cf. Table 2]: N1 [subplots (a)], N2 [subplots (b)], and N3 [subplots (c)].
three graph topologies, as in Table 2 (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 3: Biclustering problem—Average/MAD of Jaccard indices over 5 realizations of di-graphs.

4.3 Non-negative Sparse Coding (NNSC) and Sparse PCA (SPCA)

**Problem Formulation:** We consider the Non-negative Sparse Coding (NSC) formulation (5) (Hoyer, 2004) and the Sparse PCA problem (4) (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 $19 \times 19$ pixels each (i.e. $M = 361$);
- The VOC 2006 database (Everingham et al. 2010): a pool of $N = 10,000$ vectorized natural image patches of size $16 \times 16$ pixels each (i.e. $M = 256$).

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

- NNSC (5): $K = 49$, $\lambda = \mu = 1/\sqrt{M}$, and $\alpha = 1$;
- Sparse PCA (4): $K = 49$, $\lambda_X = \mu_X = 1/\sqrt{M}$, $\lambda_D = \mu_D = 1/\sqrt{M}$, 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 4. Note that if $N/I$ is not an integer, we pad zero columns to the data matrix $S$ so that all the agents own equal-size partitions $S_i$’s, thus $n_i = \lceil N/I \rceil$ in both problems (4) and (5).

4.3.1 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 (19) and (11), respectively. The other parameters of the algorithm are set to: $\gamma^\nu = \gamma^\nu - 1(1 - \epsilon \gamma^\nu)$, with $\gamma^0 = 0.2$ and $\epsilon = 10^{-2}$; and $\tau^D_{D,i} = 10$ and $\tau^X_{X,i} = \max(L_{\nabla X}(U^\nu_{(i)}), 10)$. 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
Decentralized Dictionary Learning Over Time-Varying Digraphs

| 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 4: Network setting for the NNSC and Sparse PCA problems.

initialized setting $X_i^0 = 0$ and $D_i^0$ equal to some randomly chosen columns of $S_i$. Both Plain D^4L 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 (21). For both Plain D^4L and Modified ATC, the LASSO subproblems at iteration $\nu$ 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 $J_{\nu,i} \triangleq \|\hat{X}_{\nu,i} - X_{\nu,i}\|_{\infty,\infty} \leq 10^{-4}$, where

$$\hat{X}_{\nu,i} \triangleq \arg\min_{X_i \in \mathcal{X}} \left\langle \nabla_{X_i} f_i(U_{(i)}^\nu, X_i^{\nu,r}) + \tau_{X,i}^{\nu} (X_i^{\nu,r} - X_i^\nu), X_i - X_i^{\nu,r} \right\rangle + \frac{1}{2} \|X_i - X_i^{\nu,r}\|^2 + g_i(X_i),$$

and $X_i^{\nu,r}$ denotes the value of $X_i$ at the $\nu$-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^4L and Modified ATC algorithms over different network settings, as listed in Table 4, 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 5 and Table 6, 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 forementioned 5 graph realizations) of the objective function [subplot on the left], the consensus disagreement $e^\nu$ [subplot in the center], and the distance from stationarity $\Delta^\nu$ [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^4L significantly outperforms the Modified ATC algorithm. They also show a remarkable stability of Plain D^4L with respect to the simulated network graphs, which is not observed for the Modified ATC, whose performance deteriorates significantly going from N4 to N9.
Table 5: NNSC problem (MIT-CBCL dataset)–Plain D^4L/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        | 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 (VOC 2006 dataset)–Plain D^4L/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             |

4.3.2 Sparse Principal Component Analysis (SPCA)

**Algorithms and tuning:** We test the same D^4L version as used in the Biclustering experiments (cf. Subsec. 4.2), i.e., \( \bar{f}_i \) and \( \bar{h}_i \) are chosen according to (7) and (11), respectively; we set \( \gamma^\nu = \gamma^{\nu-1}(1 - \epsilon \gamma^{\nu-1}) \), with \( \gamma^0 = 0.2 \) and \( \epsilon = 10^{-2} \); and \( \tau_{D,i}^\nu = 10 \) and \( \tau_{X,i}^\nu = \max(L_{X,i}(U_i(\nu)), 10) \). We term it Plain D^4L. We compare Plain D^4L with a modified version of the ATC algorithm, which has been adapted to solve (4) and equipped with our new consensus protocol to handle directed network topologies. All the distributed algorithm are initialized, setting \( X_0 = 0 \) and \( D_0(i) \) equal to some randomly chosen columns of \( S_i \). The subproblems (8) and (10) at iteration \( \nu \) are solved using the projected (sub)gradient algorithm; the setting is the same as that used in the Biclustering problem (cf. Subsec. 4.2). We terminate the projected subgradient algorithm in the inner loop when \( J_{D,i}^r \leq 10^{-4} \) (in solving subproblem (8)) and \( J_{X,i}^r \leq 10^{-4} \) (in solving subproblem (10)), where \( J_{D,i}^r \) and \( J_{X,i}^r \) are defined as those in Subsec. 4.2. 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^4L and the Modified ATC in different network settings, as listed in Table 4. 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 7 and Figure 12 for the MIT-CBCL dataset; and in Table 8 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.

5. 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 results showed promising performance of our scheme with respect to state-of-the-art distributed methods.
Figure 11: NNSC problem (VOC 2006 dataset)–Plain D^4L and Modified ATC algorithms: objective value [subplots on the left], consensus disagreement [subplots in the center], and distance from stationarity Δν [cf. (28)] [subplots on the right] vs. number of message exchanges. Comparison over the network settings N4 [subplots (a)] and N9 [subplots (b)] (cf. Table 4).

Acknowledgments

The work of Daneshmand, Sun, and Scutari has been supported by the the National Science Foundation (NSF grants CIF 1564044 and CAREER Award No. 1555850) and the the Office of Naval Research (ONR Grant N00014-16-1-2244). 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).
Figure 12: SPCA problem (MIT-CBCL dataset)—Plain D⁴L and Modified ATC algorithms: objective value [subplots on the left], consensus disagreement [subplots in the center], and distance from stationarity Δν [cf. (28)] [subplots on the right] vs. number of message exchanges. Comparison over the network settings N4 [subplots (a)] and N9 [subplots (b)] (cf. Table 4).

| 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 7: SPCA problem (MIT-CBCL dataset)—Plain D⁴L/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 $\Delta^\nu$ [cf. (28)] [subplots on the right] vs. number of message exchanges. Comparison over three network settings [cf. Table 4: N4 [subplots (a)], N9 [subplots (b)].

| 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 8: 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.
Appendix A.

In this section, we prove convergence of the D^4L algorithm, as stated in Theorem 2. We first introduce some useful notation (cf. Sec. A.1) along with some preliminary results (cf. Sec. A.2). Theorem 2 is then proved in Sec. A.3.

A.1 Notation

Given Problem P, define

\[ F(D, X) \triangleq \sum_{i=1}^{I} f_i(D, X_i), \]
\[ U(D, X) \triangleq F(D, X) + \sum_{i=1}^{I} g_i(X_i) + G(D). \] (29)

Given the quantities \( \phi^\nu_i \) [cf. (14)], \( \tilde{D}^\nu_i \) [cf. (8)], \( U^\nu_i \) [cf. (9)], and \( \tilde{\Theta}^\nu_i \) [cf. (18)], define the following

\[ \phi^\nu \triangleq [\phi^\nu_1, \phi^\nu_2, \ldots, \phi^\nu_I]^\top, \quad \Phi^\nu \triangleq \text{diag}(\phi^\nu), \quad \hat{\Phi}^\nu \triangleq \Phi^\nu \otimes I_M, \]
\[ U^\nu \triangleq [U^\nu(1), U^\nu(2), \ldots, U^\nu(I)]^\top, \quad \tilde{D}^\nu \triangleq [\tilde{D}^\nu(1), \tilde{D}^\nu(2), \ldots, \tilde{D}^\nu(I)]^\top, \quad \tilde{\Theta}^\nu \triangleq [\tilde{\Theta}^\nu(1), \tilde{\Theta}^\nu(2), \ldots, \tilde{\Theta}^\nu(I)]^\top, \]

where \( I_M \) is the \( M \)-by-\( M \) identity matrix and \( \text{diag}(x) \) is a diagonal matrix whose diagonal entries are the components of the vector \( x \). Let us also introduce the following weighted average quantities:

\[ \mathbb{D}_{\phi^\nu} \triangleq \frac{1}{I} \sum_{i=1}^{I} \phi^\nu_i \mathbb{D}^\nu_i, \quad \mathbb{U}_{\phi^\nu} \triangleq \frac{1}{I} \sum_{i=1}^{I} \phi^\nu_i U^\nu_i. \] (30)

Using (15), (30) and the column stochasticity of \( A^\nu \) (cf. Assumption C3), it is not difficult to check that

\[ \mathbb{D}_{\phi^\nu+1} = \mathbb{U}_{\phi^\nu}, \] (31)

which, together with (9), yields

\[ \mathbb{D}_{\phi^\nu+1} = \mathbb{D}_{\phi^\nu} + \frac{\gamma^\nu}{I} \sum_{i=1}^{I} \phi^\nu_i \left( \tilde{D}^\nu_i - D^\nu_i \right). \] (32)

Finally, we stack column-wise the gradient matrices \( \nabla D f_i \)'s as

\[ G^\nu \triangleq \left[ \nabla D f_1(D^\nu(1), X^\nu_1)^\top, \nabla D f_2(D^\nu(2), X^\nu_2)^\top, \ldots, \nabla D f_I(D^\nu(I), X^\nu_I)^\top \right]^\top. \] (33)

Transition matrices and their properties. To analyze the dynamics of the D^4L algorithm, we introduce a few transition matrices. Recalling the definition of the weight matrix \( W^\nu \) [cf. (13)]:

\[ [W^\nu]_{i,j} = w^\nu_{ij} \triangleq \begin{cases} \frac{\gamma^\nu_{ij} \phi^\nu_j}{\sum_k a^\nu_{ik} \phi^\nu_k}, & \text{if } j \in A_i^{\text{in}}[^\nu] \\ 0, & \text{otherwise}, \end{cases} \] (34)
one can see that
\[ W^\nu = (\Phi^{\nu+1})^{-1} A^\nu \Phi^\nu, \quad \text{and} \quad W^\nu 1 = 1. \] (35)

Given \( W^\nu \), let us further define
\[
W^{\nu;l} \triangleq \begin{cases} 
W^\nu \cdot W^{\nu-1} \cdots W^l, & \nu > l, \\
W^\nu, & \nu = l, \\
0_l, & \nu < l,
\end{cases}
\] (36)

\[ \tilde{W}^{\nu;l} \triangleq W^{\nu;l} \otimes I_M, \quad \tilde{\Phi}^l \triangleq \Phi^l \otimes I_M, \] (37)

\[ J_\phi^\nu \triangleq \frac{1}{I} 1 \phi^{\nu\top}, \quad \tilde{J}_\phi^\nu \triangleq J_\phi^\nu \otimes I_M. \] (38)

We define similar quantities as (36) and (37) for the matrix \( A^\nu \), namely: \( A^{\nu;l} \), \( \tilde{A}^{\nu;l} \), and \( \tilde{A}^\nu \). For notational simplicity, when \( \phi^\nu = 1 \), we use \( J \) instead of \( J_\phi^\nu \) and \( \tilde{J} \triangleq J \otimes I_M \).

Using (30) and (35), it is not difficult to check that the following hold:
\[
\tilde{J}_\phi^\nu D^\nu = 1 \otimes D_\phi^\nu, \] (39)
\[
\tilde{J}_\phi^{\nu+1} \tilde{W}^{\nu;l} = \tilde{J}_\phi^l = \tilde{J}_\phi^l, \] (40)
\[
\tilde{W}^{\nu;l} = (\Phi^{\nu+1})^{-1} A^{\nu;l} \Phi^l. \] (41)

Finally, using the above notation, the main iterates of the D^4L Algorithm, i.e., (9), (14), (15), and (18), can be rewritten in compact form as
\[
U^\nu = D^\nu + \gamma^\nu (\tilde{D}^\nu - D^\nu), \] (42)
\[
\phi^{\nu+1} = A^\nu \phi^\nu, \] (43)
\[
D^{\nu+1} = \tilde{W}^{\nu} U^\nu, \] (44)
\[
\tilde{\Phi}^{\nu+1} = \tilde{W}^{\nu} \tilde{\Theta}^\nu + (\tilde{\Phi}^{\nu+1})^{-1} (G^{\nu+1} - G^\nu). \] (45)

A.2 Preliminary results

Some useful lemmas: The following lemmas will be used to prove asymptotic consensus among the local copies of the dictionary.

Lemma 3 Given the sequences \( \{a^\nu\}_\nu \) and \( \{b^\nu\}_\nu \), and a scalar \( \rho \in [0,1) \), the following hold:

(a) If \( \lim_{\nu \to \infty} a^\nu = 0 \), then,
\[
\lim_{\nu \to \infty} \sum_{t=1}^{\nu} a^t \rho^{\nu-t} = 0. \] (46)
\( (b) \) If \( \lim_{\nu \to \infty} \sum_{t=1}^{\nu} (a_t)^2 < \infty \) and \( \lim_{\nu \to \infty} \sum_{t=1}^{\nu} (b_t)^2 < \infty \), then

\[
\lim_{\nu \to \infty} \sum_{t=1}^{\nu} \left( \sum_{l=1}^{\nu} (a_t')^2 \rho^{l-t} \right) < \infty,
\]

(47)

\[
\lim_{\nu \to \infty} \sum_{t=1}^{\nu} (a_t)^2 \rho^{l-t} < \infty,
\]

(48)

\[
\lim_{\nu, \nu' \to \infty} \sum_{t=1}^{\nu} \left( \sum_{l=1}^{\nu'} (a_t')^2 \rho^{l-t} \right)^2 < \infty.
\]

(49)

**Proof** For the proof of (a) and (47)-(48) in (b), see (Nedić et al., 2010, Lemma 7). We prove next (49). Expand the LHS of (49) as:

\[
\lim_{\nu, \nu' \to \infty} \sum_{t=1}^{\nu} \left( \sum_{l=1}^{\nu} (a_t')^2 \rho^{l-t} \right)^2 = \lim_{\nu, \nu' \to \infty} \sum_{t=1}^{\nu} \sum_{l=1}^{\nu} \sum_{k=l}^{\nu'} (a_t')^2 \rho^{l-t} \rho^{k-l} \leq \lim_{\nu, \nu' \to \infty} \sum_{t=1}^{\nu} \sum_{l=1}^{\nu} \sum_{k=l}^{\nu'} (a_t')^2 \rho^{l-t} \rho^{k-l} = \lim_{\nu, \nu' \to \infty} \sum_{t=1}^{\nu} \sum_{l=1}^{\nu} \sum_{k=l}^{\nu'} (a_t')^2 \rho^{l-t} \rho^{k-l},
\]

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

\[
\lim_{\nu, \nu' \to \infty} \sum_{t=1}^{\nu} \left( \sum_{l=1}^{\nu} (a_t')^2 \rho^{l-t} \right)^2 \leq \frac{1}{1 - \rho} \lim_{\nu, \nu' \to \infty} \sum_{t=1}^{\nu} \sum_{l=1}^{\nu} (a_t')^2 \rho^{l-t} = \frac{1}{1 - \rho} \lim_{\nu \to \infty} \sum_{l=1}^{\nu'} \sum_{t=1}^{\nu} (a_t')^2 \rho^{l-t} \leq \frac{1}{1 - \rho} \lim_{\nu \to \infty} \sum_{t=1}^{\nu} (a_t')^2 < \infty.
\]

\[
\sum_{t=1}^{\nu} \left[ (W^{\nu,t})_{i,j} - (\xi^\nu)_j \right] \leq c \rho^{\nu-l+1} \quad \forall i, j, \quad \forall \nu \geq l \geq 0,
\]

where \( c = \frac{2(1 + \bar{\kappa}^{-1}(I-1)B)}{\rho(1 - \bar{\kappa}^{-1}(I-1)B)} \), \( \rho = (1 - \bar{\kappa}^{-1}(I-1)B) \frac{1}{\mu(I-1)} < 1 \), and \( \bar{\kappa} = \kappa^{I/B+1}/I \). Furthermore, the sequence \( \{\phi_i^\nu\}_{\nu} \) in (14) satisfies

\[
\xi^\phi \triangleq \inf_{\nu \in \mathbb{N}} \left( \min_{1 \leq i \leq l} \phi_i^\nu \right) \geq \kappa^{I/B}, \quad \text{and} \quad \tilde{\epsilon}_\phi \triangleq \sup_{\nu \in \mathbb{N}} \left( \max_{1 \leq i \leq l} \phi_i^\nu \right) \leq I - (I - 1)\kappa^{I/B}.
\]

(50)

If all the matrices \( A^\nu \) are doubly-stochastic, then \( \xi^\phi = \tilde{\epsilon}_\phi = 1 \).
Lemma 5 Under the setting of Lemma 4, it holds
\[ \|W^{\nu l} - J_{\phi^l}\|_2 \leq c_W \rho^{\nu - l + 1}, \quad \forall \nu \geq l \geq 0, \]
where \( \rho \in (0, 1) \) is defined in Lemma 4, and \( c_W > 0 \) is a constant.

Proof Using (41) and the column stochasticity of \( A^\nu \), one can check that the following chain of equalities holds:
\[ (\phi^{\nu + 1})^T W^{\nu l} = (\phi^{\nu + 1})^T (\Phi^{\nu + 1})^{-1} A^{\nu l} \Phi^l = 1^T A^{\nu l} \Phi^l = 1^T \Phi^l = \phi^l. \] (51)

Let us bound next the spectral norm of \( W^{\nu l} - J_{\phi^l} \). Denoting \( E^{\nu l} = W^{\nu l} - 1 \xi^l \) (where \( \xi^l \) is the stochastic vector defined in Lemma 4), we get
\[
\|W^{\nu l} - J_{\phi^l}\|_2 = \|W^{\nu l} - \frac{1}{I} 1 \phi^l\|_2 \overset{(a)}{=} \left\| I - \frac{1}{I} (\phi^{\nu + 1})^T \right\| W^{\nu l}\|_2
\leq (1 + \frac{1}{I} (\phi^{\nu + 1})^T \| E^{\nu l}\|_2 \leq (1 + I)c \rho^{\nu - l + 1},
\]
where (a) follows from (51); in (b) we used the triangle inequality; and (c) follows from Lemma 4. This concludes the proof, with \( c_W \) any constant such that \( c_W \geq (1 + I)c \).

Boundedness of the iterates. We first remark some important properties of the surrogate functions \( \tilde{f}_i \)'s and \( \tilde{h}_i \)'s.

Remark 6 The surrogate functions \( \tilde{f}_i \)'s [cf. (7) and (19)] and \( \tilde{h}_i \)'s [cf. (11) and (20)] satisfy the following: for all \( i = 1, 2, \ldots, I \),

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

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

The following lemma shows that the iterates \((D^\nu, X^\nu)\) are uniformly bounded, for all \( \nu \geq 0 \).

Lemma 7 Let \( \{(D^\nu, X^\nu)\}_\nu \) be the sequence generated by the \( D^4L \) Algorithm. Suppose that Assumptions A, D1 and D2 hold and \( \tilde{h}_i \) is chosen as in (11) or (20). Then, the \( \{(D^\nu, X^\nu)\}_\nu \) is bounded.

Proof We prove the lemma only for \( \tilde{h}_i \) given by (20); the proof can be easily tailored to the other choice of \( \tilde{h}_i \). If the sets \( X_i \) are bounded [Assumption A5i)], the lemma follows.
readily. We prove next the lemma under Assumptions. By the optimality of $X^\nu_{i+1}$ in (10), there exist $\xi^0_i \in \partial_X g_i(X^0_i)$ and $\Xi^\nu_{i+1} \in \partial_X g_i(X^\nu_{i+1})$ such that

$$0 \leq \left\langle \nabla_X \hat{h}_i(X^\nu_{i+1}, U^\nu_{(i)}(X^\nu_i), \Xi^\nu_{i+1}, X^0_i - X^\nu_{i+1} \right\rangle = \left\langle \Xi^\nu_{i+1} - \Xi^0_i + \tau^\nu_{i,i}(X^\nu_{i+1} - X^0_i), X^0_i - X^\nu_{i+1} \right\rangle + \left\langle \nabla_X f_i(U^\nu_{(i)}(X^\nu_i)), X^0_i - X^\nu_{i+1} \right\rangle - \left\langle \tau^\nu_{i,i}(X^\nu_i - X^0_i), X^0_i - X^\nu_{i+1} \right\rangle + \left\langle \nabla_X f_i(U^\nu_{(i)}(X^\nu_i)) + \Xi^0_i, X^0_i - X^\nu_{i+1} \right\rangle.$$

Using Remark 6 and the $\mu_i$-strongly convexity of $g_i$, we obtain

$$(\tau^\nu_{i,i} + \mu_i) \|X^\nu_{i+1} - X^0_i\|^2 \leq \left\langle \tau^\nu_{i,i} X^\nu_i - \nabla_X f_i(U^\nu_{(i)}(X^\nu_i), X^\nu_{i+1} - X^0_i \right\rangle$$

$$- \left\langle \tau^\nu_{i,i} X^0_i - \nabla_X f_i(U^\nu_{(i)}(X^\nu_i), X^\nu_{i+1} - X^0_i \right\rangle$$

$$- \left\langle \nabla_X f_i(U^\nu_{(i)}(X^\nu_i)) + \Xi^0_i, X^0_i - X^\nu_{i+1} \right\rangle.$$  

(52)

Define $\Upsilon^\nu_i(X_i) \triangleq \tau^\nu_{i,i} X_i - \nabla_X f_i(U^\nu_{(i)}(X^\nu_i)$ and rewrite (52) as

$$(\tau^\nu_{i,i} + \mu_i) \|X^\nu_{i+1} - X^0_i\|^2 \leq \|\Upsilon^\nu_i(X^\nu_i) - \Upsilon^\nu_i(X^0_i)\|_F + \|Z^\nu_i\|_F.$$  

(53)

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

$$\|\Upsilon^\nu_i(X^\nu_i) - \Upsilon^\nu_i(X^0_i)\|^2 = (\tau^\nu_{i,i})^2 \|X^\nu_i - X^0_i\|^2 + \|\nabla_X f_i(U^\nu_{(i)}(X^\nu_i)) - \nabla_X f_i(U^\nu_{(i)}(X^0_i))\|^2$$

$$- 2\tau^\nu_{i,i} \left\langle \nabla_X f_i(U^\nu_{(i)}(X^\nu_i)), X^\nu_i - \nabla_X f_i(U^\nu_{(i)}(X^\nu_i), X^\nu_i - X^0_i \right\rangle$$

$$\leq (\tau^\nu_{i,i})^2 \|X^\nu_i - X^0_i\|^2_F$$

$$+ \left(1 - \frac{2\tau^\nu_{i,i}}{L_{\nabla X_i(U^\nu_{(i)})}} \right) \left\| \nabla_X f_i(U^\nu_{(i)}(X^\nu_i)) - \nabla_X f_i(U^\nu_{(i)}(X^0_i)) \right\|^2_F,$$

(54)

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

$$\left\langle \nabla_X f_i(U^\nu_{(i)}(X^\nu_i)), X^\nu_i - X^0_i \right\rangle \geq \frac{1}{L_{\nabla X_i(U^\nu_{(i)})}} \left\| \nabla_X f_i(U^\nu_{(i)}(X^\nu_i)) - \nabla_X f_i(U^\nu_{(i)}(Y^\nu_i)) \right\|^2_F, \quad \forall X^\nu_i, Y^\nu_i \in X^\nu_i.$$

Note that $\|\Upsilon^\nu_i(X^\nu_i) - \Upsilon^\nu_i(X^0_i)\|_F \leq \tau^\nu_{i,i} \|X^\nu_i - X^0_i\|_F$ as long as $\tau^\nu_{i,i} \geq \frac{1}{2} L_{\nabla X_i(U^\nu_{(i)})}$, which is satisfied under D2. Therefore, we can bound (53) as

$$(\tau^\nu_{i,i} + \mu_i) \|X^\nu_{i+1} - X^0_i\|_F \leq \tau^\nu_{i,i} \|X^\nu_i - X^0_i\|_F + B_Z.$$  

(55)
To prove (57), let us first expand and

\[ \nabla \] is that

\[ \text{Remark 8 (On the Lipschitz continuity)} \]

where the second inequality is due to \( X \). Evidently

\[ \text{Proposition 9} \]

We prove the asymptotic consensus of the local copies

\[ \text{D} \times \]

for all \( i \), it follows that \( \text{(B)} \), and \( \text{(37)} \]. Similarly, we expand the subtrahend as

\[ \text{We can now prove that, starting from} \ X^0_i, \ \text{the iterates} \ X^\nu_i \ \text{stays in the ball} \ B_i(R_i, X^0_i) \ \triangleq \ \{ X_i \in \mathbb{R}^{K \times n_i} : \| X_i - X^0_i \|_F \leq R_i \} \], for all \( \nu \geq 1 \), where \( R_i \geq B_Z / \mu_i \). Let us prove it by induction. Evidently \( X^0_i \in B_i(R_i, X^0_i) \). Let \( X^\nu_i \in B_i(R_i, X^0_i) \); by (55), we get

\[ \| X^\nu_i - X^0_i \|_F \leq \frac{\tau X_i^\nu}{\tau X_i^0 + \mu_i} \| X^\nu_i - X^0_i \|_F + \frac{B_Z}{\tau X_i^0 + \mu_i} \leq R_i, \]

where the second inequality is due to \( X^\nu_i \in B_i(R, X^0_i) \) and \( R \geq B_Z / \mu \). Hence \( X^\nu_i \in B_i(R_i, X^0_i) \). Therefore, \( X^\nu_i \in B_i(R_i, X^0_i) \), for all \( \nu \geq 0 \). Since \( D \) is bounded (cf. Assumption A3), it follows that \( (D^\nu_i, X^\nu_i) \in D \times B_i(R_i, X^0_i) \), for all \( \nu \geq 0 \).

\[ \text{Remark 8 (On the Lipschitz continuity)} \]

Since \( f_i \) is \( C^2 \), a direct consequence of Lemma 7 is that \( \nabla f_i \) [the gradient of \( f_i \) with respect to \( (D, X_i) \)] is Lipschitz continuous on \( D \times B_i(R_i, X^0_i) \), that is, there exists some positive finite constant \( L_{\nabla,i} \) such that

\[ \| \nabla f_i(D, X_i) - \nabla f_i(D', X'_i) \|_F \leq L_{\nabla,i} \| (D, X_i) - (D', X'_i) \|_F, \]

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

The above result also implies that \( \nabla D F : D \times (X_1 \times \cdots \times X_I) \to D \) [cf. (29)] is Lipschitz continuous on \( D \times (B_i(R_i, X^0_i) \times \cdots \times B_i(R_i, X^0_i)) \), with constant \( L_{\nabla,D} \triangleq I \cdot L_{\nabla} \).

\[ \text{Consensus properties.} \]

We prove the asymptotic consensus of the local copies \( D_i \)'s.

\[ \text{Proposition 9} \]

Let \( \{ (D^\nu_i, X^\nu_i) \} \) be the sequence generated by the \( D^4L \) Algorithm, under the setting of Theorem 2. Then, the following hold:

\[ \lim_{\nu \to \infty} \| D^\nu - 1 \otimes \overline{D}_{\phi^\nu} \|_F = 0, \]

\[ \lim_{\nu \to \infty} \sum_{t=1}^{\nu} \| D^t - 1 \otimes \overline{U}_{\phi^t} \|_F^2 < \infty; \]

and

\[ \lim_{\nu \to \infty} \sum_{t=1}^{\nu} \| U^t - 1 \otimes \overline{U}_{\phi^t} \|_F^2 < \infty, \]

\[ \lim_{\nu \to \infty} \sum_{t=1}^{\nu} \| D^t - D^{t-1} \|_F^2 < \infty. \]

\[ \text{Proof} \]

To prove (57), let us first expand \( D^\nu - 1 \otimes \overline{D}_{\phi^\nu} \) as follows: for any \( \nu \geq 1 \),

\[ D^\nu \overset{(a)}{=} \hat{W}^{\nu-1} U^{\nu-1} = \hat{W}^{\nu-1} D^{\nu-1} + \hat{W}^{\nu-1} (U^{\nu-1} - D^{\nu-1}) \]

\[ \overset{(b)}{=} \hat{W}^{\nu-1} D^0 + \sum_{t=0}^{\nu-1} \hat{W}^{\nu-1:t} (U^t - D^t), \]

where (a) is due to (44); and (b) follows by induction and the definition of \( \hat{W}^{\nu:t} \) [cf. (36) and (37)]. Similarly, we expand the subtrahend as

\[ 1 \otimes \overline{D}_{\phi^\nu} \overset{(a)}{=} \hat{J}_{\phi^\nu} D^\nu \overset{(b)}{=} \hat{J}_{\phi^\nu} \left( \hat{W}^{\nu-1} D^0 + \sum_{t=0}^{\nu-1} \hat{W}^{\nu-1:t} (U^t - D^t) \right) \]

\[ \overset{(c)}{=} \hat{J} \left( D^0 + \sum_{t=0}^{\nu-1} F^t (U^t - D^t) \right), \]

38
where (a) is due to (39); in (b) we used (61); and (c) follows from (40). Subtracting (62) from (61) and using (42), yields

$$\| \mathbf{D}^\nu - 1 \otimes \mathbf{D}_{\phi^\nu} \|_F \leq \| \hat{\mathbf{W}}^\nu_{\nu-1:0} - \hat{\mathbf{J}} \|_2 \| \mathbf{D}^0 \|_F + \sum_{t=0}^{\nu-1} \gamma^t \| \hat{\mathbf{W}}^\nu_{\nu-1:t} - \hat{\mathbf{J}}_{\phi^\nu} \|_2 \| \hat{\mathbf{D}}^t - \mathbf{D}^t \|_F$$

$$\leq c_1 \rho^\nu + c_2 \sum_{t=0}^{\nu-1} \gamma^t \rho^{t-t} \xrightarrow{\nu \to \infty} 0,$$

for some finite constants $c_1, c_2 > 0$, where (a) is due to Lemma 5 and the boundedness of $\{\| \hat{\mathbf{D}}^\nu - \mathbf{D}^\nu \| \}_\nu$; and (b) follows from Lemma 3(a).

Let us now proceed to prove (58). Using (63), we have

$$\lim_{\nu \to \infty} \sum_{t=1}^{\nu} \| \mathbf{D}^t - 1 \otimes \mathbf{D}_{\phi^\nu} \|_F^2 \leq \lim_{\nu \to \infty} \sum_{t=1}^{\nu} \left( c_1 \rho^t + c_2 \sum_{l=0}^{t-1} \gamma^l \rho^{t-l} \right)^2$$

$$\leq \frac{2c_1^2}{1 - \rho^2} + \frac{2c_2^2}{1 - \rho^2} \lim_{\nu \to \infty} \sum_{t=1}^{\nu} \sum_{l=0}^{t-1} \sum_{k=0}^{t-1} \gamma^l \gamma^k \rho^{t-k} \rho^{t-l}$$

$$\leq \frac{2c_1^2}{1 - \rho^2} + \frac{2c_2^2}{1 - \rho^2} \lim_{\nu \to \infty} \sum_{t=1}^{\nu} \sum_{l=0}^{t-1} \sum_{k=0}^{t-1} \gamma^l \rho^{t-k} \rho^{t-l}$$

$$\leq \frac{2c_1^2}{1 - \rho^2} + \frac{2c_2^2}{1 - \rho^2} \lim_{\nu \to \infty} \sum_{t=1}^{\nu} \sum_{l=0}^{t-1} (\gamma^l)^2 \rho^{t-l} \leq \infty,$$

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 (48) (cf. Lemma 3 (b)).

We prove now (59). Using (30) and (42), we get

$$\| \mathbf{U}^t - 1 \otimes \mathbf{U}_{\phi^\nu} \|_F^2 \leq \gamma^t \left( \mathbf{D}^t - 1 \otimes \frac{1}{T} \sum_{i=1}^{I} \phi_i^t \mathbf{D}_t^{(i)} \right) + (1 - \gamma^t) \left( \mathbf{D}^t - 1 \otimes \mathbf{D}_{\phi^\nu} \right)$$

$$\leq 2 \gamma^t \| \mathbf{D}^t - 1 \otimes \frac{1}{T} \sum_{i=1}^{I} \phi_i^t \mathbf{D}_t^{(i)} \|_F^2 + 2 \| \mathbf{D}^t - 1 \otimes \mathbf{D}_{\phi^\nu} \|_F^2,$$

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

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

where $c_3$ is a positive finite constant, and (a) follows from Assumption D1 and (58).

We finally prove (60). Using the Jensen inequality and (32), (60) can be bounded as

$$\lim_{\nu \to \infty} \sum_{t=1}^{\nu} \| \mathbf{D}^t - \mathbf{D}^{t-1} \|_F^2 \leq 3 \lim_{\nu \to \infty} \sum_{t=1}^{\nu} \| \mathbf{D}^t - 1 \otimes \mathbf{D}_{\phi^\nu} \|_F^2 + 3 \lim_{\nu \to \infty} \sum_{t=1}^{\nu} \| \mathbf{D}^{t-1} - 1 \otimes \mathbf{D}_{\phi^{t-1}} \|_F^2$$

$$+ \frac{3}{T} \lim_{\nu \to \infty} \sum_{t=1}^{\nu} (\gamma^{t-1})^2 \| \sum_{i=1}^{I} \phi_i^{t-1} (\mathbf{D}^{t-1}_t - \mathbf{D}_t^{(i)}) \|_F^2 < \infty,$$

39
there the last inequality follows from (58) and the boundedness of \( \{\phi_{i}^{\nu}\} \) and \( \{\tilde{D}_{(i)}^{\nu} - D_{(i)}^{\nu}\} \).

\( \blacksquare \)

**Remark 10 (On \( L_{\nabla X_{i}}(\tilde{U}_{\phi}^{\nu}) \) and \( L_{\nabla X_{i}}(U_{(i)}^{\nu}) \))** Recall that \( \nabla_{X_{i}} f_{i}(\tilde{U}_{\phi}^{\nu}, \bullet) \) is Lipschitz continuous on \( X_{i} \), with constant \( L_{\nabla X_{i}}(\tilde{U}_{\phi}^{\nu}) \) (cf. Assumption A2). Since \( \|U_{(i)}^{\nu} - \tilde{U}_{\phi}^{\nu}\|_{F} \xrightarrow{\nu \to \infty} 0 \) [cf. (59), Proposition 9] and \( L_{\nabla X_{i}}(D) \) is continuous, we have

\[
\|L_{\nabla X_{i}}(\tilde{U}_{\phi}^{\nu}) - L_{\nabla X_{i}}(U_{(i)}^{\nu})\|_{F} \xrightarrow{\nu \to \infty} 0, \quad i = 1, 2, \ldots, I. \tag{65}
\]

**On the properties of the best-response map.** Some key properties of the best-response maps defined in (6), (8) and (10) are needed and described next.

**Proposition 11** Let \( \{\{D^{\nu}, X^{\nu}\}\}_{\nu} \) be the sequence generated by the \( D^{4}L \) Algorithm, under Assumptions A, B, C, D1, D2 and with either choices of \( \tilde{f}_{i} \)'s [see (7) and (19)] and \( \tilde{h}_{i} \)'s [see (11) and (20)]. Given the solution maps defined in (6), (8) and (10), 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 \( \nu \geq 1 \),

\[
\left\langle \nabla D\left(\tilde{D}_{\phi}^{\nu}, X^{\nu}\right), \sum_{i=1}^{I} \phi_{i}^{\nu} \left(\tilde{D}_{(i)}^{\nu} - D_{(i)}^{\nu}\right) \right\rangle + \sum_{i=1}^{I} \phi_{i}^{\nu} \left(G(\tilde{D}_{(i)}^{\nu}) - G(D_{(i)}^{\nu})\right) \leq -s_{D}\left(\|\tilde{D}^{\nu} - D^{\nu}\|_{F} - \frac{I\bar{\epsilon}_{\phi} T^{\nu}}{2s_{D}}\right)^{2} + \eta \|\tilde{D}^{\nu} - D^{\nu}\|_{F} \sum_{i=1}^{\nu} \rho^{\nu-t} \|X^{t} - X^{t-1}\|_{F} + \frac{I^{2} \bar{\epsilon}_{\phi}^{2}}{4s_{D}} (T^{\nu})^{2}, \tag{66}
\]

where \( \bar{\epsilon}_{\phi} \) is defined in (50);

(b) There exist finite constants \( s_{X} > 0 \) and \( L_{X} > 0 \), such that, for all \( \nu \geq 1 \),

\[
\sum_{i=1}^{I} \left\langle \nabla_{X_{i}} f_{i}(\tilde{D}_{\phi}^{\nu+1}, X_{i}^{\nu+1}), X_{i}^{\nu+1} - X_{i}^{\nu} \right\rangle + \sum_{i=1}^{I} \left(g_{i}(X_{i}^{\nu+1}) - g_{i}(X_{i}^{\nu})\right) \leq -\sum_{i=1}^{I} \tau_{X,i}^{\nu+1} \|X_{i}^{\nu+1} - X_{i}^{\nu}\|_{F}^{2} + L_{X} \|U^{\nu} - 1 \times \tilde{U}_{\phi}^{\nu}\|_{F} \|X^{\nu+1} - X^{\nu}\|_{F}; \tag{67}
\]

(c) There exists some constant \( L_{D} > 0 \), such that, for any integers pair \( \nu_{1} \) and \( \nu_{2} \),

\[
\|\tilde{D}^{\nu_{2}} - \tilde{D}^{\nu_{1}}\|_{F} \leq L_{D}\left(\|D^{\nu_{2}} - D^{\nu_{1}}\|_{F} + \|X^{\nu_{2}} - X^{\nu_{1}}\|_{F} + \sum_{i=1}^{I} |\tau_{D,i}^{\nu_{2}} - \tau_{D,i}^{\nu_{1}}|\right); \tag{68}
\]

(d) There exist some constants \( p_{X}, q_{X} > 0 \) and a sufficiently large \( \nu_{X} \) such that, for all \( \nu \geq \nu_{X} \),

\[
\|X^{\nu+1} - X^{\nu}\|_{F} \leq p_{X} \|X^{\nu} - X^{\nu-1}\|_{F} + q_{X} \|U^{\nu} - U^{\nu-1}\|_{F}. \tag{69}
\]

Furthermore, if Assumption D3 holds, then \( p_{X} < 1 \).
Proof (a) It follows from the optimality of $\tilde{D}_i^\nu$ [cf. (8)] and convexity of $G$ that
\[
\left\langle \nabla_D \tilde{f}_i(D_i^\nu; D_i^\nu, X_i^\nu) + \tilde{\Pi}_i^\nu, D_i^\nu - \tilde{D}_i^\nu \right\rangle + G(D_i^\nu) - G(\tilde{D}_i^\nu) \geq 0. \tag{70}
\]

Adding and subtracting inside the first term $\nabla_D \tilde{f}_i(D_i^\nu; D_i^\nu, X_i^\nu)$ and $\sum_j \nabla_D f_j(D_i^\nu, X_i^\nu)$ and using $\nabla_D \tilde{f}_i(D_i^\nu; D_i^\nu, X_i^\nu) = \nabla_D f_i(D_i^\nu, X_i^\nu)$ [cf. Remark 6], inequality (70) becomes
\[
\left\langle \nabla_D \tilde{f}_i(D_i^\nu; D_i^\nu, X_i^\nu) - \nabla_D \tilde{f}_i(D_i^\nu; D_i^\nu, X_i^\nu), \tilde{D}_i^\nu - D_i^\nu \right\rangle \\
+ \left\langle \sum_{j=1}^I \nabla_D f_j(D_i^\nu; X_j^\nu), \tilde{D}_i^\nu - D_i^\nu \right\rangle + G(\tilde{D}_i^\nu) - G(D_i^\nu) \leq 0.
\]

Invoking the uniform strongly convexity of $\tilde{f}_i(\cdot; D_i^\nu, X_i^\nu)$, the definition of $\tilde{\Theta}_i^\nu$ in (17), and recalling that $\nabla_D F(\tilde{D}_i^\nu; X_i^\nu) = \sum_j \nabla_D f_j(D_i^\nu; X_j^\nu)$, we get
\[
\left\langle \nabla_D F(\tilde{D}_i^\nu, X_i^\nu), \tilde{D}_i^\nu - D_i^\nu \right\rangle + G(\tilde{D}_i^\nu) - G(D_i^\nu) \\
\leq -\tau_{D,i} \|\tilde{D}_i^\nu - D_i^\nu\|^2 + \frac{1}{I} \sum_{j=1}^I \nabla_D f_j(D_i^\nu; X_j^\nu) \|F\| \|\tilde{D}_i^\nu - D_i^\nu\|_F. \tag{71}
\]

Multiplying both side of the above inequality by the positive quantities $\phi_i^\nu$ (cf. Lemma 4) and summing over $i = 1, 2, \ldots, I$, and using $\phi_i^\nu \leq \bar{\varepsilon}_\phi$ (see Lemma 4), yields
\[
\left\langle \nabla_D F(\tilde{D}_i^\nu; X_i^\nu), \sum_{i=1}^I \phi_i^\nu \left( \tilde{D}_i^\nu - D_i^\nu \right) \right\rangle + \sum_{i=1}^I \phi_i^\nu \left( G(\tilde{D}_i^\nu) - G(D_i^\nu) \right) \\
\leq -s_D \|\tilde{D}^\nu - D^\nu\|^2 + I \bar{\varepsilon}_\phi \left\| \tilde{\Theta}^\nu - 1 \right\| \sum_{i=1}^I \nabla_D f_i(D_i^\nu; X_i^\nu) \|F\| \|\tilde{D}_i^\nu - D_i^\nu\|_F, \tag{71}
\]
where $s_D$ is any positive constant such that $s_D \leq \min_i \phi_i^\nu \tau_{D,i}$ [note that such a constant exists because $\phi_i^\nu \geq \xi_D$, with $\xi_D > 0$ defined in Lemma 4, and all $\tau_{D,i}$ are uniformly bounded away from zero–see Assumption D2].

Now let us bound the gradient tracking error term in (71). Using (45) recursively, $\tilde{\Theta}_i^\nu$ can be rewritten as
\[
\tilde{\Theta}_i^\nu = \tilde{\Theta}_i^{\nu-1} + \sum_{t=1}^{\nu-1} \tilde{\Theta}_i^{\nu-1:t} \left( \tilde{\Phi}_i \right)^{-1} (G_i^t - G_i^{t-1}) + \left( \tilde{\Phi}_i \right)^{-1} (G_i^\nu - G_i^{\nu-1}). \tag{72}
\]
Using the definition of $\mathbf{G}^\nu$ [cf. (33)] and $\mathbf{J}$ [cf. (38)], write

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

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

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

Using (72) and (73), the gradient tracking error term in (71) can be upper bounded as

$$\left\| \bar{\Theta}^\nu - 1 \otimes \frac{1}{T} \sum_{i=1}^{I} \nabla_D f_i(\mathbf{D}_{\phi^\nu}, \mathbf{X}_i^\nu) \right\|_F \leq \left\| \hat{\mathbf{W}}^{\nu-1:0} - \mathbf{J} \right\|_2 \left\| \bar{\Theta}^0 \right\|_F + \frac{1}{L_{\phi}} \sum_{t=1}^{\nu-1} \left\| \hat{\mathbf{W}}^{\nu-1:t} - \mathbf{J}\phi^t \right\|_2 \| \mathbf{G}_t - \mathbf{G}_{t-1} \|_F$$

$$+ \left\| \left( \hat{\mathbf{J}}^\nu \right)^{-1} - \mathbf{J} \right\|_2 \| \mathbf{G}^\nu - \mathbf{G}^{\nu-1} \|_F + \frac{1}{\sqrt{T}} \sum_{i=1}^{I} \left\| \nabla_D f_i(\mathbf{D}_{\phi^\nu}, \mathbf{X}_i^\nu) - \nabla_D f_i(\mathbf{D}_{(i)}^\nu, \mathbf{X}_i^\nu) \right\|_F$$

$$\leq c_4 \rho^\nu + c_5 L_{\nabla} \sum_{t=1}^{\nu} \rho^{\nu-t} (\left\| \mathbf{D}_t - \mathbf{D}_{t-1} \right\|_F + \left\| \mathbf{X}_t - \mathbf{X}_{t-1} \right\|_F) + L_{\nabla} \| \mathbf{D}^\nu - 1 \otimes \mathbf{D}_{\phi^\nu} \|_F$$

$$\leq T^\nu + c_5 L_{\nabla} \sum_{t=1}^{\nu} \rho^{\nu-t} \left\| \mathbf{X}_t - \mathbf{X}_{t-1} \right\|_F ,$$

for some positive finite constants $c_4$ and $c_5$, where in (a) we used the lower bound $\phi^\nu_t \geq \mathbf{L}_{\phi}$ (Lemma 4) and $\mathbf{J}_{\phi^\nu} = \mathbf{J} \hat{\Phi}^t$ [cf. (40)]; and in (b) we used (33), (56) (cf. Remark 8), and Lemma 5; and in (c) we defined $T^\nu$ as

$$T^\nu \triangleq c_4 \rho^\nu + c_5 L_{\nabla} \sum_{t=1}^{\nu} \rho^{\nu-t} \left\| \mathbf{D}_t - \mathbf{D}_{t-1} \right\|_F + L_{\nabla} \| \mathbf{D}^\nu - 1 \otimes \mathbf{D}_{\phi^\nu} \|_F .$$

Substituting (74) into (71) yields
\[ \left\langle \nabla_D F(\mathbf{D}_{\phi^0}, \mathbf{X}^\nu), \sum_{i=1}^{I} \phi_i^\nu \left( \mathbf{D}_i^\nu - \mathbf{D}_i^0 \right) \right\rangle + \sum_{i=1}^{I} \phi_i^\nu \left( G(\mathbf{D}_i^\nu) - G(\mathbf{D}_i^0) \right) \]
\[ \leq -s_D \| \mathbf{D}^\nu - \mathbf{D}^0 \|^2_F + I \bar{\epsilon}_\phi T^\nu \| \mathbf{D}^\nu - \mathbf{D}^0 \|_F \]
\[ + I \bar{\epsilon}_\phi c_5 L_\nu \| \mathbf{D}^\nu - \mathbf{D}^0 \|_F \sum_{t=1}^{\nu} \rho^{\nu-t} \| \mathbf{X}^t - \mathbf{X}^{t-1} \|_F \]
\[ = -s_D \left( \| \mathbf{D}^\nu - \mathbf{D}^0 \|_F - \frac{I \bar{\epsilon}_\phi T^\nu}{2 s_D} \right)^2 + I \frac{\bar{\epsilon}_\phi^2}{4 s_D} (T^\nu)^2 \]
\[ + \eta \| \mathbf{D}^\nu - \mathbf{D}^0 \|_F \sum_{t=1}^{\nu} \rho^{\nu-t} \| \mathbf{X}^t - \mathbf{X}^{t-1} \|_F, \]

with \( \eta = I \bar{\epsilon}_\phi c_5 L_\nu \).

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 (75) is square summable, and so is the third one, due to Proposition 9 [cf. (58)]. Invoking Jensen’s inequality, it is sufficient to show that the second term on RHS of (75) is square summable. Following the same approach used to prove (58), we have

\[ \lim_{\nu \to \infty} \sum_{t=1}^{\nu} \left( \sum_{l=1}^{t} \rho^{t-l} \| \mathbf{D}^l - \mathbf{D}^{l-1} \|_F \right)^2 \]
\[ \leq \lim_{\nu \to \infty} \sum_{t=1}^{\nu} \sum_{l=1}^{t} \sum_{k=1}^{t} \rho^{t-l} \rho^{t-k} \| \mathbf{D}^l - \mathbf{D}^{l-1} \|_F \| \mathbf{D}^k - \mathbf{D}^{k-1} \|_F \]
\[ \leq (a) \frac{1}{1 - \rho} \lim_{\nu \to \infty} \sum_{t=1}^{\nu} \sum_{l=1}^{t} \rho^{t-l} (\gamma^l)^2 \| \mathbf{D}^l - \mathbf{D}^{l-1} \|_F^2 \leq (b) \frac{\nu}{\rho} < \infty, \]

where (a) follows from \( ab \leq (a^2 + b^2)/2 \), and (b) is due to Lemma 3 and Assumption A3. Hence \( \sum_{\nu=1}^{\infty} (T^\nu)^2 < \infty \).

(b) We prove this statement using the definition (19) of \( \tilde{f}_i \); the same conclusion holds also using the alternative choice (7) of \( \tilde{f}_i \); the proof is thus omitted. Invoking the optimality of \( \mathbf{X}^{\nu+1} \) [cf. (10)] together with the convexity of \( g_i \), yield

\[ \left\langle \nabla_X \tilde{h}_i(\mathbf{X}_i^{\nu+1}; \mathbf{U}_i^\nu, \mathbf{X}_i^\nu), \mathbf{X}_i^\nu - \mathbf{X}_i^{\nu+1} \right\rangle \]
\[ + \left\langle \nabla_X \tilde{h}_i(\mathbf{X}_i^{\nu}; \mathbf{U}_i^\nu), \mathbf{X}_i^\nu - \mathbf{X}_i^{\nu+1} \right\rangle \]
\[ + \left\langle \nabla_X \tilde{h}_i(\mathbf{X}_i^{\nu}; \mathbf{U}_\phi^\nu, \mathbf{X}_i^\nu), \mathbf{X}_i^\nu - \mathbf{X}_i^{\nu+1} \right\rangle + g_i(\mathbf{X}_i^\nu) - g_i(\mathbf{X}_i^{\nu+1}) \geq 0. \]

Using Remark 6 and \( \mathbf{U}_\phi^\nu = \mathbf{D}_{\phi^\nu+1} \) [cf. (31)], we obtain

\[ \left\langle \nabla_X f_i(\mathbf{D}_{\phi^\nu+1}, \mathbf{X}_i^\nu), \mathbf{X}_i^{\nu+1} - \mathbf{X}_i^\nu \right\rangle + g_i(\mathbf{X}_i^{\nu+1}) - g_i(\mathbf{X}_i^\nu) \]
\[ \leq -\tau_X \| \mathbf{X}_i^{\nu+1} - \mathbf{X}_i^\nu \|^2_F + \left\langle \nabla_X f_i(\mathbf{U}_\phi^\nu, \mathbf{X}_i^\nu), \mathbf{X}_i^{\nu+1} - \mathbf{X}_i^\nu \right\rangle, \]

which, together with (56), leads to the desired result (67), with \( L_X = \sqrt{I} L_\nu \).
(c) Using the optimality of $\hat{D}^\nu_i$ defined in (6) together with convexity of $G$, yields

$$\begin{align*}
\langle \nabla_D \tilde{f}_i(\hat{D}^\nu_i, D^\nu_i, X^\nu_i) - \nabla_D \tilde{f}_i(\hat{D}^\nu_i, D^\nu_i, X^\nu_i), \hat{D}^\nu_i - \hat{D}^\nu_i \rangle \\
bigger\langle \nabla_D \tilde{f}_i(\hat{D}^\nu_i, D^\nu_i, X^\nu_i) - \nabla_D \tilde{f}_i(\hat{D}^\nu_i, D^\nu_i, X^\nu_i), \hat{D}^\nu_i - \hat{D}^\nu_i \rangle \bigg\rangle
\end{align*}$$

Summing the two inequalities above while adding/subtracting inside the inner product $\nabla_D \tilde{f}_i(D^\nu_i, D^\nu_i, X^\nu_i)$ and using (16), yield

$$\begin{align*}
\langle \nabla_D \tilde{f}_i(\hat{D}^\nu_i, D^\nu_i, X^\nu_i) - \nabla_D \tilde{f}_i(\hat{D}^\nu_i, D^\nu_i, X^\nu_i), \hat{D}^\nu_i - \hat{D}^\nu_i \rangle \\
bigger\langle \nabla_D \tilde{f}_i(\hat{D}^\nu_i, D^\nu_i, X^\nu_i) - \nabla_D \tilde{f}_i(\hat{D}^\nu_i, D^\nu_i, X^\nu_i), \hat{D}^\nu_i - \hat{D}^\nu_i \rangle \geq \tau_{D,i}^\nu \|\hat{D}^\nu_i - \hat{D}^\nu_i\|_F^2,
\end{align*}$$

where the second inequality follows from the $\tau_{D,i}^\nu$-strong convexity of $\tilde{f}_i(\bullet; D^\nu_i, X^\nu_i)$ [cf. Remark 6. To bound the first term on the LHS of the above inequality, let us use the expression (19) of $\tilde{f}_i$, and write

$$\begin{align*}
\nabla_D \tilde{f}_i(\hat{D}^\nu_i, D^\nu_i, X^\nu_i) - \nabla_D \tilde{f}_i(\hat{D}^\nu_i, D^\nu_i, X^\nu_i) \\
= \nabla_D \tilde{f}_i(D^\nu_i, X^\nu_i) - \nabla_D \tilde{f}_i(D^\nu_i, X^\nu_i) + \tau_{D,i}^\nu \left( D^\nu_i - \hat{D}^\nu_i \right) + \left( \tau_{D,i}^\nu - \tau_{D,i}^\nu \right) \left( \hat{D}^\nu_i - \hat{D}^\nu_i \right).
\end{align*}$$

Substituting (77) in (76) and invoking the Lipschitz continuity bound (56), we get

$$\begin{align*}
\|\hat{D}^\nu_i - \hat{D}^\nu_i\|_F \leq & \frac{\tau_{D,i}^\nu - \tau_{D,i}^\nu}{\tau_{D,i}^\nu} \|\hat{D}^\nu_i - \hat{D}^\nu_i\|_F + \frac{\tau_{D,i}^\nu + IL\nu}{\tau_{D,i}^\nu} \|D^\nu_i - D^\nu_i\|_F \\
& + \frac{IL\nu}{\tau_{D,i}^\nu} \sum_{j=1}^I \|X^\nu_j - X^\nu_j\|_F.
\end{align*}$$

By the compactness of $\mathcal{D}$, we have $\|\hat{D}^\nu_i - \hat{D}^\nu_i\|_F \leq B_D$, for some finite $B_D > 0$. Furthermore, by Assumption D2, we infer that 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 $\nu$ and $i$. Using the above facts and summing (78) over $i$, we get

$$\begin{align*}
\|\hat{D}^\nu_i - \hat{D}^\nu_i\|_F \leq & \frac{B_D}{\tilde{s}_D} \sum_{i=1}^I \tau_{D,i}^\nu - \tau_{D,i}^\nu \\
& + \frac{(\tilde{s}_D^{-1} + IL\nu)\sqrt{I}}{\tilde{s}_D} \|D^\nu_i - D^\nu_i\|_F + \frac{IL\nu\sqrt{I}}{\tilde{s}_D} \|X^\nu_i - X^\nu_i\|_F, \\
\end{align*}$$

which proves the desired result (68), with $L_D = \max(B_D, (\tilde{s}_D^{-1} + IL\nu)\sqrt{I}, IL\nu\sqrt{I})/\tilde{s}_D$. 

44
(d) We prove (69) when $\tilde{h}_i$ is given by (20); we leave the proof for the choice (11) of $\tilde{h}_i$ to the reader, since it is almost identical to that for the case of (20).

Invoking optimality of each $X_i^\nu$ and $X_i^{\nu+1}$ defined in (10) while using the strong convexity of $h_i(\bullet; U_{(i)}^\nu, X_i^\nu)$ and $g_i$’s, it is not difficult to show that the following holds:

\[
(\tau_{i,j}^\nu + \mu_i) \|X_i^{\nu+1} - X_i^\nu\|^2_F \leq \\
\left\langle \nabla_X \tilde{h}_i(X_i^\nu; U_{(i)}^\nu), X_i - \nabla_X \tilde{h}_i(X_i^\nu; U_{(i)}^\nu, X_i^{\nu-1}), X_i - X_i^\nu \right\rangle.
\]

Using (20), the definition $\Upsilon_i^\nu(X_i) \triangleq \nabla X \tilde{h}_i(X_i^\nu; U_{(i)}^\nu, X_i)$ (see the proof of Lemma 7), and the Cauchy-Schwarz inequality, the above inequality yields

\[
(\tau_{i,j}^\nu + \mu_i) \|X_i^{\nu+1} - X_i^\nu\|^2_F \leq \|\Upsilon_i^\nu(X_i^\nu) - \Upsilon_i^\nu(X_i^{\nu-1})\|_F \|X_i^{\nu+1} - X_i^\nu\|_F \\
+ \|\nabla X_i f_i(U_{(i)}^\nu, X_i^\nu) - \nabla X_i f_i(U_{(i)}^\nu, X_i^{\nu-1})\|_F \|X_i^{\nu+1} - X_i^\nu\|_F \\
+ |\tau_{i,j}^\nu - \tau_{i,j}^{\nu-1}| \|X_i - X_i^{\nu-1}\|_F \|X_i^{\nu+1} - X_i^\nu\|_F.
\]

(79)

Following the same steps used to prove (54), it is not difficult to check that, under Assumption D2, $\|\Upsilon_i^\nu(X_i^\nu) - \Upsilon_i^\nu(X_i^{\nu})\|_F \leq \tau_{i,j}^\nu \|X_i - X_i^\nu\|_F$. Using in (79) this bound together with the Lipschitz continuity of $\nabla X_i f_i(\bullet, X_i^\nu)$ [cf. Remark 8] and summing over $i$, yield

\[
\|X_i^{\nu+1} - X_i^\nu\|^2_F = \sum_{i=1}^I \|X_i^{\nu+1} - X_i^\nu\|^2_F \\
\leq \sum_{i=1}^I \frac{\tau_{i,j}^\nu + |\tau_{i,j}^\nu - \tau_{i,j}^{\nu-1}|}{\tau_{i,j}^\nu + \mu_i} \|X_i^\nu - X_i^{\nu-1}\|_F \|X_i^{\nu+1} - X_i^\nu\|_F \\
+ \sum_{i=1}^I \frac{L_{\nabla}}{\tau_{i,j}^\nu + \mu_i} \|U_{(i)}^\nu - U_{(i)}^{\nu-1}\|_F \|X_i^{\nu+1} - X_i^\nu\|_F.
\]

(80)

Define

\[
p_X^\nu \triangleq \max_i \frac{\tau_{i,j}^\nu + |\tau_{i,j}^\nu - \tau_{i,j}^{\nu-1}|}{\tau_{i,j}^\nu + \mu_i}, \text{ and } q_X^\nu \triangleq \max_i \frac{L_{\nabla}}{\tau_{i,j}^\nu + \mu_i}.
\]

(81)

Note that $0 < p_X^\nu, q_X^\nu < \infty$. Then, (80) becomes

\[
\|X_i^{\nu+1} - X_i^\nu\|^2_F \leq p_X^\nu \sum_{i=1}^I \|X_i^\nu - X_i^{\nu-1}\|_F \|X_i^{\nu+1} - X_i^\nu\|_F + q_X^\nu \sum_{i=1}^I \|U_{(i)}^\nu - U_{(i)}^{\nu-1}\|_F \|X_i^{\nu+1} - X_i^\nu\|_F \\
\leq p_X^\nu \|X_i^\nu - X_i^{\nu-1}\|_F \|X_i^{\nu+1} - X_i^\nu\|_F + q_X^\nu \|U_{(i)}^\nu - U_{(i)}^{\nu-1}\|_F \|X_i^{\nu+1} - X_i^\nu\|_F,
\]

where in the last inequality we used $\sum_i a_i b_i \leq \|a\| \cdot \|b\|$. Therefore,

\[
\|X_i^{\nu+1} - X_i^\nu\|_F \leq p_X^\nu \|X_i^\nu - X_i^{\nu-1}\|_F + q_X^\nu \|U_{(i)}^\nu - U_{(i)}^{\nu-1}\|_F.
\]

(82)

If, in addition, Assumption D3 holds, then it follows from (81) that there exists a sufficiently large $\nu_X > 0$ such that $p_X^\nu \in (0, \delta_0)$, for all $\nu \geq \nu_X$ and some $\delta_0 \in (0, 1)$.
A.3 Proof of Theorem 2

We are now ready to prove Theorem 2.

Statement (b) follows readily from Proposition 9:
\[
\lim_{\nu \to \infty} \|D' - 1 \otimes \overline{D}'\| = \lim_{\nu \to \infty} \|D' - 1 \otimes \overline{D}'\| + \lim_{\nu \to \infty} \sqrt{\nu} \\|\overline{D}' - \overline{D}'\| \leq 2 \lim_{\nu \to \infty} \|D' - 1 \otimes \overline{D}'\| \to 0.
\]

We prove next statements (a) and (a') throughout the following steps: 1) we first show that \(\lim \inf_{\nu \to \infty} \|\overline{D}' - D'\| = 0\); 2) we then prove \(\lim \inf_{\nu \to \infty} \|\overline{D}' - D'\| = 0\); and 3) finally we show that the limit point of any subsequence \(\{(\overline{D}', X')\}_{\nu \in Z}\) satisfying \(\lim_{\nu \to \infty} \|\overline{D}' - D'\| = 0\), is a stationary solution of \(P\).

Step 1: \(\lim \inf_{\nu \to \infty} \|\overline{D}' - D'\| = 0\).

We begin studying the descent properties of \(U\) along the sequence \(\{(\overline{D}'_{\nu}, X')\}_{\nu}\). We preliminary observe that
\[
G(\overline{D}'_{\nu+1}) \leq G(\overline{D}'_{\nu}) + \frac{\gamma'}{T} \sum_{i=1}^{I} \phi'_{i} \left( G(\overline{D}'(i)) - G(\overline{D}'_{\nu})) \right) + \frac{\gamma'}{T} \sum_{i=1}^{I} \phi'_{i} \left( G(D'_{\nu}) - G(\overline{D}'_{\nu})) \right),
\]
due to (32) and the convexity of \(G\) [together with \(1^{T} \phi' = 1\)].

Invoking the descent lemma for \(f_{i}(\overline{U}'_{\nu}, \bullet)\) and using \(\overline{U}'_{\nu} = \overline{D}'_{\nu+1}\) [cf. (31)], we get: for sufficiently large \(\nu\), say \(\nu \geq \nu_{0}\),
\[
U(\overline{D}'_{\nu+1}, X'_{\nu+1}) \leq \sum_{i=1}^{I} \left\{ f_{i}(\overline{D}'_{\nu+1}, X'_{i}) + g_{i}(X'_{i}) \right\} + G(\overline{D}'_{\nu+1}) + \sum_{i=1}^{I} \left( g_{i}(X'_{\nu+1}) - g_{i}(X'_{i}) \right) \\
+ \sum_{i=1}^{I} \left\langle \nabla X_{i} f_{i}(D'_{\nu+1}, X'_{i}), X'_{\nu+1} - X'_{i} \right\rangle + \frac{1}{2} \sum_{i=1}^{I} L\nabla X_{i}(\overline{U}'_{\nu}) \|X'_{\nu+1} - X'_{i}\|^{2}
\]
\[
\leq U(\overline{D}'_{\nu+1}, X'_{\nu}) - \sum_{i=1}^{I} \left\{ \frac{\tau'_{X,i} - 1}{2} L\nabla X_{i}(\overline{U}'_{\nu}) \|X'_{\nu+1} - X'_{i}\|^{2} \right\} \\
+ L_X \|U' - 1 \otimes \overline{U}'_{\nu}\|_{F} \|X'_{\nu+1} - X'_{i}\|_{F}
\]
\[
\leq U(\overline{D}'_{\nu+1}, X'_{\nu}) - s_X \|X'_{\nu+1} - X'_{i}\|_{F}^{2} + L_X \|U' - 1 \otimes \overline{U}'_{\nu}\|_{F} \|X'_{\nu+1} - X'_{i}\|_{F},
\]
where in (a) we used Proposition 11(b); and in (b) \(s_X > 0\) is a constant such that \(\inf_{\nu \geq \nu_{0}} (\tau'_{X,i} - 1/2 L\nabla X_{i}(\overline{U}'_{\nu})) \leq s_X\), for all \(i = 1, \ldots, I\). Note that such a constant exists because of (65) and Assumption D2.

To lower bound \(U(\overline{D}'_{\nu+1}, X'_{\nu})\), we apply the descent lemma to \(F(\bullet, X'_{\nu})\). Recalling that \(\nabla F(\bullet, X'_{\nu})\) is Lipschitz continuous with constant \(L_{\nabla X}\) and using (32), we get
\[
U(\mathbf{D}_{\phi^{t+1}}, X^{t+1}) \leq F(\mathbf{D}_{\phi^{t}}, X^{t}) + \frac{\gamma^{\nu}}{I} \left( \nabla_{\mathbf{D}} F(\mathbf{D}_{\phi^{t}}, X^{t}), \sum_{i=1}^{I} \phi^{\nu}_{i} \left( \mathbf{D}^{\nu}_{(i)} - \mathbf{D}^{\nu}_{(i)} \right) \right) \\
+ \frac{L_{\nabla D}}{2} \left( \frac{\gamma^{\nu}}{I} \right)^{2} \left\| \sum_{i=1}^{I} \phi^{\nu}_{i} \left( \mathbf{D}^{\nu}_{(i)} - \mathbf{D}^{\nu}_{(i)} \right) \right\|_{F}^{2} + \sum_{i=1}^{I} g_{i}(X^{t}_{i}) + G(\mathbf{D}_{\phi^{t+1}}),
\]
\[
\leq U(\mathbf{D}_{\phi^{t}}, X^{t}) - \frac{S_{D} \cdot \gamma^{\nu}}{I} \left( \left\| \mathbf{D}^{\nu} - \mathbf{D}^{\nu} \right\|_{F} - \frac{I_{\epsilon_{\phi}^{2}}}{2S_{D}} T^{\nu} \right)^{2} \\
+ \eta \cdot \gamma^{\nu} \left( \frac{\left\| \mathbf{D}^{\nu} - \mathbf{D}^{\nu} \right\|_{F}}{I} \sum_{t=1}^{\nu} \rho^{\nu-t} \left\| X^{t} - X^{t-1} \right\|_{F} + \frac{I_{\epsilon_{\phi}^{2}}^{2} \gamma^{\nu}}{4S_{D}} (T^{\nu})^{2} \\
+ \frac{L_{\nabla D}}{2} \left( \frac{\gamma^{\nu}}{I} \right)^{2} \left\| \mathbf{D}^{\nu} - \mathbf{D}^{\nu} \right\|_{F}^{2} + \frac{\gamma^{\nu}}{I} \left\| \mathbf{D}^{\nu} - \mathbf{D}^{\nu} \right\|_{F} \right) \sum_{i=1}^{I} \phi^{\nu}_{i} \left( G(\mathbf{D}^{\nu}_{(i)}) - G(\mathbf{D}_{\phi^{t}}) \right),
\]
where in (a) we used (83) and Proposition 11,’a), and the Lipschitz continuity of \( G \) (\( G \) is thus locally Lipschitz continuous) and the compactness of \( D \); we denoted by \( L_{G} > 0 \) the Lipschitz constant.

Combining (84) with (85) and defining \( \tau_{D}^{\nu} \triangleq s_{D} - \gamma^{\nu} \frac{L_{\nabla D}^{2}}{2} \), we get: for \( \nu \geq \nu_{0} \),
\[
U(\mathbf{D}_{\phi^{t+1}}, X^{t+1}) \leq U(\mathbf{D}_{\phi^{t}}, X^{t}) - \tau_{D}^{\nu} \left\| \mathbf{D}^{\nu} - \mathbf{D}^{\nu} \right\|_{F}^{2} + L_{\nabla D} \left\| \mathbf{D}^{\nu} - \mathbf{D}^{\nu} \right\|_{F} \left| X^{\nu+1} - X^{\nu} \right|_{F} \\
- \frac{\tau_{D}^{\nu} \gamma^{\nu}}{I} \left( \left\| \mathbf{D}^{\nu} - \mathbf{D}^{\nu} \right\|_{F} - \frac{I_{\epsilon_{\phi}^{2}}}{2 \tau_{D}^{\nu}} T^{\nu} \right)^{2} + \frac{L_{G} \gamma^{\nu}}{4 \tau_{D}^{\nu}} (T^{\nu})^{2} \\
+ \frac{\eta \gamma^{\nu}}{I} \left\| \mathbf{D}^{\nu} - \mathbf{D}^{\nu} \right\|_{F} \sum_{t=1}^{\nu} \rho^{\nu-t} \left\| X^{t} - X^{t-1} \right\|_{F} + L_{G} \gamma^{\nu} \left\| \mathbf{D}^{\nu} - \mathbf{D}^{\nu} \right\|_{F}
\]
\[
\leq c_{5} \rho^{\nu} + \rho^{\nu-1} \sum_{t=1}^{\nu} \rho^{\nu-t} \left\| X^{t+1} - X^{t} \right\|_{F}
\]
where \( c_{5} \triangleq \rho^{-1} \left\| X^{t} - X^{0} \right\|_{F} \). Since \( \gamma^{\nu} \downarrow 0 \), there exists an integer \( \nu_{1} \geq \nu_{0} \) and some \( \tau_{D}^{\nu} \) such that \( \tau_{D}^{\nu} \geq \tau_{D} > 0 \), for all \( \nu \geq \nu_{1} \). Let \( \tilde{\nu} \) be any integer \( \nu \geq \nu_{1} \). Then, applying (86) recursively on \( \nu, \nu-1, \ldots, \tilde{\nu}+1, \tilde{\nu} \), and using the boundedness of \( \left\{ \left\| \mathbf{D}^{\nu} - \mathbf{D}^{\nu} \right\|_{F} \right\}_{\nu} \), we obtain
\[
U(\mathbf{D}_{\phi^{t+1}}, X^{t+1})
\]
\[
\leq U(\mathbf{D}_{\phi^{t}}, X^{t}) - s_{X} \sum_{l=\tilde{\nu}}^{\nu} \left\| X^{l+1} - X^{l} \right\|_{F}^{2} + L_{\nabla D} \sum_{l=\tilde{\nu}}^{\nu} \left\| U^{l} - 1 \otimes \mathbf{U}_{\phi^{t}} \right\|_{F} \left\| X^{l+1} - X^{l} \right\|_{F} \\
- \frac{\tau_{D}}{I} \sum_{l=\tilde{\nu}}^{\nu} \gamma^{l} \left( \left\| \mathbf{D}^{l} - \mathbf{D}^{l} \right\|_{F} - \frac{I_{\epsilon_{\phi}^{2}}}{2 \tau_{D}^{l}} T^{l} \right)^{2} + \frac{L_{G}^{2}}{4 \tau_{D}^{l}} \sum_{l=\tilde{\nu}}^{\nu} \gamma^{l} \left( T^{l} \right)^{2} \\
+ c_{6} \sum_{l=\tilde{\nu}}^{\nu} \gamma^{l} \rho^{l} + c_{7} \sum_{l=\tilde{\nu}}^{\nu} \gamma^{l} \rho^{l-t} \left\| X^{l+1} - X^{l} \right\|_{F} + L_{G} \sum_{l=\tilde{\nu}}^{\nu} \gamma^{l} \left\| \mathbf{D}^{l} - 1 \otimes \mathbf{D}_{\phi^{t}} \right\|_{F}
\]
for some finite constants \( c_{6}, c_{7} > 0 \). Using the boundedness of \( \left\{ \left\| X^{\nu+1} - X^{\nu} \right\|_{F} \right\}_{\nu} \) (cf. Lemma 7)\( - \left\| X^{\nu+1} - X^{\nu} \right\|_{F} \leq B_{X}, \) for all \( \nu \) and some \( B_{X} > 0 \)--we can bound the double-sum term on the RHS of (87) as
\[
\sum_{t=1}^{\nu} \sum_{l=\bar{\nu}}^{\nu} \nu^{l-t} \|X^{t+1} - X^t\|_F = \sum_{t=1}^{\nu} \sum_{l=\max(\nu, \bar{\nu})}^{\nu} \nu^{l-t} \|X^{t+1} - X^t\|_F \\
= \sum_{t=1}^{\bar{\nu}-1} \|X^{t+1} - X^t\|_F \sum_{t=\bar{\nu}}^{\nu} \nu^{l-t} + \sum_{t=\nu}^{\nu} \|X^{t+1} - X^t\|_F \sum_{t=\nu}^{\nu} \nu^{l-t} \\
\leq B_X \left( \max_{\nu \leq t \leq \nu} \nu^t \right) \frac{(1 - \rho^\nu)(1 - \rho^{\nu-\nu} + 1)}{(1 - \rho)^2} + \sum_{t=\nu}^{\nu} \|X^{t+1} - X^t\|_F \sum_{t=\nu}^{\infty} \nu^{l-t}, \tag{88}
\]

where in (a) we used the summability of \(\sum_{t=1}^{\nu} \nu^{l-t} \), and the following bound

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

Substituting (88) in (87), yields

\[
U(D_{\phi^{\nu+1}}, X^{\nu+1}) \leq U(D_{\phi^{\nu}}, X^{\nu}) - s_X \sum_{t=\bar{\nu}}^{\nu} \|X^{t+1} - X^t\|_F^2 \\
+ \sum_{t=\bar{\nu}}^{\nu} \left( c_7 \sum_{t=\nu}^{\infty} \nu^{l-t} + L_X \|U^t - 1 \otimes \mathcal{U}_{\phi}^t\|_F \right) \|X^{t+1} - X^t\|_F \\
- \frac{T_D}{I} \sum_{t=\bar{\nu}}^{\nu} \nu^t \left( \|D^t - D^t\|_F - \frac{I \epsilon_{\phi}}{2 T_D} T^t \right)^2 + \frac{I \epsilon_{\phi}^2}{4 T_D} \sum_{t=\nu}^{\nu} \nu^t \left( T^t \right)^2 \\
+ \left[ \frac{c_6}{1 - \rho^\nu(\rho^{\nu+1})} + \frac{c_7 B_X}{(1 - \rho)^2} \right] \left( \max_{\nu \leq t \leq \nu} \nu^t \right) + L_G \sum_{t=\nu}^{\nu} \nu^t \|D^t - 1 \otimes \mathcal{D}_{\phi}^t\|_F \\
= U(D_{\phi^{\nu}}, X^{\nu}) - \sum_{t=\bar{\nu}}^{\nu} Y^t + \sum_{t=\bar{\nu}}^{\nu} W^t + E^{\nu, \bar{\nu}}, \tag{89}
\]

where we denoted

\[
Y^t \triangleq s_X \left( \|X^{t+1} - X^t\|_F - \frac{Z^t}{2 s_X} \right)^2 + \frac{T_D}{I} \gamma^t \left( \|D^t - D^t\|_F - \frac{I \epsilon_{\phi}^2}{2 T_D} T^t \right)^2,
\]

\[
W^t \triangleq \frac{I \epsilon_{\phi}^2}{4 T_D} \gamma^t \left( T^t \right)^2 + \frac{1}{4 s_X} \left( Z^t \right)^2 + L_G \gamma^t \|D^t - 1 \otimes \mathcal{D}_{\phi}^t\|_F.
\]
Note that the sequences \( \{Z_l^\nu\}_l \), \( \{W_l^\nu\}_l \), and \( \{E_{\nu,\bar{\nu}}\}_\nu \), \( \bar{\nu} \) satisfy
\[
\lim_{\nu \to \infty} \sum_{l=0}^{\nu} (Z_l^\nu)^2 < \infty, \\
\lim_{\nu \to \infty} \sum_{l=0}^{\nu} W_l^\nu < \infty, \\
\lim_{\bar{\nu} \to \infty} \left( \lim_{\nu \to \infty} E_{\nu,\bar{\nu}} \right) = 0;
\]
where (90) follows from (59) [cf. Prop. 9], Assumption D1, and Lemma 3(b) [cf. (49)]; (91) is a consequence of \( \lim_{\nu \to \infty} \sum_{l=0}^{\nu} \gamma_l (T_l^\nu)^2 < \infty \) [due to Prop. 11 (a)], (58) [cf. Prop. 9], and (90); and eq. (92) is proved by inspection.

It follows from (89), (90)-(92), and the coercivity of \( U \) that \( \{U(D_{\nu}, X_{\nu})\}_\nu \) is convergent. Indeed, taking the limsup of the LHS of (89) and using (91) and (92), we get
\[
\limsup_{\nu \to \infty} U(D_{\phi_{\nu+1}}, X_{\nu+1}) \leq U(D_{\phi}, X_{\nu}) + \sum_{l=0}^{\nu} W_l^\nu + E_{\nu,\bar{\nu}} < \infty.
\]
Additionally, taking the liminf of the RHS of the above inequality with respect to \( \bar{\nu} \) while using (92) and \( \lim_{\bar{\nu} \to \infty} \sum_{l=0}^{\nu} W_l^\nu = 0 \), yields
\[
\limsup_{\nu \to \infty} U(D_{\phi_{\nu+1}}, X_{\nu+1}) \leq \liminf_{\bar{\nu} \to \infty} U(D_{\phi}, X_{\nu}) < \infty,
\]
which implies the convergence of \( \{U(D_{\phi}, X_{\nu})\}_\nu \) to a finite value (since \( U \) is coercive), and
\[
\lim_{\nu \to \infty} \sum_{l=0}^{\nu} \left( \frac{||X_{\nu+1}^l - X_l^\nu||_F - \frac{Z_l^\nu}{2s_X}}{2s_X} \right)^2 < \infty, \\
\lim_{\nu \to \infty} \sum_{l=0}^{\nu} \gamma_l \left( \frac{||D_{\nu} - D_l^\nu||_F - \frac{I_{\bar{\nu}} T_l^\nu}{2r_D}}{2r_D} \right)^2 < \infty.
\]
It follows from (93) together with \( \lim_{\nu \to \infty} Z_{\nu} = 0 \) [due to (90)] that
\[
\lim_{\nu \to \infty} ||X_{\nu+1}^l - X_l^\nu||_F = 0.
\]
Similarly, (94) and Assumption D1 together with \( \lim_{\nu \to \infty} T_{\nu} = 0 \) [cf. Prop. 11 (a)] lead to
\[
\liminf_{\nu \to \infty} ||D_{\nu} - D_{\nu}||_F = 0.
\]

**Step 2:** \( \lim_{\nu \to \infty} ||\tilde{D}_{\nu} - D_{\nu}||_F = 0. \)

It is sufficient to show \( \limsup_{\nu \to \infty} ||\tilde{D}_{\nu} - D_{\nu}||_F = 0. \). To do so, we will use the following intermediate results.

**Lemma 12** Let \( \{(D_{\nu}, X_{\nu})\}_\nu \) be the sequence generated by the \( D^4L \) Algorithm; then,
(a) Vanishing gradient-tracking error:

\[
\lim_{\nu \to \infty} \left\| \tilde{\Theta}^\nu - 1 \otimes \frac{1}{I} \sum_{i=1}^{I} \nabla D f_i(D\delta, X_i^\nu) \right\|_F = 0,
\]

(b) Vanishing best-response error:

\[
\lim_{\nu \to \infty} \left\| \tilde{D}^\nu - \tilde{D}^\nu \right\|_F = 0.
\]

**Proof** (a) It follows readily from (74), using \( T^\nu \to 0, \) (95), and Lemma 3.

(b) Using i) the optimality of \( \tilde{D}^\nu \) and \( \tilde{D}^\nu \) defined in (6) and (8), respectively; ii) the convexity of \( G; \) and iii) the strongly convexity of \( \frac{\tilde{f}_i(\bullet; D^\nu, X_i^\nu)}{||D^\nu - 1 \otimes D\delta||_F \to 0}, \) it is not difficult to show that the following holds:

\[
\left\| \tilde{D}^\nu_{(i)} - \tilde{D}^\nu_{(i)} \right\|_F \leq \frac{I}{\tau_{D,i}^\nu} \left\| \tilde{\Theta}^\nu_{(i)} - \frac{1}{I} \sum_{j=1}^{I} \nabla D f_j(D\delta, X_j^\nu) \right\|_F \]

\[
\leq \frac{I}{\tau_{D,i}^\nu} \left\| \tilde{\Theta}^\nu_{(i)} - \frac{1}{I} \sum_{j=1}^{I} \nabla D f_j(D\delta, X_j^\nu) \right\|_F + \frac{1}{\tau_{D,i}^\nu} \left\| \sum_{j=1}^{I} \nabla D f_j(D_{(i)}, X_j^\nu) - \frac{1}{I} \sum_{j=1}^{I} \nabla D f_j(D\delta, X_j^\nu) \right\|_F \]

\[
\leq \frac{I}{\tau_{D,i}^\nu} \left\| \tilde{\Theta}^\nu_{(i)} - \frac{1}{I} \sum_{j=1}^{I} \nabla D f_j(D\delta, X_j^\nu) \right\|_F + \frac{L\nabla \sqrt{T}}{\tau_{D,i}^\nu} ||D^\nu - 1 \otimes D\delta||_F \to 0,
\]

where (a) is due to the Lipschitz continuity of \( \nabla D f_i \)’s [cf. (56)]; and (b) follows from \( \inf_{\nu} \tau_{D,i}^\nu > 0, \) for all \( i \) [cf. Assumption D2], (97) and (57) [cf. Proposition 9].

We are now ready to prove \( \lim \sup_{\nu \to \infty} \left\| \tilde{D}^\nu - D^\nu \right\|_F = 0. \) For notational simplicity, let us define \( \Delta \tilde{D}^\nu \triangleq \tilde{D}^\nu - D^\nu. \) Suppose by contradiction that \( \lim \sup_{\nu \to \infty} \left\| \Delta \tilde{D}^\nu \right\|_F > 0; \) since \( \lim \inf_{\nu \to \infty} \left\| \Delta \tilde{D}^\nu \right\|_F = 0, \) there exists \( \delta > 0 \) such that \( \left\| \Delta \tilde{D}^\nu \right\|_F > 2\delta \) and \( \left\| \Delta \tilde{D}^\nu \right\|_F < \delta \) for infinitely many \( \nu \)’s. Therefore, one can find an infinite subset of indexes, denoted by \( \mathcal{K}, \) having the following properties: for any \( \nu \in \mathcal{K}, \) there exists an index \( i_{\nu} > \nu \) such that

\[
\left\| \Delta \tilde{D}^\nu \right\|_F < \delta, \quad \left\| \Delta \tilde{D}^i_{\nu} \right\|_F > 2\delta, \quad \delta \leq \left\| \Delta \tilde{D}^i \right\|_F \leq 2\delta, \quad \nu < j < i_{\nu}.
\]

Let \( \nu_2 \) be a sufficiently large integer such that (89) holds and \( T^\nu < \frac{2\nu \delta}{T_i^\nu}, \) for all \( \nu \geq \nu_2 \) [such \( \nu_2 \) exists since \( T^\nu \to 0, \) see Proposition 11(a)]. Note that there exits a \( \bar{\delta} \) such that \( \delta - \frac{R_i^\nu T^\nu}{T_i^\nu} \geq \bar{\delta}, \) for all \( \nu \geq \nu_2. \) Choose \( \mathcal{K} \ni \nu \geq \nu_2; \) using (89), with \( \nu = i_{\nu} \) and \( \bar{\nu} = \nu + 1, \) yields

\[
U(D_{\delta i_{\nu} + 1}, X^{i_{\nu} + 1}) \leq U(D_{\delta i_{\nu} + 1}, X^{i_{\nu} + 1}) - c_8 \delta^2 \sum_{l=\nu+1}^{i_{\nu}} \gamma^{l} + \sum_{l=\nu+1}^{i_{\nu}} W^{l} + E^{i_{\nu}, \nu + 1},
\]
for some finite constant $c_8 > 0$. Using the convergence of $\{U(D_{\phi^\nu}, X^\nu)\}_\nu$, $\sum_{i=1}^{\infty} W^i < \infty$ [cf. (91)], and $\lim_{K \to \infty} E_{i^\nu, \nu+1} = 0$ [cf. (92)], inequality (101) implies

$$\lim_{K \to \infty} \sum_{i=\nu+1}^{\infty} \gamma^i = 0. \quad (102)$$

We show next that (102) leads to a contradiction.

It follows from (99) and (100) that, for all $K \ni \nu \geq \nu_2$,

$$\delta < \|\Delta \hat{D}^i\|_F - \|\Delta \hat{D}^\nu\|_F$$

$$(a) \leq \|\hat{D}^\nu - D^\nu - \hat{D}^i + \hat{D}^\nu\|_F + \|(|\tilde{D}^\nu - \hat{D}^\nu| - (\hat{D}^i - \hat{D}^\nu)|\|_F$$

$$\leq \|\hat{D}^i - \hat{D}^\nu\|_F + \|D^\nu - D^i\|_F + \|\tilde{D}^\nu - \hat{D}^\nu\|_F + \|\hat{D}^\nu - \hat{D}^\nu\|_F,$$

$$(b) \leq (L_D + 1)\|D^\nu - D^i\|_F + L_D |X^i - X^\nu|_F + L_D \sum_{t=1}^{T} |\tau^i_{D,i} - \tau^\nu_{D,i}|$$

$$+ \|\hat{D}^i - \tilde{D}^\nu\|_F + \|\tilde{D}^\nu - \hat{D}^\nu\|_F$$

$$(c) \leq (L_D + 1) \sqrt{\|D_{\phi^\nu} - \hat{D}_{\phi^\nu}\|_F + L_D |X^i - X^\nu|_F + \tilde{E}_{i^\nu, \nu}.$$}

where in (a) we used the reverse triangle inequality; (b) follows from Proposition 11(c); and in (c) we added/subtracted $1 \otimes \hat{D}_{\phi^\nu}$ and $1 \otimes \hat{D}_{\phi^\nu}$ in $\|D^\nu - D^i\|$, and defined

$$\tilde{E}_{i^\nu, \nu} \triangleq (L_D + 1) \left( \|D^\nu - 1 \otimes \hat{D}_{\phi^\nu}\|_F + \|D^i - 1 \otimes \hat{D}_{\phi^\nu}\|_F \right)$$

$$+ L_D \sum_{t=1}^{T} \sum_{i=\nu}^{i-1} |\tau^i_{D,i} - \tau^\nu_{D,i}| + \|\hat{D}^i - \tilde{D}^\nu\|_F + \|\tilde{D}^\nu - \hat{D}^\nu\|_F.$$
Rewrite first \( \| U^t - U^{t-1} \|_F \) as

\[
\| U^t - U^{t-1} \|_F \leq \| U^t - 1 \otimes U_{\phi^t} \|_F + \| U^{t-1} - 1 \otimes U_{\phi^{t-1}} \|_F + \sqrt{T} \| \phi^t \Delta \bar{D}^t - \bar{D}^t \|_F,
\]

where (a) follows from (31) and (32). Since \( \| \sum_{i=1}^{i_v} \phi_i^t (\bar{D}_{(i)}^t - D_{(i)}^t) \|_F \) is bounded (due to \( \phi_i^t \leq \bar{c}_\phi \) and compactness of \( D \)) and \( \lim_{K \to \infty} \sum_{i=1}^{i_v} \phi_i^t \| D_{(i)}^t - D_{(i)}^t \|_F = 0 \) [cf. (102)], there holds

\[
\lim_{K \to \infty} \sum_{i=1}^{i_v} \phi_i^t \| D_{(i)}^t - D_{(i)}^t \|_F = 0.
\]

Therefore, to prove \( \lim_{K \to \infty} \sum_{i=1}^{i_v} \| U^t - U^{t-1} \|_F = 0 \), it is sufficient to show that \( \lim_{K \to \infty} \sum_{i=1}^{i_v} \| U^t - 1 \otimes U_{\phi^t} \|_F = 0 \) which implies also \( \lim_{K \to \infty} \sum_{i=1}^{i_v} \| U^{t-1} - 1 \otimes U_{\phi^{t-1}} \|_F = 0 \), due to \( \lim_{K \to \infty} \| U^t - 1 \otimes U_{\phi^t} \|_F = 0 \), see (59). By (64), the boundedness of \( \{ \bar{D}^\nu \}_\nu \), and (102), it is sufficient to show that \( \lim_{K \to \infty} \sum_{i=1}^{i_v} \| D^t - 1 \otimes \bar{D}_{\phi^t} \|_F = 0 \). We have

\[
\lim_{K \to \infty} \sum_{i=1}^{i_v} \| D^t - 1 \otimes \bar{D}_{\phi^t} \|_F \leq \sum_{i=1}^{i_v} \left( c_1 \rho^t + c_2 \sum_{l=0}^{l-1} \gamma^t \rho^{l-t} \right)
\]

\[
= c_2 \lim_{K \to \infty} \sum_{i=1}^{i_v} \sum_{t=0}^{i_v-1} \gamma^t \rho^{l-t} = c_2 \lim_{K \to \infty} \sum_{i=1}^{i_v} \sum_{l=\max(t+1,t)}^{i_v-1} \gamma^t \rho^{l-t}
\]

\[
\leq c_2 \lim_{K \to \infty} \sum_{i=1}^{i_v} \sum_{t=0}^{i_v-1} \gamma^t \rho^{l-t} + c_2 \lim_{K \to \infty} \sum_{i=1}^{i_v} \sum_{t=1}^{i_v-1} \gamma^t \rho^{l-t}
\]

\[
\leq c_2 \left( \lim_{K \to \infty} \sum_{i=1}^{i_v} \gamma^t \rho^{l-t} \right) = 0.
\]

This proves \( \lim_{\nu \to \infty} \| X^\nu - X^{\nu} \|_F = 0 \) and thus \( \lim_{\nu \to \infty} \bar{E}^{i_v}\nu = 0 \).

We can now prove that (102) leads to a contradiction. Since \( \bar{E}^{i_v}\nu \to 0 \) as \( \nu \to \infty \), there exists a sufficiently large integer \( \nu_3 \in \mathbb{K} \), such that \( \nu_3 > \nu_2 \) and \( \bar{E}^{i_v}\nu < \delta \), for all \( \nu > \nu_3 \). Define \( \delta' \) such that \( 0 < \delta' \leq \delta - \bar{E}^{i_v}\nu \). Using (32) and \( 1^T \phi^\nu = I \), (103) implies

\[
\frac{\delta'}{(L_D + 1)\sqrt{T}} < \sum_{t=\nu}^{i_v-1} \gamma^t \| \Delta \bar{D}^t \|_F \leq 2\delta \sum_{t=\nu}^{i_v-1} \gamma^t,
\]

for all \( \mathbb{K} \ni \nu > \nu_3 \), where the last inequality follows from (99) and (100). Equation (104) contradicts (102). Hence, it must be \( \limsup_{\nu \to \infty} \| \Delta \bar{D}^\nu \|_F = 0 \), and thus

\[
\lim_{\nu \to \infty} \| \bar{D}^\nu - D^\nu \|_F = 0.
\]
Step 3: Stationarity of the limit points.

Let \((\mathcal{D}^\infty, \mathbf{X}^\infty)\) be the limit point of the subsequence \(\{(\mathcal{D}^\nu, \mathbf{X}^\nu)\}_{\nu \in \mathcal{Z}}\) satisfying \(\lim_{\mathcal{Z}\ni \nu \to \infty} ||\mathcal{D}^\nu - \mathcal{D}^\nu||_F = 0\); and let \(\mathcal{D}_{\phi}^\infty = \lim_{\mathcal{Z}\ni \nu \to \infty} \mathcal{D}_{\phi}^\nu\). The optimality of \(\mathbf{X}_{i}^{\nu+1}\) [cf. (10)] implies

\[
\tilde{h}_i(\mathbf{X}_{i}^{\nu+1}, \mathbf{U}_{(i)}^{\nu}, \mathbf{X}_{i}^{\nu}) + g_i(\mathbf{X}_{i}^{\nu+1}) \leq \tilde{h}_i(\mathbf{X}_{i}; \mathbf{U}_{(i)}^{\nu}, \mathbf{X}_{i}^{\nu}) + g_i(\mathbf{X}_{i}), \quad \forall \mathbf{X}_{i} \in \mathcal{X}_i,
\]

for all \(\mathbf{X}_{i} \in \mathcal{X}_i\). Taking the limit \(\mathcal{Z}\ni \nu \to \infty\) and using \(\mathbf{U}_{(i)}^\infty \equiv \lim_{\mathcal{Z}\ni \nu \to \infty} \mathbf{U}_{(i)}^\nu\) [due to (9), (105), and Proposition 9], \(\lim_{\mathcal{Z}\ni \nu \to \infty} ||\mathbf{X}_{i}^{\nu+1} - \mathbf{X}_{i}||_F = 0\), and the continuity of \(\tilde{h}_i\) and \(g_i\), yields

\[
\tilde{h}_i(\mathbf{X}_{i}^\infty, \mathcal{D}_{\phi}^\infty, \mathbf{X}_{i}^\infty) + g_i(\mathbf{X}_{i}^\infty) \leq \tilde{h}_i(\mathbf{X}_{i}; \mathcal{D}_{\phi}^\infty, \mathbf{X}_{i}^\infty) + g_i(\mathbf{X}_{i}), \quad \forall \mathbf{X}_{i} \in \mathcal{X}_i,
\]

which, together with \(\nabla_{\mathbf{X}_i} \tilde{h}_i(\mathbf{X}_{i}^\infty, \mathcal{D}_{\phi}^\infty, \mathbf{X}_{i}^\infty) = \nabla_{\mathbf{X}_i} f_i(\mathbf{X}_{i}^\infty)\) [cf. Remark 6], yields

\[
\langle \nabla_{\mathbf{X}_i} f_i(\mathcal{D}_{\phi}^\infty, \mathbf{X}_{i}^\infty) + \Xi_{i}^\infty, \mathbf{X}_{i} - \mathbf{X}_{i}^\infty \rangle \geq 0, \quad \forall \mathbf{X}_{i} \in \mathcal{X}_i,
\]

for some \(\Xi_{i}^\infty \in \partial g_i(\mathbf{X}_{i}^\infty)\). Using similar arguments together with the optimality of \(\mathcal{D}_{\phi}^\nu\) in (8), the continuous differentiability of \(f_i\)’s and \(\tilde{f}_i\)’s, (57), and (97) [cf. Lemma 12(a)], one can show that

\[
0 \leq \left\langle \nabla_{\mathbf{D}_i} \tilde{f}_i(\mathcal{D}_{\phi}^\infty, \mathbf{X}_{i}^\infty) + \sum_{j \neq i} \nabla_{\mathbf{D}_j} f_j(\mathcal{D}_{\phi}^\infty, \mathbf{X}_{j}^\infty) + \Omega_{i}^\infty, \mathbf{D}_{(i)} - \mathcal{D}_{\phi}^\infty \right\rangle
\]

\[
= \left\langle \nabla_{\mathbf{D}_i} F(\mathcal{D}_{\phi}^\infty, \mathbf{X}_{i}^\infty) + \Omega_{i}^\infty, \mathbf{D}_{(i)} - \mathcal{D}_{\phi}^\infty \right\rangle, \quad \forall \mathbf{D}_{(i)} \in \mathcal{D}, \quad \text{with} \ i = 1, \ldots, I, \quad \text{and some} \ \Omega_{i}^\infty \in \partial G(\mathcal{D}_{\phi}^\infty). \quad \text{By} \ (107) \ \text{and} \ (108), \ \text{it follows that} \ (\mathcal{D}_{\phi}^\infty, \mathbf{X}_{i}^\infty) \ \text{is a stationary solution of Problem} \ \text{P and, by} \ (57), \ \text{so is} \ (\mathcal{D}_{\phi}^\infty, \mathbf{X}_{i}^\infty).
\]

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.

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, Y. Sun, G. Scutari, and F. Facchinei. D2L: Decentralized dictionary learning over dynamic networks. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 4084–4088, March 2017.

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.

A. G. Dimakis, S. Kar, J. M. F. Moura, M. G. Rabbat, and A. Scaglione. Gossip algorithms for distributed signal processing. Proceedings of the IEEE, 98(11):1847–1864, November 2010.

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.

B. Gharesifard and J. Cortés. When does a digraph admit a doubly stochastic adjacency matrix? In Proceedings of the 2010 American Control Conference, pages 2440–2445, Baltimore, Maryland, USA, June 2010.

T. Hastie, R. Tibshirani, and M. Wainwright. Statistical Learning with Sparsity. CRC Press, Taylor & Francis Group, 2015.
M. Hong. Decomposing linearly constrained nonconvex problems by a proximal primal dual approach: Algorithms, convergence, and applications. *arXiv:1604.00543*, November 2016.

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 *Proceedings 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 and A. Olshevsky. Distributed optimization over time-varying directed graphs. *IEEE Transactions on Automatic Control*, 60(3):601–615, March 2015.

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, 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 2014.

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.

Ali H. Sayed. Adaptation, learning, and optimization over networks. Foundations and Trends in Machine Learning, 7(4-5):311–801, 2014.

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. arXiv:1512.00895, December 2016.

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.

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.

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.