Distributed iterative thresholding for 0/1-regularized linear inverse problems

Distributed iterative thresholding for ℓ_0/ℓ_1-regularized linear inverse problems

C. Ravazzi, S.M. Fosson, and E. Magli
Department of Electronics and Telecommunications,
Politecnico di Torino, Italy

***Abstract*—The ℓ_0/ℓ_1-regularized least-squares approach is used to deal with linear inverse problems under sparsity constraints, which arise in mathematical and engineering fields. In particular, multi-agent models have recently emerged in this context to describe diverse kinds of networked systems, ranging from medical databases to wireless sensor networks. In this paper, we study methods for solving ℓ_0/ℓ_1-regularized least-squares problems in such multi-agent systems. We propose a novel class of distributed protocols based on iterative thresholding and input driven consensus techniques, which are well-suited to work in-network when the communication to a central processing unit is not allowed. Estimation is performed by the agents themselves, which typically consist of devices with limited computational capabilities. This motivates us to develop low-complexity and low-memory algorithms that are feasible in real applications. Our main result is a rigorous proof of the convergence of these methods in regular networks. We introduce suitable distributed, regularized, least-squares functionals and we prove that our algorithms reach their minima, using results from dynamical systems theory. Furthermore, we propose numerical comparisons with the alternating direction method of multipliers and the distributed subgradient methods, in terms of performance, complexity, and memory usage. We conclude that our techniques are preferable for their good memory-accuracy tradeoff.**

***Index Terms*—Distributed optimization, input driven consensus algorithms, multi-agent systems, regularized linear inverse problems, sparse estimation.**

I. Introduction

Linear inverse problems arise in several areas of engineering and mathematical sciences. A standard linear inverse problem considers an affine system of the form

$$A\widetilde{x} = y \tag{1}$$

where $A \in \mathbb{R}^{m\times n}$ and $y \in \mathbb{R}^m$ are known, while $\widetilde{x} \in \mathbb{R}^n$ is the unknown vector to be estimated. The goal is to provide an estimate of $\widetilde{x}$ starting from the data (y, A). For this purpose, a widely used methodology is the least-squares (LS) approach [1]. However, in most applications, the problem in (1) is ill-conditioned or under-determined, namely the number of available measurements is smaller than the dimension of the model to be estimated; for this motivation, regularizing constraints are often added in order to obtain stable solutions.

In many practical situations, models are constrained structurally so that only few degrees of freedom compared to their ambient dimension are significant [2]. In the last decades, inverse problems under sparsity constraints have attracted an increasing attention, especially in statistics, signal processing, machine learning, and coding theory. Examples include sparse linear regressions [3], [4], approximation of functions [5], signal recovery in compressed sensing [6], [7], image denoising and restoration [8], and channel decoding [9]. The reader can refer to [10] and references therein for an overview of possible applications. Methods that seek approximate solutions to linear systems of the form (1) in which only the most relevant variables are chosen, have been developed in different areas. One of the most popular techniques is the Tikhonov regularization [11] in which a quadratic penalty is added to the LS function. However, this method leads to a solution which is generally nonsparse.

In this paper we consider two methods capable of providing a parsimonious estimate of solutions to (1): the ℓ_0 and the ℓ_1 regularized estimators [4]. The ℓ_0-regularized estimator [12] is defined as the minimizer of the cost function

$$\mathcal{J}_0(x) = \|y - Ax\|_2^2 + \frac{2\lambda}{\tau}\|x\|_0 \tag{2}$$

where $\lambda, \tau > 0$, and $\|x\|_0 = |\{i \in \{1,\dots,n\} | x_i \neq 0\}|$ counts the number of nonzero entries of x. It should be noted that the function in (2) is not convex and standard algorithms for convex optimization cannot be used [5]. The ℓ_1-regularized optimization problem relaxes the ℓ_0 penalty and replace $\|x\|_0$ with its convex envelope $\|x\|_1 = \sum_{i=1}^n |x_i|$. The ℓ_1-regularized estimator is the minimizer of the following nonsmooth convex function

$$\mathcal{J}_1(x) = \|y - Ax\|_2^2 + \frac{2\lambda}{\tau}\|x\|_1 \tag{3}$$

where $\lambda, \tau > 0$ is a parameter that controls the amount of sparsity. The minimizer of problem (3) is also known as the *least-absolute shrinkage and selection operator* (LASSO) estimator [13].

The literature describes a large number of approaches to estimate the minimizers of (2) and (3). Examples include quadratic programming methods [5], including interior-point methods [5], projected gradient methods [14], and iterative hard thresholding algorithms (IHTA, [15], [16]) and iterative soft thresholding algorithms (ISTA, [17]). Iterative thresholding algorithms have lower computational complexity per iteration and lower storage requirements than interior-point methods. Notice that these types of recursions are a modification of the gradient method to solve a linear system: the only difference consists in the application of a (hard or soft)

A preliminary version of some of the results has appeared in the proceedings of the IEEE Global Communications Conference 2013, held in Atlanta, GA, USA.

shrinkage operator, which promotes the sparsity of the estimate at each iteration.

In the inverse problems, the observed data are typically assumed to be centrally available, so that they can be jointly processed to minimize functions in (2) or (3). However, distributed inverse problems commonly arise in many applications, where data are inherently scattered across a large geographical area [18]. This scenario applies, for example, to sparse event detection in wireless networks [19], distributed indoor localization [20], and distributed tracking in sensor networks [21]. These problems consider a network with N nodes that individually store data (y_i, A_i), with $i \in \{1, \ldots, N\}$. In this setting, the most used paradigm can be summarized as follows: agents transmit $(y_i, A_i)_{i=1}^N$ to a central processing unit that performs joint estimation minimizing (2) or (3) with $A = (A_1^\mathsf{T}, \ldots, A_N^\mathsf{T})^\mathsf{T}$ and $y = (y_1^\mathsf{T}, \ldots, y_N^\mathsf{T})^\mathsf{T}$. A drawback of this model is that, particularly in large-scale networks, gathering all data to a central processing unit may be inefficient, as a large number of hops have to be taken, requiring a significant amount of energy for communication over the wireless channel. Moreover, this may also introduce delays, severely reducing the sensor network performance. In other applications, agents providing private data may not be willing to share them, but only to build the learning results. This occurs, for example, in classification [22], one of the fundamental problems in machine learning, or in data fitting in statistics [23], where the training data consists of sensitive information of agents (such as medical data, or data flow across the Internet). We also notice that parallel computing can be recasted in this distributed context. Let us consider, for example, the implementation of the algorithms that minimize (2) and (3) on GPUs [24], [25]. Even if the data are centrally available, in this situation we aim at decentralizing them on the GPU cores and threads in order to distribute the computational load and accelerate the estimation procedure.

A common element to many distributed applications is the necessity of limiting computations and memory usage, as the nodes generally consist of devices with scarce capabilities in this sense. An evident example is given by sensor networks: computations may require a lot of energy while sensors are generally endowed with small memories of the order of a few kB. Also in parallel computing this issue is crucial, as the memory is known to be a bottleneck to the system performance.

So far the distributed minimization of (2) and (3) has received less attention than the centralized problem and the literature is very recent [26], [27], [28], [18], [22]. These first contributions propose natural ways to distribute known centralized methods, and obtain interesting results in terms of convergence and estimation performance. However they do not consider the problem of the insufficient computation and memory resources. In particular, in [26] a distributed pursuit recovery algorithm is proposed, assuming that every node i knows the matrix A_i of every other agent. This estimation scheme is clearly unpractical in large-scale networks, since individual agents do not have the capacity to store and process a large number of these matrices. Distributed basis pursuit algorithms for sparse approximations when the measurement matrices are not globally known have been studied in [27], [28], [18]. In these algorithms, sensors collaborate to estimate the original vector, and, at each iteration, they update this estimate based on communication with its neighbors in the network. These methods are based on Distributed Subgradient Methods (DSM, [29]) or Alternating Direction Method of Multipliers (ADMM, [30]). DSM have low memory requirements but are extremely slow and, consequently, many transmission between nodes are needed in order to obtain a good estimate. ADMM is fast but requires the inversion of $n \times n$ matrices at each unit. Although the inversion of the matrices in ADMM can be performed off-line, the storage of a $n \times n$ matrix may be critical.

It is worth mentioning that efficient methods have been proposed recently to minimize an objective function composed of local convex and differentiable functions and a common nondifferentiable function as in (3). In particular, in [31] a distributed algorithm based on Nesterov acceleration techniques is developed and analyzed, under the assumption that the gradient of the differentiable component is bounded (see [31, Assumption 1.b]). This assumption is not satisfied for (3). Moreover, in [32] a modified function of (3) is considered, adding an ℓ_2 regularization which makes the problem solvable by the linearized Bregman algorithm. In [32], a decentralized version of the Bregman algorithm is studied and proved to reach a neighborhood of the unknown desired signal [32, Theorem 1]; nevertheless, no point convergence is guaranteed. Numerical results [32, Section V] show that a Nesterov accelerated variant of the analyzed algorithm is faster than ADMM based techniques.

In this paper, we propose distributed, consensus-based versions of (2) and (3) and we derive new algorithms to reach their minima. Our purpose is to develop low complexity techniques that require very little memory, in order to fill the feasibility gap left open by the previous works, while achieving performance as close as possible to that of the ADMM estimation [18].

As mentioned before, in the centralized case a good tradeoff between complexity and performance is obtained by iterative thresholding methods, which has motivated us to investigate a similar paradigm in the distributed scenario. In particular, we present a class of *distributed iterative thresholding algorithms* for ℓ_0/ℓ_1 regularized optimization problems. Our approach builds on the seminal work of Daubechies *et* al. [17] and of Blumensath *et* al. (see [15], [16]), who developed iterative thresholding methods for solving regularized optimization problems. Moreover, our work is related to the literature on distributed computation and estimation, which has attracted recent interest in the scientific community [33], [29], [34], and whose main goal is to design distributed iterative algorithms to cooperatively minimize a common cost function. The techniques that we introduce, in fact, work in virtue of their cooperative characteristics. We propose two different algorithms: *distributed iterative hard thresholding algorithm* (DIHTA) and *distributed iterative soft thresholding algorithm* (DISTA). They consist of a gradient step that seek to minimize the LS functional, a (hard or soft) thresholding step that promotes sparsity and, as a key ingredient, a consensus step

to share information among neighboring sensors.

Besides the design of the algorithms, our main contributions include (1) a rigorous proof of their convergence for regular networks, (2) the numerical evaluation of their performance, and (3) an analysis in terms of complexity and memory requirements. Our intuition is that these hypotheses on the network regularity, that are useful to prove the convergence of the proposed algorithms, are not really necessary.

As will be seen, the proposed methods achieve extremely good performance and there is no significant loss compared to the centralized implementations [15] and [35]. Extensive simulations show that DIHTA and DISTA are satisfactory in the following sense: when the product of the number of agents in the network times the number of data for each unit exceeds a given threshold, accurate estimation is achieved. Moreover, the total number of available data required for the estimation is comparable to that required by joint estimation. This implies that decentralization is not a drawback. We assess their performance on the basis of a number of numerical results, comparing with that of existing methods, such as DSM [36] and consensus ADMM [30]: compared to DSM, DISTA requires equal memory storage and communication cost but is extremely faster. This implies that the total number of communications is minimized. Indeed, DISTA is only slightly suboptimal with respect to consensus ADMM. On the other hand, it features a much lower memory usage, making it suitable also for low-energy environments such as wireless sensor networks. For what concerns communication, we will show that our methods are comparable with the existing ones. However, we mention that different variants on the proposed protocols are possible in order to further minimize the communications between agents. For example, asynchronous and randomized adaptations are subject of our current research. Theoretical contributions consist of the convergence proof of DIHTA and DISTA to the consensus-based estimator of (2) and (3), respectively. These results are obtained for regular networks. For both DIHTA and DISTA, we show that they reach the fixed points of the maps that rule their dynamics, which coincide with the minima of the consensus-based reformulated cost functions. Even if the algorithms' patterns are similar, the convergence proof is based on different mathematical tools: for DIHTA we use the LaSalle invariance principle, while for DISTA the Opial's Theorem. We notice that the mathematical tools used in the proof provide also a general framework in analyzing new distributed algorithms for linear inverse problems.

A. Outline of the paper

The paper is organized as follows. In Section II, the general linear inverse problem and the classical iterative thresholding algorithms for ℓ_0/ℓ_1-regularized LS problems are described. Section III addresses the problem of distributed estimation. In particular, the optimization problem is formulated in a separable form based on consensus techniques. In this way, the distributed iterative thresholding algorithms are developed and compared with related literature in Section IV. In Section V theoretical results on convergence are stated and proved. Numerical experiments with simulated and real data are presented in Section VI. Some concluding remarks (Section VII) and an Appendix collecting the most technical steps of the proof complete the paper.

B. Notation

Throughout this paper, we use the following notation. We denote column vectors with small letters, and matrices with capital letters. Given a matrix X, X^T denotes its transpose, $(X)_v$ (or x_v) denotes the v-th column of X, and $(X_{jv})_{j,v\in\mathcal{V}}$ (or $(x_{jv})_{j,v\in\mathcal{V}}$) are its entries. We consider $\mathbb{R}^n$ as a euclidean space endowed with the following norms:

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}$$

with $p = 1, 2$. Given $x \in \mathbb{R}^n$, we denote with

$$\|x\|_0 = \sum_{i=1}^{n} |x_i|^0,$$

where we use the convention $0^0 = 0$. For a rectangular matrix $M \in \mathbb{R}^{m\times n}$, we consider the Frobenius norm, which is defined as follows

$$\|M\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} M_{ij}^2} = \sqrt{\sum_{j=1}^{n} \|(M)_j\|_2^2},$$

and the operator norm

$$\|M\|_2 = \sup_{z\neq 0} \frac{\|Mz\|_2}{\|z\|_2}.$$

We denote the sign function as

$$\operatorname{sgn}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{otherwise.} \end{cases}$$

If x is a vector in $\mathbb{R}^n$, $\operatorname{sgn}(x)$ is intended as a function to be applied elementwise. If $f : \Omega \to \mathbb{R}$ is a real-valued convex function defined on a convex open set in the Euclidean space $\mathbb{R}^n$, a vector v in that space is called a subgradient at a point x_0 in Ω if for any $x \in \Omega$ the following inequality holds:

$$f(x) \geq f(x_0) + v \cdot (x - x_0).$$

An undirected graph is a pair $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ is the set of vertices (or nodes), and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of edges with the property $(i, j) \in \mathcal{E}$ implies $(j, i) \in \mathcal{E}$. In this paper, we use the convention that $(i, i) \in \mathcal{E}$ for all $i \in \mathcal{V}$. A path in a graph is a sequence of edges which connect a sequence of vertices. In an undirected graph $\mathcal{G}$, two vertices u and v are called connected if there exists a path from u to v. A graph is said to be connected if every pair of vertexes in the graph is connected. A graph is said to be regular when each vertex is connected to the same number of nodes. Using the convention that each node has a self-loop, we call d-regular a graph in which each vertex is connected to exactly $d-1$ nodes different from itself. A matrix with non-negative elements P is said to be stochastic if $\sum_{j\in\mathcal{V}} P_{ij} = 1$ for every $i \in \mathcal{V}$. Equivalently, P

is stochastic if $P\mathbb{1} = \mathbb{1}$. The matrix P is said to be adapted to a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ if $P_{v,w} = 0$ for all $(w, v) \notin \mathcal{E}$. We finally define the neighborhood of v as the set $\mathcal{N}_v$ that contains all $w \in \mathcal{V}$ such that $P_{v,w} \neq 0$. According to our notation, $v \in \mathcal{N}_v$ as $(v, v) \in \mathcal{E}$.

II. Preliminaries

A. Linear inverse problems under sparsity constraints

In our model, we consider a network represented by an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. The vertices $\mathcal{V}$ is the set of nodes and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of available communication links. We assume that each node $v \in \mathcal{V}$ has available local data $y_v \in \mathbb{R}^m$ and matrix $A_v \in \mathbb{R}^{m \times n}$. Nodes seek to find the vector $\widetilde{x} \in \mathbb{R}^n$ such that

$$y_v = A_v \widetilde{x}, \tag{4}$$

where the model parameters vector $\widetilde{x}$ is sparse (*i.e.*, it consists of a small number of nonzero elements). (y_v, A_v) are assumed to be available only at the v-th node and not sharable with other nodes. We remark that $\widetilde{x}$ couples all the nodes in the network.

If the data stored by the network were available at once in a single central processing unit that performs joint estimation, an approximation of $\widetilde{x}$ could be obtained solving the following ℓ_p-regularized optimization problem:

$$\min_{x \in \mathbb{R}^n} \mathcal{J}_p(x) := \min_{x \in \mathbb{R}^n} \sum_{v \in \mathcal{V}} \|y_v - A_v x\|_2^2 + \frac{2\lambda}{\tau} \|x\|_p, \tag{5}$$

with $p = 0$ or $p = 1$ for some $\lambda, \tau > 0$. It should be noted that for problems in which the number of variables n exceeds the number of observations m the function $\mathcal{J}_1(x)$ is not strictly convex, and hence it may not have a unique minimum. Sufficient conditions guaranteeing the uniqueness of the solution of (5) are derived for $p = 1$ in [37] and for $p = 0$ in [38]. We make the following assumption throughout the paper.

Assumption 1. *The problems in* (5) *with* $p \in \{0, 1\}$ *admit a unique solution.*

The solutions of the problems in (5) provide an approximation of the model parameters with a bounded error, which is controlled by λ (see [7], [39]). We now review the iterative thresholding algorithms, that have been recently proposed to solve (5), *e.g.*, [15], [16], [17].

B. Iterative thresholding algorithms

Popular approaches to solve the optimization problem in (5) are IHTA when $p = 0$ and ISTA for $p = 1$. These methods are based on moving at each iteration in the direction of the steepest descent and thresholding to promote sparsity [17].

Let us collect the training data in the vector $y = (y_1^\mathsf{T}, \dots, y_{\mathcal{V}}^\mathsf{T})^\mathsf{T}$ and $A = (A_1^\mathsf{T}, \dots, A_{\mathcal{V}}^\mathsf{T})^\mathsf{T}$. Given an initial estimate $x(0)$, the iterate for $t \in \mathbb{N}$ can be written as

$$x(t+1) = \eta_{p,\lambda}[x(t) + \tau A^\mathsf{T}(y - Ax)]$$

where $\tau > 0$ is the step-size in the direction of the steepest descent. The operator η is a thresholding function to be applied elementwise, *i.e.*

$$\eta_{0,\lambda}[x] = \begin{cases} x & \text{if } |x| > \sqrt{2\lambda} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

and

$$\eta_{1,\lambda}[x] = \begin{cases} \operatorname{sgn}(x)(|x| - \lambda) & \text{if } |x| > \lambda \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

The convergence of these algorithms was proved under the assumption that $\|A\|_2^2 < 1/\tau$ in [17] (for ISTA) and [16] (for IHTA). A dissertation about the convergence results can be found in [40].

III. Consensus-based reformulation of the ℓ_p-regularized LS problem

In this work, we design iterative algorithms to solve (5), in which the nodes only exchange information with their nearest neighbors at each iteration, without any central coordination. Before presenting the proposed protocols, we recast the optimization problem in (5) into a separable form, that facilitates distributed implementation. The goal is to split the problem into simpler subtasks that can be executed locally at each node.

From now on, we adopt the following assumption.

Assumption 2. $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ *is connected and* d*-regular.*

Let us replace the global variable x in (5) with local variables $\{x_v\}_{v \in \mathcal{V}}$, representing estimates of the model parameters $\widetilde{x}$, provided by each node. We rewrite the distributed problem as follows

$$\begin{aligned} &\min_{x_1, \dots, x_{|\mathcal{V}|} \in \mathbb{R}^n} \sum_{v \in \mathcal{V}} \left[\|y_v - A_v x_v\|_2^2 + \frac{2\lambda}{\tau |\mathcal{V}|} \|x_v\|_p \right], \\ &\text{s.t. } x_v = \overline{x}_w, \quad \forall w \in \mathcal{N}_v, \forall v \in \mathcal{V}, \end{aligned} \tag{8}$$

where

$$\overline{x}_w = \sum_{u \in \mathcal{V}} P_{wu} x_u$$

and P is adapted to the graph with $P_{u,v} = 1/d$ if $(u, v) \in \mathcal{E}$ and zero otherwise. Since P is adapted to the graph, it should be noted that

$$\overline{x}_w = \sum_{u \in \mathcal{N}_w} P_{wu} x_u.$$

The following should be easily guessed.

Proposition 1. *If* $\mathcal{G}$ *is a connected graph, then the optimization problems in* (5) *and* (8) *are equivalent, in the sense that any solution of* (5) *is a minimizer for* (8) *and vice versa.*

Proof. Since $v \in \mathcal{N}_v$, from the constraints in (8) we have $x_v = \overline{x}_v$ and $x_w = \overline{x}_v$ for all $w \in \mathcal{N}_v$. This implies by transitivity that $x_v = x_w$ for all $w \in \mathcal{N}_v$. If the graph is connected, then there exists a path connecting every pair of vertexes. We can conclude that $x_v = x$ for any $v \in \mathcal{N}_v$, in which case the cost function (8) reduces to the one in (5). □

We now relax the problem in (8). Let us consider the minimization of the functional $\mathcal{F}_p : \mathbb{R}^{n\times|\mathcal{V}|} \longmapsto \mathbb{R}^+$ defined as follows

$$\mathcal{F}_p(x_1,\dots,x_{|\mathcal{V}|}) := \sum_{v\in\mathcal{V}} \Bigg[q\|y_v - A_v x_v\|_2^2 + \frac{2q\lambda}{\tau|\mathcal{V}|}\|x_v\|_p + \frac{1-q}{\tau}\sum_{w\in\mathcal{V}} P_{vw}\|\overline{x}_w - x_v\|_2^2 \Bigg] \tag{9}$$

for some $q \in (0,1)$. By minimizing $\mathcal{F}_p$, each node seeks to estimate the sparse vector $\widetilde{x}$ and to enforce agreement with the estimates calculated by other nodes in the network. It should also be noted that $\mathcal{F}_p(x,\dots,x) = q\mathcal{J}_p(x)$.

The following theorem ensures that, under certain conditions on the parameter τ, the problem of minimizing the functions in (9) is well-posed. We derive necessary optimality conditions of (9) and discuss the relationships with the original problem (5).

Let $\Gamma_p : \mathbb{R}^{n\times|\mathcal{V}|} \longmapsto \mathbb{R}^{n\times|\mathcal{V}|}$ be the operator defined as

$$(\Gamma_p X)_v = \eta_{p,\alpha}\left[(1-q)(X(P^{\mathsf{T}})^2)_v + q(x_v + \tau A_v^{\mathsf{T}}(y_v - A_v x_v))\right] \tag{10}$$

where $v \in \mathcal{V}$, $\alpha = q\lambda/|\mathcal{V}|$ and $\eta_{p,\alpha}$ is defined in (6) and (7).

Theorem 1 (Characterization of minima)**.** *If $\tau < \|A_v\|_2^{-2}$ for all $v \in \mathcal{V}$, the sets of minimizers of the functions $\mathcal{F}_p$, defined in* (9)*, are not empty and coincide with the sets*

$$\mathrm{Fix}(\Gamma_p) := \{Z \in \mathbb{R}^{n\times\mathcal{V}} : \Gamma_p Z = Z\}$$

where $p \in \{0,1\}$.

Theorem 1 is proved in Section III-A through variational techniques. The more technical points are postponed to Appendix A.

Theorem 2. *Let us denote as $\{\widehat{x}_v^q\}_{v\in\mathcal{V}}$ the minimizing value of $\mathcal{F}_p(x_1,\dots,x_{|\mathcal{V}|})$ in* (9)*. If $\mathcal{G}$ is connected, then $\lim_{q\to 0}\widehat{x}_v^q = \widehat{x}$, $\forall v \in \mathcal{V}$, where $\widehat{x}$ is the minimizing value of $\mathcal{J}_p(x)$ in* (5)*.*

Theorem 2 states that q can be interpreted as a temperature; as q decreases, estimates x_v's associated with adjacent nodes become increasingly correlated. This fact suggests that if q is sufficiently small, then each vector $\widehat{x}_v^q$ can be used as an estimate of the model parameters $\widetilde{x}$. The proof of Theorem 2 is reported in Section III-B.

A. Proof of Theorem 1

We now prove rigorously Theorem 1 through intermediate steps. Since for the most part of the proof computations are the same for $p = 0$ and $p = 1$, we will distinguish the two cases only when necessary, and just use p otherwise.

Instead of optimizing $\mathcal{F}_p$, let us introduce a surrogate objective function [41]:

$$\mathcal{F}_p^{\mathcal{S}}(X,C,B) := \sum_{v\in\mathcal{V}} \Bigg(q\|A_v x_v - y_v\|_2^2 + \frac{2\alpha}{\tau}\|x_v\|_p + \frac{1-q}{d\tau}\sum_{w\in\mathcal{N}_v}\|x_v - c_w\|_2^2 + \frac{q}{\tau}\|x_v - b_v\|_2^2 - q\|A_v(x_v - b_v)\|_2^2 \Bigg) \tag{11}$$

where $C = (c_1,\dots,c_{|\mathcal{V}|}) \in \mathbb{R}^{n\times|\mathcal{V}|}$, $B = (b_1,\dots,b_{|\mathcal{V}|}) \in \mathbb{R}^{n\times|\mathcal{V}|}$. It should be noted that, defining $\overline{X} = XP^{\mathsf{T}}$,

$$\mathcal{F}_p^{\mathcal{S}}(X,\overline{X},X) = \mathcal{F}_p(X)$$

and that if $\tau < \|A_v\|_2^{-2}$ for all $v \in \mathcal{V}$ then this surrogate objective function is a majorization of $\mathcal{F}_p$ [42]. The optimization of (11) can be computed by minimizing with respect to each x_v separately.

Proposition 2. *The following fact holds*

$$\operatorname*{argmin}_{x_v\in\mathbb{R}^n} \mathcal{F}_p^{\mathcal{S}}(X,C,B) = \eta_{p,\alpha}\left[(1-q)\overline{c}_v + qb_v + q\tau A_v^{\mathsf{T}}(y_v - A_v b_v)\right]$$

where $\overline{c}_v = \frac{1}{d}\sum_{w\in\mathcal{N}_v} c_w$.

Proof. We can write

$$\mathcal{F}_p^{\mathcal{S}}(X,C,B) = \frac{1}{\tau}\left\|x_v - \left[(1-q)\overline{c}_v + q(b_v + \tau A_v^{\mathsf{T}}(y_v - A_v b_v))\right]\right\|_2^2 + \frac{2\alpha}{\tau}\|x_v\|_p + K \tag{12}$$

where $K \in \mathbb{R}$ is a term independent of x_v. The statement follows from the fact that $\eta_{p,\alpha}[z] = \operatorname*{argmin}_x \|x - z\|_2^2 + 2\alpha\|x\|_p$. □

Proposition 3. *If $\tau < \|A_v\|^{-2}$ for all $v \in \mathcal{V}$ the following facts hold:*

$$\operatorname*{argmin}_{c_v\in\mathbb{R}^n} \mathcal{F}_p^{\mathcal{S}}(X,C,B) = \frac{1}{d}\sum_{w\in\mathcal{N}_v} x_w, \tag{13}$$

$$\operatorname*{argmin}_{b_v\in\mathbb{R}^n} \mathcal{F}_p^{\mathcal{S}}(X,C,B) = x_v. \tag{14}$$

Proof. Since the graph is regular, the first statement follows from the fact that

$$\mathcal{F}_p^{\mathcal{S}}(X,C,B) = \sum_{v\in\mathcal{V}}\frac{1-q}{d\tau}\sum_{w\in\mathcal{N}_v}\|x_v - c_w\|_2^2 + K = \sum_{w\in\mathcal{V}}\frac{1-q}{d\tau}\sum_{v\in\mathcal{N}_w}\|x_v - c_w\|_2^2 + K$$

where K is independent of c_v. The proof of the second statement is immediate, assuming that $\tau < \|A_v\|^{-2}$. □

Proposition 4. *There exists $\widehat{\alpha} \in \mathbb{R}$ such that if the parameter $\alpha < \widehat{\alpha}$, then the set of fixed points $\mathrm{Fix}(\Gamma_0)$ is not empty.*

Proof. If $\alpha < \widehat{\alpha} := \min_{i\in\mathrm{supp}(\widetilde{x})}|\widetilde{x}_i|$ then $\eta_{0,\alpha}(\widetilde{x}) = \widetilde{x}$. We observe that $X_0 = \widetilde{x}\mathbb{1}^{\mathsf{T}} \in \mathrm{Fix}(\Gamma_0)$. In fact

$$\Gamma_0 X_0 = (1-q)X_0 P^{\mathsf{T}} + qX_0 = (1-q)\widetilde{x}\mathbb{1}^{\mathsf{T}}P^{\mathsf{T}} + q\widetilde{x}\mathbb{1}^{\mathsf{T}} = X_0$$

where the last inequality follows from the fact that P is row-stochastic. □

Proposition 5. *The set of minimizers of $\mathcal{F}_1(X)$ is not empty.*

Proof. In order to prove that $\mathcal{F}_1(X)$ admits minimizers, it is sufficient to notice that, for any $K > 0$, the set $\{X \in$

$\mathbb{R}^{n\times|\mathcal{V}|} : \mathcal{F}_1(X) \leq K\}$ is closed set (as it is counterimage of a closed through a continuous function) and bounded (as $\sum_{v\in\mathcal{V}} \frac{2\alpha}{\tau}\|x_v\|_1 \leq \mathcal{F}_1(X) \leq K$). The compactness guarantees that $\mathcal{F}_1(X)$ has at least one minimizer. □

Lemma 1. *The following facts hold*

1) If $U^\star \in \mathrm{Fix}(\Gamma_p)$, there exists $\epsilon > 0$ such that, for $H = (h_1,\dots,h_{|\mathcal{V}|}) \in \mathbb{R}^{n\times|\mathcal{V}|}$, $|h_{jv}| < \epsilon$, $v \in \mathcal{V}$, $j = 1,\dots,n$, then

$$\mathcal{F}_p^{\mathcal{S}}(U^\star + H, \overline{U^\star}, U^\star) \geq \mathcal{F}_p(U^\star) + \frac{1}{\tau}\|H\|_F^2.$$

2) If $U^\star \in \mathrm{Fix}(\Gamma_p)$, there exists $\epsilon > 0$ such that, for $H = (h_1,\dots,h_{|\mathcal{V}|}) \in \mathbb{R}^{n\times|\mathcal{V}|}$, $|h_{jv}| < \epsilon$, $v \in \mathcal{V}$, $j = 1,\dots,n$, then

$$\mathcal{F}_p^{\mathcal{S}}(U^\star + H, \overline{U^\star}, U^\star) \leq \mathcal{F}_p(U^\star + H) + \frac{1}{\tau}\|H\|_F^2.$$

The proof is rather technical and for this reason is postponed to Appendix A.

We conclude the proof of Theorem 1 by showing the minimizers of $\mathcal{F}_p(X)$ coincide with the fixed points of Γ_p. Merging assertions 1) and 2) in Lemma 1, we obtain that if $U^\star \in \mathrm{Fix}(\Gamma_p)$ for a sufficiently small increment H holds

$$\mathcal{F}_p(U^\star + H) \geq \mathcal{F}_p(U^\star)$$

which means that $U^\star$ is a minimum for $\mathcal{F}_p(\cdot)$. If $U^\star$ is a minimizer of $\mathcal{F}_p(\cdot)$, then

$$\begin{aligned}\mathcal{F}_p(U^\star) = \mathcal{F}_p^{\mathcal{S}}(U^\star, \overline{U^\star}, U^\star) &\leq \mathcal{F}_p^{\mathcal{S}}(U^\star + H, \overline{U^\star} + \overline{H}, U^\star + H)\\ &\leq \mathcal{F}_p^{\mathcal{S}}(U^\star + H, \overline{U^\star}, U^\star + H) \leq \mathcal{F}_p^{\mathcal{S}}(U^\star + H, \overline{U^\star}, U^\star)\end{aligned}$$

Therefore, $U^\star$ is a minimizer for $\mathcal{F}_p^{\mathcal{S}}(\cdot, \overline{U^\star}, U^\star)$. This implies that

$$u_v^\star = \eta_{p,\alpha}\left((1-q)\overline{\overline{u_v^\star}} + q(u_v^\star + \tau A_v^{\mathsf{T}}(y_v - A_v u_v^\star)\right)$$

where $\overline{\overline{u_v^\star}} = \frac{1}{d}\sum_{w\in\mathcal{N}_v}\overline{u_w^\star}$ which means that $U^\star \in \mathrm{Fix}(\Gamma_p)$. The thesis is then obtained using Proposition 4 and Proposition 5. □

B. Proof of Theorem 2

Let

$$\widehat{X}^q = [\widehat{x}_1^q,\dots,\widehat{x}_{|\mathcal{V}|}^q] = \operatorname*{argmin}_{x_1,\dots,x_{|\mathcal{V}|}} \frac{1}{q}\mathcal{F}_1(x_1,\dots,x_{|\mathcal{V}|}).$$

We prove the assertion by showing the following facts:

i. first, the convergence to a consensus, *i.e.*

$$\lim_{q\to 0}\|\widehat{x}_v^q - \widehat{x}_w^q\| = 0 \quad \forall v,w \in \mathcal{V};$$

ii. second, the convergence to a common value

$$\forall v \in \mathcal{V} \quad \lim_{q\to 0}\widehat{x}_v^q = \widehat{x},$$

which is the minimum of function $\mathcal{J}_p(x)$, *i.e.* $\mathcal{J}_p(\widehat{x}) \leq \mathcal{J}_p(x), \forall x \in \mathbb{R}^n$.

We start with point i.: suppose ad absurdum that there exist $(v,w) \in \mathcal{E}$, a sequence $\{q_\ell\}_{\ell\in\mathbb{N}}$ converging to zero and $\epsilon > 0$ such that there exist infinitely many $\ell \in \mathbb{N} : \|x_v^{q_\ell} - \overline{x}_w^{q_\ell}\| > \epsilon$ and, consequently,

$$\frac{1}{q_\ell}\mathcal{F}_p(\widehat{X}^{q_\ell}) > \frac{(1-q_\ell)P_{vw}\epsilon}{\tau q_\ell}.$$

Since $P_{vw} > 0$, then $\lim_{q\to 0}\frac{(1-q)P_{vw}\epsilon}{q\tau} = +\infty$ or, equivalently by definition, for any $\chi > 0$ there exists $\ell_0 > 0$ such that if $\ell > \ell_0$ then

$$\frac{(1-q_\ell)P_{vw}\epsilon}{q_\ell\tau} > \chi.$$

Let us fix the constant $\chi = \sum_{v\in\mathcal{V}}\|y_v\|_2^2$, then

$$\frac{1}{q_\ell}\mathcal{F}_p(\widehat{X}^{q_\ell}) > \frac{(1-q_\ell)P_{vw}\epsilon}{q_\ell\tau} > \chi = \frac{1}{q_\ell}\mathcal{F}_1(0)$$

and we obtain the contradiction that $\widehat{X}^{q_\ell}$ is not the minimizing value of (9). Since the graph is connected and applying similar arguments used for Proposition 1, we deduce that

$$\lim_{q\to 0}\|\widehat{x}_v^q - \widehat{x}_w^q\| = 0 \quad \forall w \in \mathcal{N}_v.$$

We now prove point ii.. Suppose ad absurdum that there exists $v \in \mathcal{V}$ and two sequences $\{q_\ell\}_{\ell\in\mathbb{N}}$ and $\{m_\ell\}_{\ell\in\mathbb{N}}$ converging to zero such that

$$\lim_{\ell\to\infty}\widehat{x}_v^{q_\ell} = \xi_1 \quad \lim_{\ell\to\infty}\widehat{x}_v^{m_\ell} = \xi_2.$$

From point i. we deduce that

$$\lim_{\ell\to\infty}\widehat{x}_w^{q_\ell} = \xi_1 \quad \lim_{\ell\to\infty}\widehat{x}_w^{m_\ell} = \xi_2.$$

for all $w \in \mathcal{V}$. Then,

$$\frac{\mathcal{F}_1(\widehat{X}^q)}{q} \geq \sum_{v\in\mathcal{V}}\left[\|y_v - A_v\widehat{x}_v^q\|_2^2 + \frac{2\lambda}{\tau|\mathcal{V}|}\|\widehat{x}_v^q\|_1\right].$$

By definition of $\widehat{X}^q$ we also have for all $X \in \mathbb{R}^{n\times|\mathcal{V}|}$

$$\begin{aligned}\frac{\mathcal{F}_1(\widehat{X}^q)}{q} \leq \sum_{v\in\mathcal{V}}\Bigg[&\|y_v - A_v x_v\|_2^2 + \frac{2\lambda}{\tau|\mathcal{V}|}\|x_v\|_1\\ &+\frac{1-q}{q\tau}\sum_{w\in\mathcal{V}}P_{vw}\|\overline{x}_w - x_v\|_2^2\Bigg]\end{aligned}$$

and, in particular,

$$\frac{\mathcal{F}_1(\widehat{X}^q)}{q} \leq \sum_{v\in\mathcal{V}}\left[\|y_v - A_v x\|_2^2 + \frac{2\lambda}{\tau|\mathcal{V}|}\|x\|_1\right], \quad \forall x \in \mathbb{R}^n$$

We now distinguish the cases $p = 1$ and $p = 0$.

1) If $p = 1$, by letting $\ell \to \infty$ and considering that $\lim_{\ell\to\infty}\widehat{X}^{q_\ell} = \xi_1\mathbb{1}^\top$ and $\mathcal{F}_1$ is a continuous function, then

$$\begin{aligned}\mathcal{J}_1(\xi_1) &= \lim_{\ell\to\infty}\sum_{v\in\mathcal{V}}\left[\|y_v - A_v\widehat{x}_v^{q_\ell}\|_2^2 + \frac{2\lambda}{\tau|\mathcal{V}|}\|\widehat{x}_v^{q_\ell}\|_1\right]\\ &\leq \lim_{\ell\to\infty}\frac{\mathcal{F}_1(\widehat{X}^{q_\ell})}{q_\ell} \leq \sum_{v\in\mathcal{V}}\left[\|y_v - A_v x\|_2^2 + \frac{2\lambda}{\tau|\mathcal{V}|}\|x\|_1\right]\\ &= \mathcal{J}_1(x), \quad \forall x \in \mathbb{R}^n\end{aligned}$$

and, analogously,

$$\mathcal{J}_1(\xi_2) \leq \mathcal{J}_1(x), \quad \forall x \in \mathbb{R}^n$$

From Assumption 1 we conclude that $\widehat{x} = \xi_1 = \xi_2$.

2) If $p = 0$, from theorem of sign permanence we have that if $\xi_{1,i} \neq 0$ then there exists $\ell_0 \in \mathbb{N}$ such that also $\widehat{x}_{i,v}^{q_\ell} \neq 0$, $\forall \ell > \ell_0$. This implies that $\|\xi_1\|_0 \leq \|x_v^{q_\ell}\|_0$ $\forall \ell > \ell_0$ and

$$\begin{aligned}\mathcal{J}_0(\xi_1) &\leq \lim_{\ell\to\infty} \sum_{v\in\mathcal{V}} \left[\|y_v - A_v \widehat{x}_v^{q_\ell}\|_2^2 + \frac{2\lambda}{\tau|\mathcal{V}|}\|\widehat{x}_v^{q_\ell}\|_0\right] \\ &\leq \lim_{\ell\to\infty} \frac{\mathcal{F}_0(\widehat{X}^{q_\ell})}{q_\ell} \leq \sum_{v\in\mathcal{V}} \left[\|y_v - A_v x\|_2^2 + \frac{2\lambda}{\tau|\mathcal{V}|}\|x\|_0\right] \\ &= \mathcal{J}_0(x), \quad \forall x \in \mathbb{R}^n\end{aligned}$$

From Theorem 1 it can be deduced that the first inequality is actually an equality $\forall \ell > \ell_0$. Using similar arguments for the sequence $\widehat{x}_v^{m_\ell}$ and Assumption 1 we conclude that $\xi_1 = \xi_2$.

□

IV. Proposed distributed estimation algorithms

In this section we introduce two distributed iterative algorithms to solve (5), in which the nodes only exchange information with their nearest neighbors at each iteration, without any central coordination. In particular, we describe a family of low-complexity subgradient thresholding methods.

A. Algorithms description

Distributed iterative thresholding algorithms seek to minimize (9) in an iterative, distributed way. The key idea is as follows.

The algorithm is parametrized by a stochastic transition matrix P which is adapted to the graph. All nodes v store two messages at each time $t \in \mathbb{N}$, $x_v(t)$ and $\overline{x}_v(t)$. Starting from $x_v(0) = 0$ for all $v \in \mathcal{V}$, the update is performed in an alternating fashion. More specifically, the update consists of two stages; for convenience, the first stage is identified with even times $t \in 2\mathbb{N}$, and the second stage with odd times $t \in 2\mathbb{N}+1$, so that one complete iteration spans two time units. At even time $t \in 2\mathbb{N}$, each node $v \in \mathcal{V}$ receives the estimates $x_w(t)$ for each $w \in \mathcal{N}_v$, which is communicating with v, and $\overline{x}_v(t+1)$ is obtained by a convex combination of these estimates. At odd time $t \in 2\mathbb{N}+1$, each node receives the vectors $\overline{x}_w(t)$ from their neighbors, the estimate $x_v(t+1)$ is then obtained applying the thresholding operator to a convex combination of the received messages and of the subgradient step.

The coefficients of the convex combination are obtained through the matrix P and the temperature parameter $q \in (0,1)$. In a simple case the nodes compute simply the average of the received messages, giving equal weight to each of them and setting $q = 1/2$. Other solutions are also possible, in which the weights are computed in an optimal manner according to some cost function. *E.g.*, in sensor networks, nodes may be affected by different noise, or some sensors may be partially damaged. The coefficients in P could take into account the noise by adjusting the weights. Noise may also be noise on the wireless channel communication, in which case its variance is proportional to the distance. The design parameter $q \in (0,1)$ balances, instead, cooperation and the gradient descent, and can be suitably tuned to optimize different performance parameters (mean-squared error, detection error on support and so on).

The thresholding operation can be hard or soft as described in (6) and (7). We refer to DIHTA in the case of (6) and DISTA if the thresholding operator is (7).

More precisely, the patterns are described in Algorithm 1 and Algorithm 2.

Algorithm 1 DIHTA

Given a row-stochastic matrix P adapted to the graph, $\alpha = q\lambda/|\mathcal{V}| > 0$, $\tau > 0$, $x_v(0) = 0$, $y_v = A_v\widetilde{x}$ for any $v \in \mathcal{V}$, iterate

- $t \in 2\mathbb{N}$, $v \in \mathcal{V}$,

$$\begin{aligned}\overline{x}_v(t+1) &= \sum_{w\in\mathcal{V}} P_{v,w} x_w(t) \\ x_v(t+1) &= x_v(t)\end{aligned}$$

- $t \in 2\mathbb{N}+1$, $v \in \mathcal{V}$,

$$\begin{aligned}\overline{x}_v(t+1) &= \overline{x}_v(t) \\ x_v(t+1) &= \eta_{0,\alpha}\Bigg[(1-q)\sum_{w\in\mathcal{V}} P_{v,w}\overline{x}_w(t) \\ &\qquad + q\Big(x_v(t) + \tau A_v^{\mathsf{T}}(y_v - A_v x_v(t))\Big)\Bigg].\end{aligned}$$

Algorithm 2 DISTA

Given a row-stochastic matrix P adapted to the graph, $\alpha = q\lambda/|\mathcal{V}| > 0$, $\tau > 0$, $x_v(0) = 0$, $y_v = A_v\widetilde{x}$ for any $v \in \mathcal{V}$, iterate

- $t \in 2\mathbb{N}$, $v \in \mathcal{V}$,

$$\begin{aligned}\overline{x}_v(t+1) &= \sum_{w\in\mathcal{V}} P_{v,w} x_w(t) \\ x_v(t+1) &= x_v(t)\end{aligned}$$

- $t \in 2\mathbb{N}+1$, $v \in \mathcal{V}$,

$$\begin{aligned}\overline{x}_v(t+1) &= \overline{x}_v(t) \\ x_v(t+1) &= \eta_{1,\alpha}\Bigg[(1-q)\sum_{w\in\mathcal{V}} P_{v,w}\overline{x}_w(t) \\ &\qquad + q\Big(x_v(t) + \tau A_v^{\mathsf{T}}(y_v - A_v x_v(t))\Big)\Bigg].\end{aligned}$$

The proposed methods define a distributed protocol: each node only needs to be aware of its neighbors and no further information about the network topology is required. It should be noted that if $|\mathcal{V}| = 1$, DIHTA and DISTA coincide with IHTA and ISTA, respectively.

B. Discussion on related literature

Algorithms for distributed sparse recovery (with no central processing unit) in sensor networks have been proposed in the literature in the last few years. We distinguish two classes:

1) algorithms based on the decentralization of subgradient methods for convex optimization (DSM, [29], [33], [34]);
2) consensus and distributed ADMM ([30], [43], [44]);

1) DSM: As has been said, our proposed approach leverages distributed algorithms for multi-agent optimization that have been proposed in the literature in the last few years [29], [33]. The main goal of these algorithms is to minimize over a convex set the sum of cost functions that are convex and differentiable almost everywhere. Formally, the following problem is considered:

$$\min_{x\in\mathbb{R}^n} \sum_{v\in\mathcal{V}} f_v(x) \tag{15}$$

where $f_v : \mathbb{R}^n \to \mathbb{R}$ are convex functions. The main idea behind distributed algorithms is to use consensus as a mechanism for spreading information through the network. In particular, each agent, starting from an initial estimate, updates it by first combining the estimates received from its neighbors, then taking a subgradient step, in order to minimize its objective function. More formally,

$$\begin{aligned} \overline{x}_v(t+1) &= \sum_{w\in\mathcal{V}} P_{v,w} x_w \\ x_v(t+1) &= \overline{x}_v(t+1) - r\partial f_v(x_v(t)) \end{aligned} \tag{16}$$

where P is a stochastic matrix, $r \in (0,1)$, and $\partial f_v(x_v(t))$ is a subgradient of the function f_v, evaluated at $x_v(t)$.

It should be noted that these subgradient methods can be applied to the Lasso functional in (3) but not to solve the ℓ_0-regularized LS problem in (2), which is not a convex function.

As has been said, DIHTA and DISTA leverage the projected subgradient methods for multi-agent optimization. The memory storage requirements and the computational complexity are similar, as it does not require to solve linear systems, to invert matrices, or to operate on the matrices A_v. However, we emphasize some substantial differences.

The protocol in (16) is not guaranteed to converge while both DIHTA and DISTA will be proved to converge to a minimum of (9). The convergence of (16) can be achieved by considering the related "stopped" model (see page 56 in [45]), whereby the nodes stop computing the subgradient at some time, but they keep exchanging their information and averaging their estimates only with neighboring messages for subsequent time. However, the tricky point of such techniques is the optimal choice of the number of iterations to stop the computation of the subgradient. Moreover, the limit point can not be variationally characterized and depends on the time we stop the model.

In the analysis of the algorithm in [29], the functions f_v are assumed polyhedral and the subgradient sets of each f_v bounded over the set Ω. This is not true for Lasso cost in general linear inverse problems [6].

In [34] a version of DSM is proposed for optimization of (15) over a convex set, which in general is not the case of LS problems under sparsity constraints, whose feasible sets are not convex. Moreover, the thresholding operators defined in (2) are not projections on convex sets, *i.e.* $\eta_{0,\alpha}$ is not nonexpansive and $\eta_{1,\alpha}(x) \neq x$ for all $x \neq 0$.

In [46], the stepsize r in (16) decreases to zero along the iterations. This choice, however, requires to fix an initial time and is not be feasible in case of time-variant input: introducing a new input would require some resynchronization. For this reason the parameters q, τ in distributed iterative thresholding algorithms are kept fixed and will be compared with (16) with r fixed.

2) Consensus ADMM: The consensus ADMM [18], [30] is a method for solving problems in which the objective and the constraints are distributed across multiple processors. The problem in (5) is tackled by introducing dual variables ω_v and minimizing the augmented Lagrangian in a iterative way with respect to the primal and dual variables. The algorithm entails the following steps for each $t \in \mathbb{N}$: node v receives the local estimates from its neighbors, uses them to evaluate the dual price vector and calculate the new estimate via coordinate descent and thresholding operations. Formally, choosing some $\rho > 1$, the update equations would be typically be performed according to the following rules:

$$\begin{aligned} x_v(t+1) &= (A_v^{\mathsf{T}} A_v + \rho I)^{-1}(A_v^{\mathsf{T}} y + \rho z(t) - \omega_v(t)) \\ z(t+1) &= \eta_{\lambda/\rho}\left(\frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{N}} x_v(t+1) + \omega_v(t)/\rho\right) \\ \omega_v(t+1) &= \omega_v(t) + \rho(x_v(t+1) - z(t+1)). \end{aligned} \tag{17}$$

More recently, distributed versions of ADMM that just require local communications have been proposed in the literature [43], [44]. These methods however address problems slightly different from ours, and in particular they require stronger convexity assumptions on the cost functional to guarantee the convergence.

C. Complexity and memory analysis

The bottleneck of the consensus ADMM is the inversion of the $n \times n$ matrices $(A_v^{\mathsf{T}} A_v + \rho I)$. Even if assuming that such inversion can performed off-line and does not affect the procedure, the storage of the inverse matrix may be prohibitive as well for a node with a small amount of available memory. Specifically, for the consensus ADMM each node has to store $O(n^2)$ real values. DIHTA and DISTA, instead, require only $O(n)$ real values. Just to do a practical example, nodes with 16kB of RAM are widely used for wireless sensor networks, *e.g.* as STM32 F3-series microcontrollers with Contiki operating system. As the static memory occupied by ADMM and DISTA is almost the same, we neglect it along with the memory used by the operating system (the total is of the order of hundreds of byte). Using a single-precision floating-point format, 2^{12} real values can be stored in 16 kB. Therefore, even assuming just one measurement per node ($m = 1$), ADMM can handle signals with length of some tens, while DISTA up to thousands. This illustrates the greater efficiency of DISTA in low memory devices.

V. Convergence analysis

In Section IV we have derived DIHTA and DISTA to minimize the cost function $\mathcal{F}_p(x_1,\dots,x_{|\mathcal{V}|})$ defined in (9) with $p=0$ and $p=1$, respectively. Theorem 1 guarantees that, under suitable conditions, the minima of the cost functions $\mathcal{F}_0$ and $\mathcal{F}_1$ defined in (9) coincide with the fixed points of the map that rule the dynamics of DIHTA and DISTA, respectively. In this section, we present our theoretical results regarding the convergence of the proposed algorithms, conveniently organized into two theorems: Theorem 3 and Theorem 4 define sufficient conditions to guarantee the convergence of DIHTA and DISTA to a fixed point. It follows that DIHTA and DISTA converge to a minimum of $\mathcal{F}_0$ and $\mathcal{F}_1$, respectively. It is worth mentioning that $\mathcal{F}_0$ is not convex, then DIHTA is actually proved to converge to a local minimum.

In order to state our results in formal way, it is convenient to rewrite the dynamics of the proposed algorithms. In particular, we express the iterations as follows. Let

$$X(t) = (x_1(t),\dots,x_{|\mathcal{V}|}(t))$$
$$\overline{X}(t) = X(t)P^{\mathsf{T}}, \qquad t\in\mathbb{N}.$$

The updates of DIHTA and DISTA can thus be rewritten as

$$X(t+1) = \Gamma_p X(t) \tag{18}$$

with $p=0$ and $p=1$, respectively. Notice that this recursive formula joins in one step the operations that in the algorithms 1 and 2 are splitted into two steps, but the dynamics is actually the same. $X(0)$ can be arbitrarily initialized.

In our analysis, in addition to Assumption 2, we adopt the following.

Assumption 3. *The nodes of the graph $\mathcal{G}=(\mathcal{V},\mathcal{E})$ use uniform weights,* i.e.*, $P_{u,v}=1/d$ if $(u,v)\in\mathcal{E}$ and zero otherwise.*

Theorem 1 ensures that, under certain conditions on the stepsize τ, the problem of minimizing the functions in (9) is well-posed. Moreover, it states the equivalence between the minima and the fixed points of Γ_p.

Moreover, the sequences $\{X(t)\}_{t\in\mathbb{N}}$ given by (18) for both $p=0$ and $p=1$ are guaranteed to be bounded and to converge to a finite limit point.

Theorem 3 (DIHTA convergence)**.** *There exists $\widehat{\alpha}\in\mathbb{R}$ such that if $\alpha<\widehat{\alpha}$ and $\tau<\|A_v\|_2^{-2}$ for all $v\in\mathcal{V}$, DIHTA produces a sequence $\{X(t)\}_{t\in\mathbb{N}}$ (defined by* (18) *with $p=0$) such that*

$$\lim_{t\to\infty}\inf_{Z\in\mathrm{Fix}(\Gamma_0)}\|X(t)-Z\|_F=0.$$

Moreover, defined $A=[A_1^{\mathsf{T}},A_2^{\mathsf{T}},\dots,A_{|\mathcal{V}|}^{\mathsf{T}}]^{\mathsf{T}}$, if $A^{\mathsf{T}}A$ is positive definite then

$$\lim_{t\to\infty}\|X(t)-X^\star\|_F=0$$

where $X^\star\in\mathrm{Fix}(\Gamma_0)$.

Theorem 4 (DISTA convergence)**.** *If $\tau<\|A_v\|_2^{-2}$ for all $v\in\mathcal{V}$, DISTA produces a sequence $\{X(t)\}_{t\in\mathbb{N}}$ (defined by* (18) *with $p=1$) such that*

$$\lim_{t\to\infty}\|X(t)-X^\star\|_F=0$$

where $X^\star\in\mathrm{Fix}(\Gamma_1)$.

These theorems guarantee that both DIHTA and DISTA produce sequences of estimates converging to minima of $\mathcal{F}_0$ and $\mathcal{F}_1$, respectively.

It is worth mentioning that Theorem 3 and 4 do not imply necessarily the convergence of the algorithm to a consensus. If $p=0$ the nodes in the graph achieve a consensus in the estimate of the original model parameters; this is not theoretically proved and is left for future research. On the other hand, if $p=1$ local estimates do not necessary coincide at convergence. However, the disagreement among the nodes is controlled by temperature parameter q. The consensus can be achieved by letting q go to zero as suggested by Theorem 2 or considering the related "stopped" model [29], whereby the nodes stop computing the subgradient at some time, but they keep exchanging their information and averaging their estimates only with neighboring messages for subsequent time. It should be noted that in the literature several consensus-based estimation algorithms have been proposed, which do not reach consensus but where the agreement can be induced by using a temperature parameter, *e.g.*, [34], [47].

Theorem 3 and 4 instead are proved through the intermediate steps:

- first, we show that DIHTA and DISTA produce sequences that do not increase $\mathcal{F}_p$, respectively for $p=0$ and $p=1$, which leads to numerical convergence (see Section V-A);
- once numerical convergence is achieved, convergence is shown in two different ways for DIHTA and DISTA, respectively using the LaSalle invariance principle and the Opial's Theorem (Section V-B and Section V-C, respectively). The more technical lemmas are postponed to Appendix B.

A. Numerical convergence

This section gives strong hints that both DIHTA and DISTA converge to a limit point. In particular, we now prove that two successive iterations of both algorithms become closer and closer, which implies the numerical convergence when the number of iterations goes to infinity. More technical work is further needed in order to prove that the full sequence converges to a minimizer of the cost function $\mathcal{F}_p$ (see Sections V-B and V-C).

Our starting point is to show that Γ_p is asymptotically regular, *i.e.* $X(t+1)-X(t)\to 0$ for $t\to\infty$. In particular, this property guarantees the numerical convergence of the algorithm.

Lemma 2. *If $\tau<\|A_v\|_2^{-2}$ for all $v\in\mathcal{V}$, then the sequence $\{\mathcal{F}_p(X(t))\}_{t\in\mathbb{N}}$ is non increasing and admits the limit.*

Proof. The function is lower bounded ($\mathcal{F}(X)\geq 0$) and the sequence $\{\mathcal{F}_p(X(t)\}_{t\in\mathbb{N}}$ is decreasing and therefore admits the limit.

From Proposition 2 and Proposition 3 we obtain the follow-

ing inequalities:

$$\begin{aligned}\mathcal{F}_p(X(t+1)) &\leq \mathcal{F}_p^{\mathcal{S}}(X(t+1), \overline{X}(t+1), X(t)) \\ &\leq \mathcal{F}_p^{\mathcal{S}}(X(t+1), \overline{X}(t), X(t)) \\ &\leq \mathcal{F}_p^{\mathcal{S}}(X(t), \overline{X}(t), X(t)) = \mathcal{F}_p(X(t)).\end{aligned}$$

□

At this point, we can conclude the asymptotic regularity of Γ_p:

Proposition 6. *For any $\tau \leq \min_{v\in\mathcal{V}} \|A_v\|_2^{-2}$ the sequence $\{X(t)\}_{t\in\mathbb{N}}$ is bounded and*

$$\lim_{t\to+\infty} \|X(t+1) - X(t)\|_F^2 = 0.$$

Proof. By Proposition 2 we obtain

$$\begin{aligned}&\mathcal{F}(X(t)) - \mathcal{F}(X(t+1)) \\ &= \mathcal{F}_p^{\mathcal{S}}(X(t), \overline{X}(t), X(t)) \\ &- \mathcal{F}_p^{\mathcal{S}}(X(t+1), \overline{X}(t+1), X(t+1)) \\ &\geq \mathcal{F}_p^{\mathcal{S}}(X(t+1), \overline{X}(t), X(t)) \\ &- \mathcal{F}_p^{\mathcal{S}}(X(t+1), \overline{X}(t+1), X(t+1)) \\ &\geq \mathcal{F}_p^{\mathcal{S}}(X(t+1), \overline{X}(t+1), X(t)) \\ &- \mathcal{F}_p^{\mathcal{S}}(X(t+1), \overline{X}(t+1), X(t+1)) \\ &\geq \frac{q}{\tau} \sum_{v\in\mathcal{V}} (x_v(t+1) - x_v(t))^{\mathsf{T}} M_v (x_v(t+1) - x_v(t)) \geq 0.\end{aligned}$$

Notice that the last expression is nonnegative as $M_v = I - \tau A_v^{\mathsf{T}} A_v$ are positive definite for all $v \in \mathcal{V}$.

As $\mathcal{F}_p^{\mathcal{S}}(X(t)) - \mathcal{F}_p^{\mathcal{S}}(X(t+1) \to 0$ we thus conclude that $\|x_v(t+1) - x_v(t)\|_2^2 \to 0$ for any $v \in \mathcal{V}$ and

$$\lim_{t\to+\infty} \|X(t+1) - X(t)\|_F^2 = 0.$$

□

B. Convergence of DIHTA (Proof of Theorem 3)

In the previous section we have proved that two successive iterations of DIHTA become closer and closer. We now prove that the full sequence converges to a minimizer of the cost function $\mathcal{F}_0$. This part concludes the proof of Theorem 3.

More precisely, the proof of the convergence is obtained through intermediate steps:

- the support of each estimate $x_v(t)$ stabilizes at a finite time (Corollary 1);
- the sequence $\{x_v(t)\}_{t\in\mathbb{N}}$ is bounded (Lemma 3);
- the conclusion is obtained applying the LaSalle invariance principle (Theorem 5) and the fact that the minima of cost function $\mathcal{F}$ are isolated (Lemma 4).

We start proving the stabilization of the support for each estimate $x_v(t)$.

Corollary 1. *There exists $t_0 \in \mathbb{N}$ such that $\operatorname{supp}(x_v(t)) = x_v(t_0)$, for all $t \geq t_0$.*

Proof. This is a direct consequence of the asymptotic regularity: for all $v \in \mathcal{V}$,

$$\lim_{t\to+\infty} \|x_v(t) - x_v(t+1)\|_2 = 0.$$

Let us fix $\epsilon \in (0, \alpha)$, then there exists $t_0 \in \mathbb{N}$ such that for all $t > t_0$ holds

$$\|x_v(t) - x_v(t+1)\|_2 < \epsilon.$$

This implies that $\forall v \in \mathcal{V}$ and $\forall j \in \{1, \dots, n\}$

$$|x_{jv}(t) - x_{jv}(t+1)| \leq \|x_v(t) - x_v(t+1)\|_2 < \alpha.$$

It should be noted that if $x_{jv}(t_0) = 0$ at the time $t_0 \in \mathbb{N}$, then $|x_{jv}(t)| < \alpha$ for any $t > t_0$ and because of the effect of thresholding then $x_{jv}(t_0 + 1) = 0$ and the same is repeated at any following step. Vice versa, if $|x_{jv}(t_0 + 1)| < \alpha$, then $|x_{jv}(t_0)| < \alpha$; this means that if $|x_{jv}(t_0)| > \alpha$, then $|x_{jv}(t_0 + 1)| > \alpha$ and the same is repeated at any following step. □

In the next, we will refer to $T^\star$ as a time after which all the sensors have stabilized:

$$\begin{aligned}T^\star = \min \big\{ \widetilde{t} \in \mathbb{N} | &\operatorname{supp}(x_v(t)) = \operatorname{supp}(x_v(\widetilde{t})), \\ &\forall t \geq \widetilde{t} \text{ and } \forall v \in \mathcal{V} \big\}\end{aligned}$$

Lemma 3. *The sequence $\{X(t)\}_{t\in\mathbb{N}}$ is bounded.*

Proof. As has already been noticed, the support of each column in $X(t)$ does not change for $t \geq T^\star$. Let us denote $S_v = \operatorname{supp}(x_v(T^\star))$ for all $v \in \mathcal{V}$ and let $\widetilde{x}_v^{S_v}$ be such

$$\widetilde{x}_v^{S_v} = \operatorname*{argmin}_{z\in\mathbb{R}^{S_v}} \|A_v z - y_v\|_2^2 \quad \forall v \in \mathcal{V}.$$

Then the update of the non-zero components of v-th column in $X(t)$ (which is denoted with $x_v(t+1)|_{S_v}$) becomes a convex combination between some of the columns in $X(t)$ restricted to the indexes in S_v (denoted with $x_v(t)|_{S_v}$) and $\widetilde{x}_v^{S_v}$. We can therefore write that

$$x_v(t+1)|_{S_v} \in \operatorname{co}\left(\{\widetilde{x}_v^{S_v}\} \bigcup_{w\in N_v} \{x_w(t)|_{S_v}\} \right)$$

and

$$x_v(t+1)|_{S_v^{\mathsf{C}}} = 0,$$

where $\operatorname{co}(\Omega)$ is the convex hull of the set Ω, i.e. the smallest convex set that contains points in Ω. If we iterate the argument for $v \in \mathcal{V}$ then we conclude that $\{x_v(t)\}_{t\in\mathbb{N}}$ is bounded for all $v \in \mathcal{V}$ and so is $\{X(t)\}_{t\in\mathbb{N}}$.

□

An important consequence of Corollary 1 is that the term $\sum_{v\in\mathcal{V}} \|x_v(t)\|_0$ in the functional $\mathcal{F}(X(t))$ remains constant for all $t \geq T^\star$. Then we can define a new function

$$\mathcal{G}(X) = \sum_{v\in\mathcal{V}} \left[q\|A_v x_v - y_v\|_2^2 + \frac{1-q}{d\tau} \sum_{w\in\mathcal{N}_v} \|\overline{x}_w - x_v\|_2^2 \right] \tag{19}$$

and a constant χ such that

$$\chi + \mathcal{G}(X(t)) = \mathcal{F}(X(t)) \qquad \forall t \geq T^\star. \tag{20}$$

In an analogous way we can define a new map Ψ which act element-wise on X:

$$\begin{aligned}[\Psi(X)]_{i,v} = (1-q) \sum_{w\in\mathcal{N}_i} (P^2)_{v,w} x_{i,w}(t) + q x_{i,v}(t) \\ + q\tau A_{v,i}^{\mathsf{T}} (y_v - A_v x_v)\end{aligned} \tag{21}$$

It should be noticed that Ψ is a continuous map and $\Psi X(t) = \Gamma X(t)$ for all $t \geq T^\star$.

If we consider the system after time $T^\star$, we can thus apply the LaSalle invariance principle for discrete-time dynamical systems:

Theorem 5 (LaSalle invariance principle, [48])**.** *Let Ψ be a discrete-time dynamical system on a space Ω. Assume that*

(i) there exists a closed set $W \subset \Omega$ such that any evolution of the system with initial condition in W remains in W for all subsequent time;

(ii) there exists a function $\mathcal{G} : \Omega \to \mathbb{R}$ from which is nonincreasing along Ψ in W;

(iii) the evolution of Ψ in W is bounded;

(iv) Ψ and $\mathcal{G}$ are continuous functions.

Then any evolution of the system in W tends to a subset of $\{w \in W | \mathcal{G}(\Psi w) = \mathcal{G}(w)\}$.

This is a formulation of the principle which is sufficient for our purposes; for more details, see [48, Theorem 1.19].

Before providing the proof of Theorem 3 for hard thresholding, we present the following technical lemma, which is proved in Appendix B.

Lemma 4. *The minima of $\mathcal{F}(X)$ are isolated.*

We now conclude the proof of Theorem 3. First, let us notice that our system $X(t)$ fulfills the hypotheses of the LaSalle invariance principle: $X(t)$ is bounded by Lemma 3, hence we can define a compact set W from which it never exits; $\{\mathcal{G}(X(t))\}_{t\in\mathbb{N}}$ is nonincreasing by (19), (20) and Lemma 2; Ψ and $\mathcal{G}$ are continuous. Therefore, $X(t)$ converges to a set Ω such that, for any $\omega \in \Omega$, $\mathcal{G}(\Gamma\omega) = \mathcal{G}(\omega)$.

Now, let us show that $\mathcal{G}(X(t)) > \mathcal{G}(X(t+1))$ unless $X(t) = X(t+1)$. First, we notice that $X(t)$ is the unique minimizer of

$$\sum_{v\in\mathcal{V}} \left(\frac{1}{\tau} \|x_v(t) - b_v\|_2^2 - \|A_v(x_v(t) - b_v)\|_2^2 \right)$$

with respect to variable $B \in \mathbb{R}^{n\times|\mathcal{V}|}$. In fact, each term in the summation is nonnegative, and is equal to zero if and only if $b_v = x_v(t)$. As a consequence, if $X(t) \neq X(t+1)$, then

$$\begin{aligned}\mathcal{G}(X(t)) &\geq \mathcal{G}(X(t+1), \overline{X}(t+1), X(t)) \\ &> \mathcal{G}(X(t+1), \overline{X}(t+1), X(t+1)) \\ &= \mathcal{G}(X(t)).\end{aligned}$$

In conclusion, $\mathcal{G}(X(t+1)) = \mathcal{G}(X(t))$ if and only if $X(t+1) = X(t)$. Then, $X(t)$ converges to the set (or to a subset) of the fixed points of Γ. This does not imply the convergence of $X(t)$, as $X(t)$ might approach the set without never entering it.

However, fixed points are isolated, thus $X(t)$ necessarily tends to a single point. In fact, as the fixed points correspond to the minima of $\mathcal{F}(X)$ (see Theorem 1) and the minima of $\mathcal{F}(X)$ are isolated (see Lemma 4), i.e., if $U^\star \in \mathbb{R}^{n\times|\mathcal{V}|}$ is a minimum for $\mathcal{F}$, then $\mathcal{F}(U^\star + H) > F(U^\star)$ for a sufficiently small increment $H \in \mathbb{R}^{n\times|\mathcal{V}|} \neq 0$. In other terms, each minimum has a neighborhood in which no other minima occur.

C. Convergence of DISTA (Proof of Theorem 4)

In Section V-A we have proved that two successive iterations of DISTA become closer and closer. We now prove that the full sequence converges to a minimizer of the cost function $\mathcal{F}_1$. This part concludes the proof of Theorem 4.

More precisely, the proof of the convergence is obtained through intermediate steps:

- the sequence $\{x_v(t)\}_{t\in\mathbb{N}}$ is bounded (Lemma 5);
- the map Γ is nonexpansive (Lemma 6);
- the proof is concluded by applying the Opial's Theorem (see Theorem 6).

Lemma 5. *The sequence $\{X(t)\}_{t\in\mathbb{N}}$ is bounded.*

Proof. It is easy to show that

$$\mathcal{F}(X(0)) \geq \mathcal{F}(X(t)) \geq \sum_v \|x_v(t)\|_1 = \|X(t)\|_\infty.$$

Therefore, $\{X(t)\}_{t\in\mathcal{N}}$ is bounded. □

We prove that the sequence of the $\{X(t)\}_{t\in\mathbb{N}}$ converges to a fixed point of Γ, applying the Opial's Theorem to the operator Γ;

Theorem 6 (Opial's Theorem [49])**.** *Let T be an operator from a finite-dimensional space $\mathbb{S}$ to itself that satisfies the following conditions:*

1) *T is asymptotically regular (i.e., for any $x \in \mathbb{S}$, and for $t \in \mathbb{N}$, $\left\|T^{t+1}x - T^t x\right\| \to 0$ as $t \to \infty$);*
2) *T is non expansive (i.e., $\|Tx - Tz\| \leq \|x - z\|$ for any $x, z \in \mathbb{S}$);*
3) *$\mathrm{Fix}(T) \neq \emptyset$, $\mathrm{Fix}(T)$ being the set of fixed point of T.*

Then, for any $x \in \mathbb{S}$, the sequence $\{T^t(x)\}_{t\in\mathbb{N}}$ converges weakly to a fixed point of T.

It should be noticed that in $\mathbb{R}^n$ the weak convergence coincides with the strong convergence.

Let us now prove that Γ satisfies the Opial's conditions. It should be noted that the asymptotic regularity of Γ has already been proved in Section V-A. We now prove that Γ is nonexpansive.

Lemma 6. *For any $\tau \leq \min_{v\in\mathcal{V}} \|A_v\|_2^{-2}$, Γ_1 defined in* (10) *is nonexpansive.*

Proof. Since η_α is nonexpansive, for any $X, Z \in \mathbb{R}^{n\times|\mathcal{V}|}$,

$$\begin{aligned}&\|(\Gamma X)_v - (\Gamma Z)_v\|_2^2 \\ &\leq \left\|(1-q)(\overline{\overline{x}}_v - \overline{\overline{z}}_v) + q(I - \tau A_v^\mathsf{T} A_v)(x_v - z_v)\right\|_2^2 \\ &\leq \left[(1-q)\left\|\overline{\overline{x}}_v - \overline{\overline{z}}_v\right\|_2 + q\left\|I - \tau A_v^\mathsf{T} A_v\right\|_2 \|x_v - z_v\|_2\right]^2.\end{aligned}$$

Notice that $I - \tau A_v^\mathsf{T} A_v$ always has the eigenvalue 1 with algebraic multiplicity $n-m$, as the rank of A_v is m. Moreover, if $\tau < \|A_v\|_2^{-2}$, $I - \tau A_v^\mathsf{T} A_v$ is positive definite and its spectral radius is 1. Since $I - \tau A_v^\mathsf{T} A_v$ is a symmetric matrix, we

then have $\left\|I-\tau A_v^{\mathsf{T}}A_v\right\|_2=1$. Thus, applying the triangular inequality,

$$\begin{aligned}
&\|(\Gamma X)_v-(\Gamma Z)_v\|_2^2\\
&\leq \left[(1-q)\left\|(\overline{\overline{x}}_v-\overline{\overline{z}}_v)\right\|_2+q\left\|x_v-z_v\right\|_2\right]^2\\
&\leq \left[\frac{1-q}{d^2}\sum_{w\in\mathcal{N}_v}\sum_{w'\in\mathcal{N}_w}\|x_{w'}-z_{w'}\|_2+q\left\|x_v-z_v\right\|_2\right]^2\\
&\leq \frac{(1-q)^2}{d^4}\left(\sum_{w\in\mathcal{N}_v}\sum_{w'\in\mathcal{N}_w}\|x_{w'}-z_{w'}\|_2\right)^2+q^2\left\|x_v-z_v\right\|_2^2\\
&\quad+\frac{2(1-q)q}{d^2}\sum_{w\in\mathcal{N}_v}\sum_{w'\in\mathcal{N}_w}\|x_{w'}-z_{w'}\|_2\left\|x_v-z_v\right\|_2.
\end{aligned}$$

Applying the Cauchy-Schwarz inequality

$$\left(\sum_{i=1}^N\beta_i\right)^2=\langle\beta,\mathbb{1}_N\rangle^2\leq\|\beta\|_2^2\|\mathbb{1}_N\|_2^2=N\sum_{i=1}^N\beta_i^2 \quad (22)$$

which holds for any $\beta=(\beta_1,\ldots,\beta_N)\in\mathbb{R}^N$, we obtain

$$\begin{aligned}
&\|(\Gamma X)_v-(\Gamma Z)_v\|_2^2\\
&\leq \frac{(1-q)^2}{d^2}\sum_{w\in\mathcal{N}_v}\sum_{w'\in\mathcal{N}_w}\|x_{w'}-z_{w'}\|_2^2+q^2\left\|x_v-z_v\right\|_2^2\\
&\quad+\frac{2(1-q)q}{d^2}\sum_{w\in\mathcal{N}_v}\sum_{w'\in\mathcal{N}_w}\|x_{w'}-z_{w'}\|_2\left\|x_v-z_v\right\|_2.
\end{aligned}$$

Finally, summing over all $v\in\mathcal{V}$ and considering that $2\|x_{w'}-z_{w'}\|_2\|x_v-z_v\|_2\leq\|x_{w'}-z_{w'}\|_2^2+\|x_v-z_v\|_2^2$,

$$\begin{aligned}
\|(\Gamma X)-(\Gamma Z)\|_F^2&=\sum_{v\in\mathcal{V}}\|(\Gamma X)_v-(\Gamma Z)_v\|_2^2\\
&\leq(1-q)^2\|X-Z\|_F^2+q^2\|X-Z\|_F^2\\
&\quad+2(1-q)q\|X-Z\|_F^2\\
&\leq\|X-Z\|_F^2.
\end{aligned}$$

□

We now conclude the proof of Theorem 4. The assertion follows by a direct application of Opial's Theorem, the numerical convergence (proved in Section V-A), Theorem 1 and Lemma 6. □

VI. Applications and numerical experiments

In this section, we describe two applications in which the performance of DIHTA and DISTA can be assessed. In the first example, we perform linear regression analysis on a prostate cancer dataset proposed in [4] and studied also in [18]; in the second one, we consider a sparse signal recovery problem using compressed sensing.

A. Analysis of a medical dataset

In [4], a real set of prostate cancer medical data is used to perform linear regression and infer the values of some parameters given some predictors. More precisely, $p=8$ predictors are considered: the log cancer volume (`lcavol`), the log prostate weight (`lweight`), the age, the log of the amount of benign prostatic hyperplasia (`lbph`), the seminal vesicle invasion (`svi`), the log of capsular penetration (`lcp`), the Gleason score (`gleason`), and the percent of Gleason scores 4 or 5 (`pgg45`). The purpose is to get the correlation between these predictors, measured in 97 patients who were going to receive a prostatectomy, and the level of prostate-specific antigen. The dataset is randomly splitted in a training subset of 67 patients and 30 test patients. The objective is to use the 67 training data in order to fit a sparse linear model after standardizing the predictors. The non-zero coefficients of the linear regression suggest which factors are more relevant in the generation of the antigen.

In [18], the problem is tackled in a distributed way, by subdividing the training data into 7 groups and performing distributed Lasso on them. The distributed setting is motivated as follows. Each group of data may be owned by a laboratory or hospital that does not want to share them for privacy or secrecy reasons. Nevertheless, each laboratory aims at improving its analysis by sharing its partial linear regression results with other laboratories. A network is then raised up, in which the estimated linear regression coefficients are repeatedly transmitted among laboratories and updated using its own measured data and coefficients estimated by the others. In this way, there is no sharing of the personal data of the patients, but only of the coefficients that describe the correlation between predictors and antigen.

As in [18], we assume that the network is composed by 7 laboratories, 6 of which have a dataset of 10 patients, and one with a dataset of 7 patients. We then apply DIHTA and DISTA to recover the linear regression coefficients, assuming the network can be modelled as a complete graph (but analogous results can be obtained with less connected network).

Tables I and II show the coefficients estimated by a number of different centralized and distributed methods. Both DISTA and DIHTA are run with $q=1/2,1/10,1/100$. For the soft algorithms, we set $\lambda/\tau=8.53$, using the cross validation strategy [18], while for the hard thresholding, $\lambda/\tau=1$. The values of τ are optimized according to the values of q.

A test set of $P=30$ patients is used to judge the performance of the selected model, in terms of Test Error and Standard Error, respectively defined as $\|y-\widehat{y}\|_2^2/P$ and $\|y-\widehat{y}\|_2/(P-p)$, where y is the vector of measured antigen values and $\widehat{y}$ its estimate according to the computed coefficients. DIHTA is compared (see Table I) with LS methods and iterative hard thresholding algorithm (IHTA, [16]), while DISTA (Table II) is set against LS, iterative soft thresholding algorithm (ISTA, [17]), ADMM and DSM. As explained in paragraph IV-B1, we implement a version of DSM which is not stopped [29] and with constant stepsize r [46], in order to perform a fair comparison with DISTA. However, such DSM is not proved to converge and actually numerical experiments show it oscillates after some time. We then choose a value of r sufficiently small to avoid oscillations in the considered time range.

Notice that LS does not seek a sparse solution, but is considered as well as optimal method when it is not necessary to identify only the most significant coefficients. From the presented results, we can infer that sparsification leads to

Term	LS	IHTA	DIHTA	DIHTA	DIHTA
q			$1/2$	$1/10$	$1/100$
Stepsize τ		10^{-3}	$1/150$	$1/35$	$2/7$
`Intercept`	2.464933	2.452345	2.452345	2.452345	2.452345
`lcavol`	0.679528	0.711628	0.705936	0.712205	0.712155
`lweight`	0.263053	0	0.013304	0	0
`age`	-0.141465	0	-0.007709	0	0
`lbph`	0.210147	0	0	0	0
`svi`	0.305201	0.224599	0.220353	0.224233	0.224265
`lcp`	-0.288493	0	0	0	0
`gleason`	-0.021305	0	0	0	0
`pgg45`	0.266956	0	0	0	0
Test Error	0.521274	0.387325	0.387369	0.387413	0.387402
Std Error	0.179751	0.154944	0.154953	0.154962	0.154960

Table I
PROSTATE CANCER DATASET: ESTIMATED COEFFICIENTS AND ERRORS FOR ℓ_0-ESTIMATION.

Term	LS	Lasso	ADMM	DSM	ISTA	DISTA	DISTA	DISTA
q						$1/2$	$1/10$	$1/100$
Stepsize τ or r				$2.27 \cdot 10^{-5}$	$8 \cdot 10^{-3}$	10^{-4}	$4 \cdot 10^{-4}$	$4 \cdot 10^{-3}$
`Intercept`	2.464933	2.452345	2.452345	2.452345	2.452345	2.452345	2.452345	2.452345
`lcavol`	0.679528	0.544762	0.544762	0.542247	0.544762	0.543956	0.544328	0.544360
`lweight`	0.263053	0.211593	0.211593	0.211666	0.211593	0.211396	0.211499	0.211507
`age`	-0.141465	0	0	0	0	0	0	0
`lbph`	0.210147	0.071391	0.071391	0.071986	0.071391	0.071285	0.071375	0.071379
`svi`	0.305201	0.143954	0.143954	0.146147	0.143954	0.144036	0.144075	0.144072
`lcp`	-0.288493	0	0	0	0	0.000215	0	0
`gleason`	-0.021305	0	0	0	0	0.000195	0	0
`pgg45`	0.266956	0.052545	0.052545	0.052814	0.052545	0.052335	0.052450	0.052458
Test Error	0.521274	0.454660	0.454660	0.454596	0.454660	0.454703	0.454672	0.454671
Std Error	0.179751	0.167873	0.167873	0.167861	0.167873	0.167881	0.167875	0.167875

Table II
PROSTATE CANCER DATASET: ESTIMATED COEFFICIENTS AND ERRORS FOR ℓ_1-ESTIMATION.

a smaller error and that our distributed methods achieve performance analogous to that of centralized procedures, not only when q is very small and then expected to approximate the ℓ_0/ℓ_1 estimators as proved in Theorem 1. In particular, we notice that better results are obtained by the hard thresholding procedures.

Let us also analyze the speed of convergence. In Figure 1 the mean-squared error from the LASSO solution of ADMM, ISTA, and DISTA are depicted as a function of the iterations. The decreasing curves confirm that all local estimates of DISTA approach the LASSO coefficients and converge to them when $q \to 0$ as stated in Theorem 1. This figure confirms that DISTA is a faster alternative to DSM, whose speed of convergence is shown to be slow [18].

In Figure 2, the evolution of the mean-squared error from the ℓ_0-estimator of DIHTA is depicted along time. As expected, increasing q we speed the convergence, and by a suitable choice of the parameters the final accuracy is comparable to that of IHTA for any choice of q. We finally remark that DIHTA turns out to be faster than DISTA, and as already noticed, the Test and Standard Errors are smaller. In this setting, we can then conclude than DIHTA works better than DISTA, at the price of a more sophisticated parameters' calibration. Finally, Figures 3 and 4 illustrate the evolution of the coefficients estimated by each laboratory in the network with $q = 1/2$ and $q = 1/10$, respectively, and $\tau = 10^{-3}$. It should be noted that the local estimates approach common values which are equal to LASSO coefficients (see black circles) and that disagreement is reduced when q is sufficiently small, as stated in Theorem 1. In other terms, at the end of the experiment, each laboratory will get approximately the same coefficients. The more consensus is needed, the more q has to be kept small, which may affects the convergence time. A suitable tradeoff may then be found according to the experiment requirements.

B. Sparse signal recovery

Let us consider a network of interconnected sensors modeled as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Each sensor acquires a sparse signal, represented in vector form as $x \in \mathbb{R}^n$. The sampling is performed at a rate below the Nyquist rate, using random linear projections as suggested by the compressed sensing theory [6]. One can represent the measurements $y_v \in \mathbb{R}^m$ (with $m \ll n$) as

$$y_v = A_v \widetilde{x} + \xi_v.$$

Under certain conditions [39], it is possible to recover $\widetilde{x}$ by solving (2) or (3) with $A^{\mathsf{T}} = (A_1^{\mathsf{T}}, \dots, A_{|\mathcal{V}|}^{\mathsf{T}})$ and $y^{\mathsf{T}} = (y_1^{\mathsf{T}}, \dots, y_{|\mathcal{V}|}^{\mathsf{T}})$; further details about the properties of (y, A) can be found in [6]. For purpose of illustration, the signal to be recovered is generated choosing k nonzero components uniformly at random among the $n = 150$ elements and drawing the amplitude of each nonzero component from a Gaussian distribution $\mathsf{N}(0, 1)$. The sensing matrices $(A_v)_{v \in \mathcal{V}}$ are sampled from the Gaussian ensemble with m rows, n

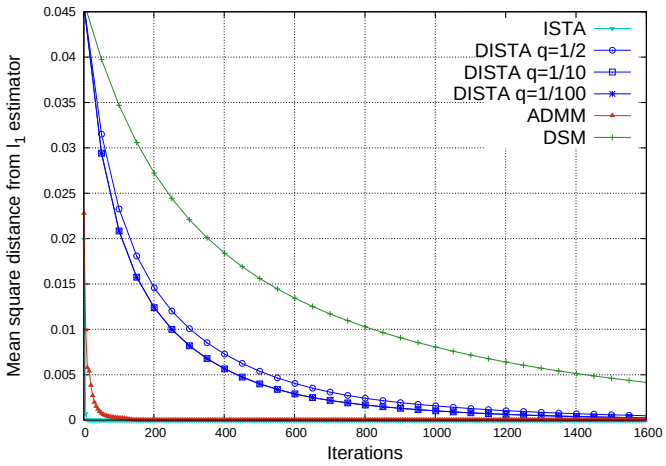

Figure 1. Prostate cancer dataset: evolution of the mean-squared error from the ℓ_1-estimator as a function of the iterations.

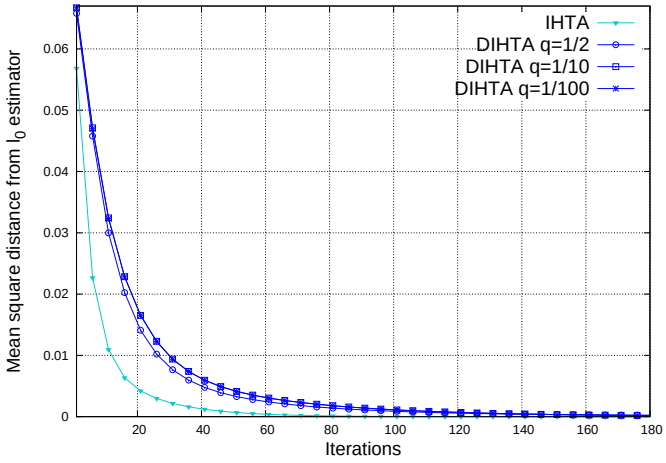

Figure 2. Prostate cancer dataset: evolution of the mean-squared error from the ℓ_0-estimator as a function of the iterations.

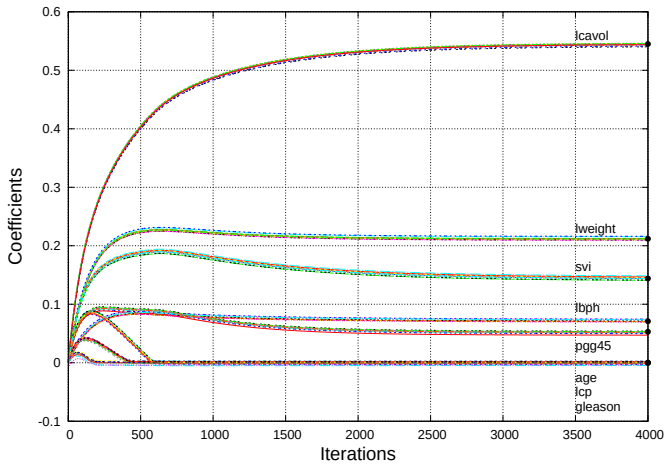

Figure 3. Prostate cancer dataset: evolution of the coefficients estimated by DISTA with $q = 0.5$. The black circles are the LASSO coefficients.

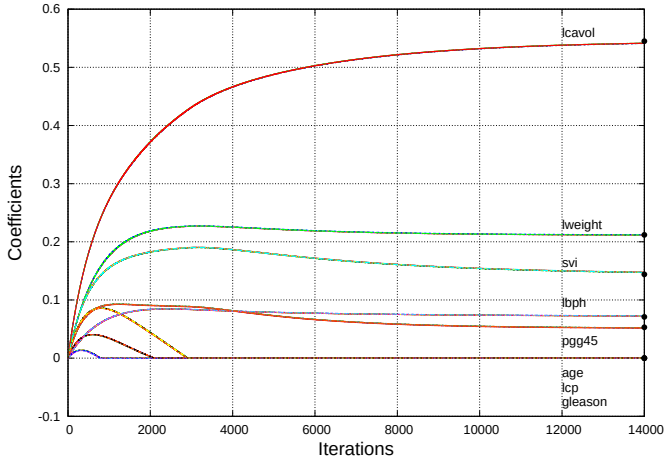

Figure 4. Prostate cancer dataset: evolution of the coefficients estimated by DISTA with $q = 0.1$. The black circles are the LASSO coefficients.

columns, zero mean and variance $\frac{1}{m}$. This is known to be a suitable choice from compressed sensing theory [6].

We now conduct a series of experiments for different architectures and for a variety of total number of measurements. Here, we are interested in the performance of the algorithms as a function of the number of data to store in the memory, which we try to minimize, and of the size of the graph. As already said, the available memory of wireless sensors is often very limited, typically few kB. We then recover the original signal employing DIHTA and DISTA. For each n, we vary the number of measurements m per node and the number of nodes in the network. For each $(n, m, |\mathcal{V}|)$ 3-tuple, we repeat the following procedures 50 times.

The measurements are then taken according to the model in (4). We use the so-called Metropolis random walk construction for P (see [50]) which amounts to the following: if $i \neq j$,

$$P_{ij} = \begin{cases} 0 & \text{if } (i,j) \notin \mathcal{E} \\ (\max\{\deg(i), \deg(j)\})^{-1} & \text{if } (i,j) \in \mathcal{E} \end{cases}$$

where $\deg(i)$ denotes the degree (the number of neighbors) of unit i in the graph $\mathcal{G}$.

In particular, we consider the following topologies:

1) *Complete graph*: $P_{ij} = \frac{1}{|\mathcal{V}|}$ for every $i, j = 1, \ldots, |\mathcal{V}|$.
2) *Ring graph*: $|\mathcal{V}|$ sensors are disposed on a circle, and each node communicates with its first neighbor on each side (left and right). The corresponding circulant symmetric matrix P is given by $P_{ij} = \frac{1}{3}$ for every $i = 2, \ldots, |\mathcal{V}|-1$ and $j \in \{i-1, i, i+1\}$; $P_{11} = P_{12} = P_{1|\mathcal{V}|} = \frac{1}{3}$; $P_{|\mathcal{V}|1} = P_{|\mathcal{V}||\mathcal{V}|-1} = P_{|\mathcal{V}||\mathcal{V}|} = \frac{1}{3}$; $P_{ij} = 0$ elsewhere.
3) *Random geometric graph*: sensors are assumed to be deployed uniformly at random in a square $[0,1] \times [0,1]$, and communication is allowed between sensors with distance below a certain radius (here we fix the radius to 0.75).

We show different numerical experiments to illustrate DIHTA and DISTA performance. For both, we define a success the case when there is a time t such that

$$\sum_{v \in \mathcal{V}} \|\widetilde{x} - x_v(t)\|_2^2 / (n|\mathcal{V}|) < 10^{-4}$$

where $\widetilde{x}$ is the original signal to be recovered and $x_v(t)$ is the estimate at time t given by our algorithms.

1) DIHTA performance: In the first experiment, we evaluate the performance of DIHTA in a noise-free setting, in

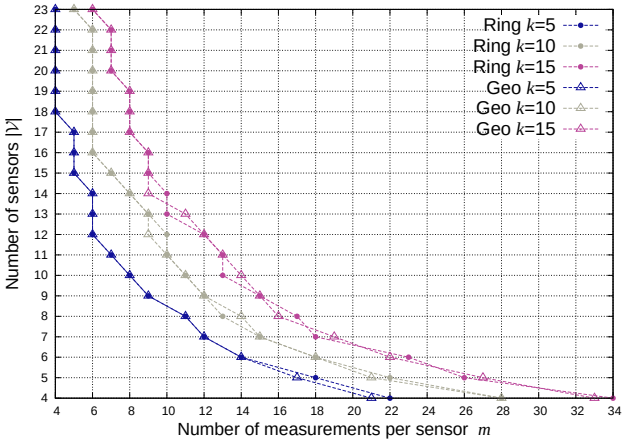

Figure 5. DIHTA performance (noise-free case, $n = 150$, $k \in \{5, 10, 15\}$): phase transition points in the $(m, \mathcal{V})$ plane. For the couples $(m, \mathcal{V})$ above the curves, the probability of success overcomes $\frac{1}{2}$.

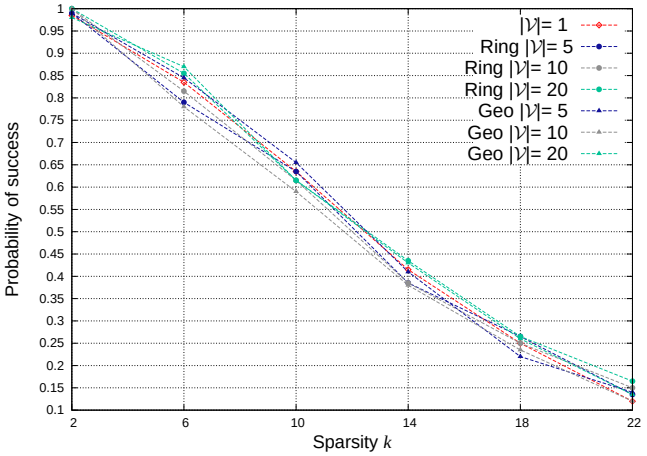

Figure 6. DIHTA performance (noise-free case, $n = 150$, $m|\mathcal{V}| = 120$): probability of success in function of the sparsity degree k.

terms of probability of success as a function of the number of sensors and of the measurements. We assess the probability of success by counting the number of successful events over 200 different runs. We consider ring and random geometric (radius 0.75) topologies; the used parameters are $\alpha \approx 0.005$ and $\tau \approx 1/|\mathcal{V}|$.

In Figure 5 we depict in the $(m, |\mathcal{V}|)$ plane the phase transition points for the probability of success, that is, the combination of m and $|\mathcal{V}|$ for which the probability of success overcomes $\frac{1}{2}$. The results are obtained over 100 runs. The signals to be reconstructed have sparsity degree $k \in \{5, 10, 15\}$; their supports are generated uniformly at random. Observing Figure 5, we evince that the phase transitions occur at $m|\mathcal{V}| \approx 80, 110, 140$ on average for $k = 5, 10, 15$, respectively. No significant difference is visible for ring and random geometric graphs: the topology does not dramatically affects the probability of success. In particular, we remark that using a non regular graph as the random geometric we always get convergence (which is not guaranteed by the theoretical analysis in Section V-B). We finally observe that for a fixed sparsity degree the total number $m|\mathcal{V}|$ of data required is approximately constant, and can be achieved equivalently by few sensors that train many data, or by many sensors that train few data. In other terms, the algorithm scales nicely to very large networks without suffering a performance loss.

The good scaling properties are further appreciable in Figure 6, where the probability of success is depicted as a function of the sparsity degree k (in the interval from 2 to 30). We consider different network sizes $|\mathcal{V}|$ and keep constant the total number of measurements $m|\mathcal{V}| = 120$. In Figure 6 we can see that a larger number of sensors does not affect the probability of success, that is, a larger decentralization is not a drawback. Similar results are obtained with different, even non regular, topologies (specifically, ring and random geometric with radius 0.75).

In conclusion, our experiments suggest that decentralization is not a drawback for the performance of the iterative hard thresholding, if the reconstruction is performed with DIHTA.

2) DISTA performance: In the compressed sensing setting, we now present the performance of DISTA, which behaves significantly better than DIHTA in terms of estimation accuracy. We show experiments in the noise-free and noisy cases, and we compare DISTA with DSM and consensus ADMM.

We first consider a noise-free scenario, and we illustrate the probability of success as a function of the number of measurements over complete, ring, and random geometric (radius 0.75) topologies (respectively, Figures 7, 8, and 9). The color of the cell in the figures reflects the empirical success rate: white denotes perfect reconstruction in all the experiments, while black represents no success occurrence. It should be noted that the number of total measurements $m|\mathcal{V}|$ which are sufficient for successful estimation is constant: the red curve collects the points $(m, |\mathcal{V}|)$ such that $m|\mathcal{V}| = 70$, which turns out to be a sufficient value to obtain good estimation (probability greater then 0.95). We observe that the performance of DISTA is not strongly affected by the graph topology. One could expect worse results with less connected graphs, since low connectivity may cause problems of scarce communication and slowness. However, our results show that only a slight degradation affects DISTA over the ring, while the behavior is very similar for the complete and random geometric graph. This also highlights that no loss occurs due to non-regularity of the graph.

In Figure 10 the probability of success of DISTA, ADMM, and DSM are compared as a function of the number of measurements per node. The curves are depicted for different numbers of sensors.

Finally, let us consider the noisy case. In Figure 11, the mean-squared error

$$\mathsf{MSE} = \frac{\sum_{v\in\mathcal{V}} \|\widetilde{x} - x_v^\star\|_2^2}{n|\mathcal{V}|},$$

averaged over 50 runs is plotted as a function of the signal-to-noise ratio

$$\mathsf{SNR} = \frac{\mathbb{E}\left[\sum_{v\in\mathcal{V}} \|y_v\|_2^2\right]}{\mathbb{E}\left[\sum_{v\in\mathcal{V}} \|\xi_v\|_2^2\right]}.$$

The number of sensors is $|\mathcal{V}| = 10$. The graph shows that DISTA performs better then DSM, even at larger compression

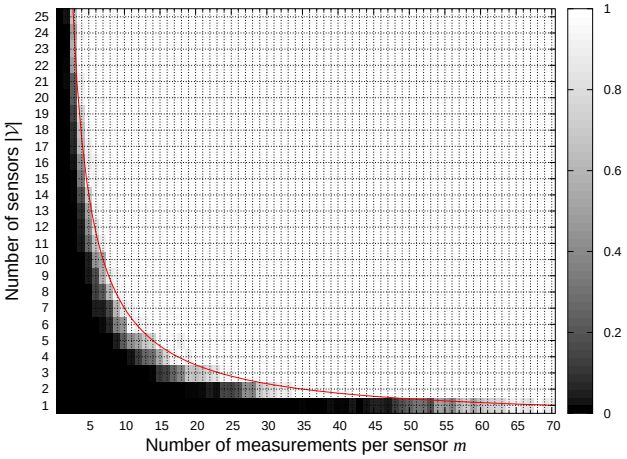

Figure 7. DISTA performance (noise-free case, $n = 150$, $k = 15$): probability of success over a complete graph.

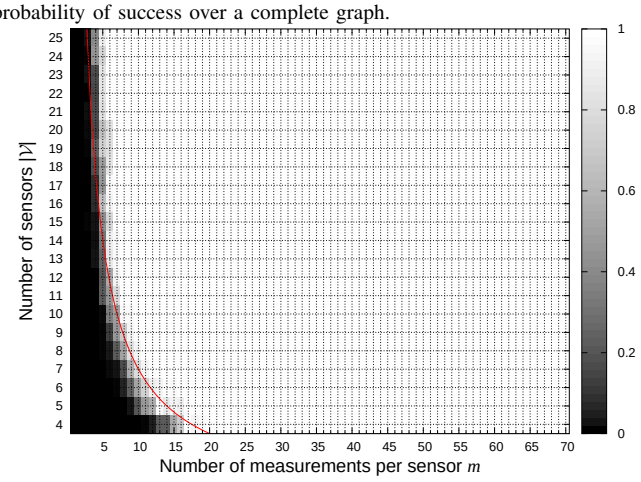

Figure 8. DISTA performance (noise-free case, $n = 150$, $k = 15$): probability of success over a ring graph.

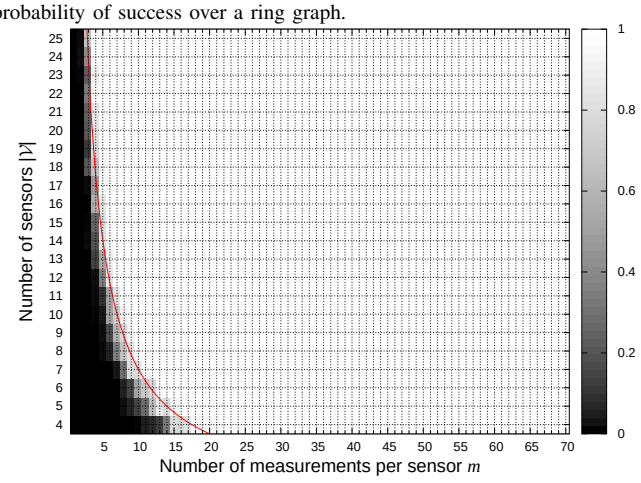

Figure 9. DISTA performance (noise-free case, $n = 150$, $k = 15$): probability of success over a random geometric graph (radius 0.75).

level: taking $m = 8$ is sufficient for DISTA to obtain a MSE lower than that obtained by DSM with $m = 12$ measurements. Notice that this is the best performance that can be obtained by DSM in this setting, that is, even without compression we do not see any improvement. On the other hand, DISTA with $m =$

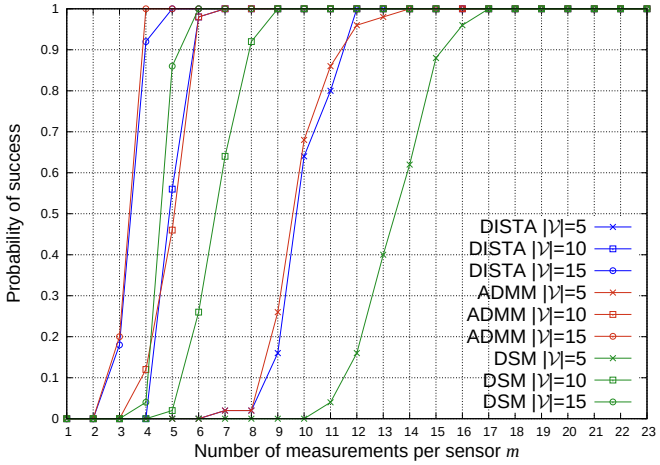

Figure 10. Noise-free case: DISTA vs ADMM vs DSM, complete graph, $n = 150$, $k = 15$.

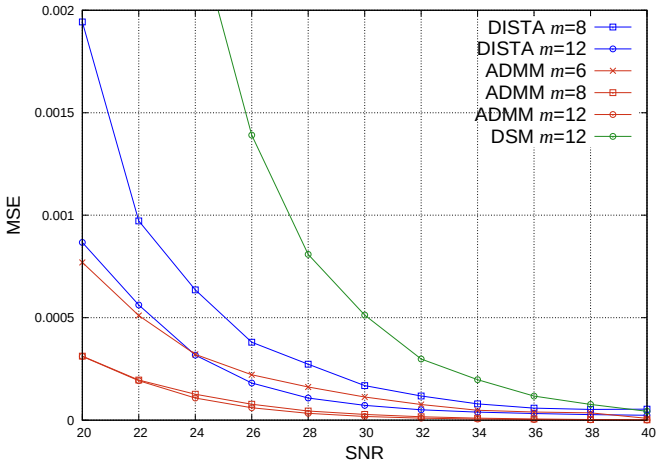

Figure 11. Noise case: DISTA vs ADMM vs DSM, complete graph, $n = 150$, $k = 15$, $|\mathcal{V}| = 10$.

12 is worse than ADMM with same m or with $m = 8$, but it is better than ADMM with $m = 6$ for sufficiently large SNR. In conclusion, DISTA can achieve the optimal performance of ADMM at the price of a smaller compression level, which is not achievable by DSM.

VII. Concluding Remarks

In this paper, we presented distributed algorithms for ℓ_0/ℓ_1-regularized linear inverse problems in multi-agent systems with limited communication capability. In this class of algorithms, each agent maintains an approximation of the model parameters. These estimates are communicated to the neighbors synchronously over a fixed connectivity structure. Each agent updates its own current estimate based on local information and its training data using an iterative thresholding method. This algorithm has low complexity and memory requirements, making it suitable for low-energy scenarios such as wireless sensor networks. Numerical results show that the proposed algorithm outperforms existing distributed schemes in terms of memory and complexity, and is almost as good as the ADMM method, but has much lower memory

requirements, making it more suitable for the target application. The main theoretical contribution includes the proof of convergence of the algorithm to a local minimum of the distributed regularized LS estimator.

The algorithms we considered assume that nodes can send and process signals synchronously. This is a restrictive assumption in many applications. The development of randomized algorithms for sparse approximations is the focus of our current work. Because of the randomization and the nonlinearity of the updates, the proof will necessitate different lines of analysis.

REFERENCES

[1] A. Björck, *Numerical methods for least squares problems*. SIAM, 1996.

[2] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The convex geometry of linear inverse problems.," *Foundations of Computational Mathematics*, vol. 12, no. 6, pp. 805 – 849, 2012.

[3] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "A sparse-group lasso," *J. Comput. Graph. Stat.*, vol 22,no. 2, pp.231 – 245, 2013.

[4] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning*. Springer, New York, 2003.

[5] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, New York, 2004.

[6] E. J. Candès, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Comm. Pure Appl. Math.,* , vol. 59, no. 8, pp. 1207 – 1223, 2006.

[7] E. J. Candès, "The restricted isometry property and its implications for compressed sensing," *C. R. Math. Acad. Sci. Paris, Ser. I*, vol. 346, pp. 589 – 592, 2008.

[8] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comp.*, vol. 20, pp. 33 – 61, 1998.

[9] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inform. Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[10] J. Tropp, "Just relax: convex programming methods for identifying sparse signals in noise," *IEEE Trans. Inform. Theory*, vol. 52, no. 3, pp. 1030 – 1051, 2006.

[11] A. N. Tikhonov and V. Y. Arsenin, *Solutions of Ill-Posed Problems*. V. H. Winston & Sons, Washington, D.C.: John Wiley & Sons, New York,, 1977.

[12] E. Chouzenoux, A. Jezierska, J. Pesquet, and H. Talbot, "A majorize-minimize subspace approach for $\ell_2 - \ell_0$ image regularization," *SIAM J. Imaging Sci.*, vol. 6, no. 1, pp. 563 – 591, 2013.

[13] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *J. Roy. Stat. Soc., Series B*, vol. 58, pp. 267 – 288, 1994.

[14] E. Huebner and R. Tichatschke, "Relaxed proximal point algorithms for variational inequalities with multi-valued operators," *Optimization Methods Software*, vol. 23, no. 6, pp. 847 – 877, 2008.

[15] T. Blumensath and M. E. Davies, "Iterative thresholding for sparse approximations," *J. Fourier Anal. Appl.*, vol. 14, no. 5, pp. 629 – 654, 2004.

[16] M. E. Blumensath, T.and Davies, "Iterative hard thresholding for compressed sensing," *Appl. Comput. Harmonic Anal.*, vol. 27, no. 3, pp. 265 – 274, 2009.

[17] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Comm. Pure Appl. Math.*, vol. 57, no. 11, pp. 1413 – 1457, 2004.

[18] G. Mateos, J. A. Bazerque, and G. B. Giannakis, "Distributed sparse linear regression," *IEEE Trans. Signal Proc.*, vol. 58, no. 10, pp. 5262 – 5276, 2010.

[19] J. Meng, H. Li, and Z. Han, "Sparse event detection in wireless sensor networks using compressive sensing."

[20] C. Feng, W. S. A. Au, S. Valaee, and Z. Tan, "Received-signal-strength-based indoor positioning using compressive sensing," *IEEE Trans. Mobile Computing*, vol. 11, no. 12, pp. 1983 – 1993, 2012.

[21] V. Cevher, M. F. Duarte, and R. G. Baraniuk, "Distributed target localization via spatial sparsity," in *in European Signal Processing Conference (EUSIPCO*, 2008.

[22] P. A. Forero, A. Cano, and G. B. Giannakis, "Consensus-based distributed support vector machines," *J. Mach. Learn. Res.*, vol. 99, pp. 1663 – 1707, 2010.

[23] K. Chaudhuri and C. Monteleoni, "Privacy-preserving logistic regression," in *NIPS* (D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, eds.), pp. 289 – 296, Curran Associates, Inc., 2008.

[24] J. D. Owens, D. Luebke, N. Govindaraju, M. Harris, J. Krüger, A. Lefohn, and T. J. Purcell, "A survey of general-purpose computation on graphics hardware," *Computer Graphics Forum*, vol. 26, no. 1, pp. 80 – 113, 2007.

[25] J. Krüger and R. Westermann, "Linear algebra operators for GPU implementation of numerical algorithms," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 908 – 916, 2003.

[26] D. Sundman, S. Chatterjee, and M. Skoglund, "A greedy pursuit algorithm for distributed compressed sensing," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2729 – 2732, 2012.

[27] J. Mota, J. Xavier, P. Aguiar, and M. Puschel, "Basis pursuit in sensor networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2916 – 2919, 2011.

[28] J. Mota, J. Xavier, P. Aguiar, and M. Puschel, "Distributed basis pursuit," *IEEE Trans. Signal Proc.*, vol. 60, no. 4, pp. 1942 – 1956, 2012.

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

[30] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1 – 122, 2011.

[31] A. I. Chen and A. Ozdaglar, "A fast distributed proximal-gradient method," in *Allerton Conference*, pp. 601 – 608, 2012.

[32] K. Yuan, Q. Ling, W. Yin, and A. Ribeiro, "A linearized Bregman algorithm for decentralized basis pursuit," in *Proceedings of the 21st European Signal Processing Conference (EUSIPCO)*, pp. 1 – 5, 2013.

[33] I. Lobel and A. Ozdaglar, "Distributed subgradient methods for convex optimization over random networks," *IEEE Trans. Automat. Contr.*, vol. 56, no. 6, pp. 1291 – 1306, 2011.

[34] S. S. Ram, A. Nedic, and V. V. Veeravalli, "Distributed subgradient projection algorithm for convex optimization," in *IEEE ICASSP*, pp. 3653 – 3656, 2009.

[35] I. Daubechies, M. Fornasier, and I. Loris, "Accelerated projected gradient method for linear inverse problems with sparsity constraints," 2004.

[36] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, "Algorithms for simultaneous sparse approximation: part i: Greedy pursuit," *Signal Process.*, vol. 86, no. 3, pp. 572 – 588, 2006.

[37] R. J. Tibshirani, "The lasso problem and uniqueness," *Electron. J. Statist.*, vol. 7, no. 0, pp. 1456 – 1490, 2013.

[38] M. Nikolova, "Description of the minimizers of least squares regularized with ℓ_0-norm. Uniqueness of the global minimizer," *SIAM J. Imaging Sci.*, vol. 6, no. 2, pp. 904 – 937, 2013.

[39] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, pp. 1289 – 1306, 2006.

[40] M. Fornasier, *Theoretical Foundations and Numerical Methods for Sparse Recovery*. Radon Series on Computational and Applied Mathematics, 2010.

[41] K. Lange, D. Hunter, and I. Yang, "Optimization transfer using surrogate objective functions," *J. Comput. Graph. Stat.*, vol. 9, pp. 1 – 20, 2006.

[42] K. Lange, *Optimization*. Springer Verlag, 2004.

[43] W. E. and A. Ozdaglar, "Distributed alternating direction method of multipliers," in *IEEE 51st Annual Conference on Decision and Control (CDC)*, pp. 5445 – 5450, 2012.

[44] J. Mota, J. Xavier, P. Aguiar, and M. Puschel, "D-admm: A communication-efficient distributed algorithm for separable optimization," *IEEE Trans. Signal Proc.*, vol. 61, no. 10, pp. 2718 – 2723, 2013.

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

[46] S. S. Ram, A. Nedic, and V. V. Veeravalli, "Distributed stochastic subgradient projection algorithms for convex optimization," *J. Optim. Theory Appl.*, pp. 516 – 545, 2010.

[47] C. Moallemi and B. Van Roy, "Consensus propagation," *IEEE Trans. Inform. Theory*, vol. 52, no. 11, pp. 4753 – 4766, Nov.

[48] F. Bullo, J. Cortes, and S. Martinez, *Distributed Control of Robotic Networks*. Applied Mathematics Series, Princeton University Press, 2009.

[49] Z. Opial, "Weak convergence of the sequence of successive approximations for nonexpansive mappings," *Bull. Amer. Math. Soc.*, vol. 73, pp. 591 – 597, 1967.

[50] L. Xiao, S. Boyd, and S. Lall, "Distributed average consensus with time-varying metropolis weights." http://web.stanford.edu/∼boyd/papers/avg_metropolis.html, 2006.

APPENDIX

A. Proof of Lemma 1

We provide the proof of statements 1) and 2) in Lemma 1 for DIHTA ($p = 0$). The validity of the statements can be verified for DISTA ($p = 1$) with similar arguments.

1) Let $U = (u_1, \dots, u_{|\mathcal{V}|})$ be a minimizer of $\mathcal{F}_0^{\mathcal{S}}(\cdot, C, B)$ for fixed $C, B \in \mathbb{R}^{n\times|\mathcal{V}|}$, that is

$$u_v = \eta_\alpha \left[(1-q)\overline{c}_v + q(b_v + \tau A_v^{\mathsf{T}}(y_v - A_v b_v))\right].$$

For any $H = (h_1, \dots, h_{|\mathcal{V}|}) \in \mathbb{R}^{n\times|\mathcal{V}|}$,

$$\begin{aligned}
&\mathcal{F}_0^{\mathcal{S}}(U+H, C, B) - \mathcal{F}_0^{\mathcal{S}}(U, C, B)\\
&\quad= \frac{1}{\tau}\|H\|_F^2 + \sum_{v\in\mathcal{V}}\left[\frac{2\alpha}{\tau}\|u_v + h_v\|_0 - \frac{2\alpha}{\tau}\|u_v\|_0\right.\\
&\quad\left. + 2\langle h_v, qA_v^{\mathsf{T}}(A_v b_v - y_v) - \frac{1-q}{\tau}\overline{c}_v + \frac{1}{\tau}u_v - \frac{q}{\tau}b_v\rangle\right].
\end{aligned} \tag{23}$$

Now, let $I_{1,v} = \mathrm{supp}(u_v)$ and $I_{0,v} = \{1, \dots, n\} \setminus I_{1,v}$. Then,

$$\begin{aligned}
&\mathcal{F}_0^{\mathcal{S}}(U+H, C, B) - \mathcal{F}_0^{\mathcal{S}}(U, C, B)\\
&= \frac{1}{\tau}\|H\|_F^2 + \sum_{v\in\mathcal{V}}\sum_{j\in I_{0,v}}\frac{2\alpha}{\tau}|h_{jv}|^0 +\\
&+ \sum_{v\in\mathcal{V}}\sum_{j\in I_{1,v}}\frac{2\alpha}{\tau}\left(|u_{jv}+h_{jv}|^0 - |u_{jv}|^0\right)\\
&+ \sum_{v\in\mathcal{V}}\left\{\sum_{j\in I_{0,v}} 2h_{jv}\left(qA_v^{\mathsf{T}}(A_v b_v - y_v) - \frac{1-q}{\tau}\overline{c}_v - \frac{q}{\tau}b_v\right)_j\right.\\
&\left.+\sum_{j\in I_{1,v}} 2h_{jv}\left(qA_v^{\mathsf{T}}(A_v b_v - y_v) - \frac{1-q}{\tau}\overline{c}_v + \frac{1}{\tau}u_v - \frac{q}{\tau}b_v\right)_j\right\}
\end{aligned}$$

Since $u_v = \eta_{0,\alpha}\left[(1-q)\overline{c}_v + q\left(b_v + \tau A_v^{\mathsf{T}}(y_v - A_v b_v)\right)\right]$, if $u_{jv} = 0$, then

$$\left|\left(qA_v^{\mathsf{T}}(A_v b_v - y_v) - \frac{1-q}{\tau}\overline{c}_v - \frac{q}{\tau}b_v\right)_j\right| \le \frac{\sqrt{2\alpha}}{\tau}.$$

Thus,

$$\begin{aligned}
&\sum_{j\in I_{0,v}} 2h_{jv}\left(qA_v^{\mathsf{T}}(A_v b_v - y_v) - \frac{1-q}{\tau}\overline{c}_v - \frac{q}{\tau}b_v\right)_j + \frac{2\alpha}{\tau}|h_{jv}|^0\\
&\ge \sum_{j\in I_{0,v}} -2\frac{\sqrt{2\alpha}}{\tau}|h_{jv}| + \frac{2\alpha}{\tau}|h_{jv}|^0 \ge 0
\end{aligned}$$

whenever $|h_{jv}| \le \frac{\sqrt{2\alpha}}{2}$. Otherwise, if $u_{jv} \neq 0$,

$$u_{jv} = \left((1-q)\overline{c}_v + q\left(b_v + \tau A_v^{\mathsf{T}}(y_v - A_v b_v)\right)\right)_j,$$

hence

$$\begin{aligned}
&2h_{jv}\left(qA_v^{\mathsf{T}}(A_v b_v - y_v) - \frac{1-q}{\tau}\overline{c}_v + \frac{1}{\tau}u_v - \frac{q}{\tau}b_v\right)_j\\
&+ \frac{2\alpha}{\tau}|u_{jv}+h_{jv}|^0 - \frac{2\alpha}{\tau}|u_{jv}|^0\\
&= \frac{2\alpha}{\tau}|u_{jv}+h_{jv}|^0 - \frac{2\alpha}{\tau}|u_{jv}|^0 = 0
\end{aligned}$$

whenever $|h_{jv}| < |u_{jv}|$. In conclusion, for any $v \in \mathcal{V}$, the sums over $I_{0,v}$ and $I_{1,v}$ are both non negative. Therefore,

$$\mathcal{F}_0^{\mathcal{S}}(U+H, C, B) \ge \mathcal{F}_0^{\mathcal{S}}(U, C, B) + \frac{1}{\tau}\|H\|_F^2. \tag{24}$$

Let us now consider a fixed point $U^\star$ of Γ, that is,

$$u_v^\star = \eta_\alpha\left[(1-q)\overline{\overline{u^\star}}_v + q\left(u_v^\star + A_v^{\mathsf{T}}(y_v - A_v u_v^\star)\right)\right],$$

where $\overline{\overline{u^\star}}_v = [U^\star(P^{\mathsf{T}})^2]_v$. We know that $U^\star$ is a minimizer of $\mathcal{F}_0^{\mathcal{S}}(\cdot, \overline{U^\star}, U^\star)$ and by (24)

$$\begin{aligned}
&\mathcal{F}_0^{\mathcal{S}}(U^\star + H, \overline{U^\star}, U^\star)\\
&\quad\ge \mathcal{F}_0^{\mathcal{S}}(U^\star, \overline{U^\star}, U^\star) + \frac{1}{\tau}\|H\|_F^2\\
&\quad= \mathcal{F}(U^\star) + \frac{1}{\tau}\|H\|_F^2.
\end{aligned} \tag{25}$$

2) Let $U^\star \in \mathrm{Fix}(\Gamma)$ and $\overline{h}_v = \frac{1}{d}\sum_{w\in\mathcal{N}_v} h_w$.

$$\begin{aligned}
&\mathcal{F}_0^{\mathcal{S}}(U^\star + H, \overline{U^\star}, U^\star) - \mathcal{F}_0^{\mathcal{S}}(U^\star + H, \overline{U^\star} + \overline{H}, U^\star + H)\\
&\quad= \sum_{v\in\mathcal{V}}\left[\frac{1-q}{d\tau}\sum_{w\in\mathcal{N}_v}\left\|u_v^\star + h_v - \overline{u_w^\star}\right\|_2^2 + \frac{q}{\tau}\|h_v\|_2^2\right.\\
&\quad\left. - q\|A_v h_v\|_2^2 - \frac{1-q}{d\tau}\sum_{w\in\mathcal{N}_v}\left\|u_v^\star + h_v - \overline{u^\star}_w - \overline{h}_w\right\|_2^2\right]\\
&\quad= \sum_{v\in\mathcal{V}}\left[2\frac{1-q}{d\tau}\sum_{w\in\mathcal{N}_v}\langle\overline{h}_w, u_v^\star + h_v - \overline{u_w^\star}\rangle\right.\\
&\quad- \frac{1-q}{d\tau}\sum_{w\in\mathcal{N}_v}\left\|\overline{h}_w\right\|_2^2 + \frac{q}{\tau}\|h_v\|_2^2 - q\|A_v h_v\|_2^2\Big]\\
&\quad\le \frac{1-q}{d\tau}\sum_{v\in\mathcal{V}}\sum_{w\in\mathcal{N}_v}\left[-\left\|\overline{h}_w\right\|_2^2 + 2\langle\overline{h}_w, u_v^\star + h_v - \overline{u^\star}_w\rangle\right]\\
&\quad+ \frac{q}{\tau}\sum_{v\in\mathcal{V}}\|h_v\|_2^2.
\end{aligned} \tag{26}$$

Now, notice that, being the graph regular

$$\begin{aligned}
&\sum_{v\in\mathcal{V}}\sum_{w\in\mathcal{N}_v}\langle\overline{h}_w, u_v^\star - \overline{u^\star}_w\rangle\\
&\quad= \sum_{v\in\mathcal{V}}\sum_{w\in\mathcal{N}_v}\langle\overline{h}_w, u_v^\star\rangle - \sum_{v\in\mathcal{V}}\sum_{w\in\mathcal{N}_v}\langle\overline{h}_w, \overline{u^\star}_w\rangle\\
&\quad= \sum_{v\in\mathcal{V}}\langle\overline{h}_v, \sum_{w\in\mathcal{N}_v} u_w^\star\rangle - \sum_{v\in\mathcal{V}}\sum_{w\in\mathcal{N}_v}\langle\overline{h}_w, \overline{u^\star}_w\rangle\\
&\quad= \sum_{v\in\mathcal{V}}\langle\overline{h}_v, \sum_{w\in\mathcal{N}_v} u_w^\star\rangle - d\sum_{v\in\mathcal{V}}\langle\overline{h}_v, \overline{u^\star}_v\rangle = 0.
\end{aligned} \tag{27}$$

Hence by the last expression of (26) we have:

$$\begin{aligned}
&\mathcal{F}_0^{\mathcal{S}}(U^\star+H,\overline{U^\star},U^\star)-\mathcal{F}_0^{\mathcal{S}}(U^\star+H,\overline{U^\star}+\overline{H},U^\star+H)\\
&\sum_{v\in\mathcal{V}}\left[\frac{1-q}{d\tau}\sum_{w\in\mathcal{N}_v}\left(-\left\|\overline{h}_w\right\|_2^2+2\langle\overline{h}_w,h_v\rangle\right)+\frac{q}{\tau}\left\|h_v\right\|_2^2\right]\\
&=\sum_{v\in\mathcal{V}}\left[\frac{1-q}{d\tau}\sum_{w\in\mathcal{N}_v}\left(-\left\|\overline{h}_w\right\|_2^2+2\langle\overline{h}_w,h_v\rangle\pm\left\|h_v\right\|_2^2\right)+\right.\\
&\qquad\left.+\frac{q}{\tau}\left\|h_v\right\|_2^2\right]\\
&=\sum_{v\in\mathcal{V}}\left[\frac{1-q}{d\tau}\sum_{w\in\mathcal{N}_v}\left(-\left\|\overline{h}_w-h_v\right\|_2^2+\left\|h_v\right\|_2^2\right)+\frac{q}{\tau}\left\|h_v\right\|_2^2\right]\\
&\leq\frac{1}{\tau}\sum_{v\in\mathcal{V}}\left\|h_v\right\|_2^2=\frac{1}{\tau}\left\|H\right\|_F^2
\end{aligned}$$

which concludes the proof. □

B. Proof of Lemma 4

$$\begin{aligned}
&\mathcal{F}(U^\star+H)-F(U^\star)=\\
&=\sum_{v\in\mathcal{V}}\left[\frac{2\alpha}{\tau}\left\|u_v^\star+h_v\right\|_0-\frac{2\alpha}{\tau}\left\|u_v^\star\right\|_0\right.\\
&\quad+q\left\|A_v(u_v^\star+h_v)-y_v\right\|_2^2-q\left\|A_vu_v^\star-y_v\right\|_2^2+\\
&\quad+\frac{1-q}{d\tau}\sum_{w\in\mathcal{N}_v}\left\|u_v^\star+h_v-\overline{u^\star}_w-\overline{h}_w\right\|_2^2\\
&\quad\left.-\frac{1-q}{d\tau}\sum_{w\in\mathcal{N}_v}\left\|u_v^\star-\overline{u^\star}_w\right\|_2^2\right]\\
&=\sum_{v\in\mathcal{V}}\left[\frac{2\alpha}{\tau}\left\|u_v^\star+h_v\right\|_0-\frac{2\alpha}{\tau}\left\|u_v^\star\right\|_0+q\left\|A_vh_v\right\|_2^2\right.\\
&\quad+2q\langle h_v,A_v^{\mathsf{T}}(A_vu_v^\star-y_v)\rangle+\\
&\quad+\frac{1-q}{d\tau}\sum_{w\in\mathcal{N}_v}\left\|h_v-\overline{h}_w\right\|_2^2\\
&\quad\left.+2\frac{1-q}{d\tau}\sum_{w\in\mathcal{N}_v}\langle h_v-\overline{h}_w,u_v^\star-\overline{u^\star}_w\rangle\right].
\end{aligned}\tag{28}$$

From (27), $\sum_{v\in\mathcal{V}}\sum_{w\in\mathcal{N}_v}\langle\overline{h}_w,u_v^\star-\overline{u^\star}_w\rangle=0$. Hence,

$$\begin{aligned}
&\mathcal{F}(U^\star+H)-\mathcal{F}(U^\star)=\\
&=\sum_{v\in\mathcal{V}}\left[\frac{2\alpha}{\tau}\left\|u_v^\star+h_v\right\|_0-\frac{2\alpha}{\tau}\left\|u_v^\star\right\|_0+q\left\|A_vh_v\right\|_2^2+\right.\\
&+\frac{1-q}{d\tau}\sum_{w\in\mathcal{N}_v}\left\|h_v-\overline{h}_w\right\|_2^2+2\langle h_v,\frac{1-q}{\tau}(u_v^\star-\overline{\overline{u^\star}}_v)\rangle\\
&\left.+2\langle h_v,qA_v^{\mathsf{T}}(A_vu_v^\star-y_v)\rangle\right].
\end{aligned}\tag{29}$$

Recalling that $U^\star$ is also a fixed point for Γ and considering the sets $I_{1,v}=\operatorname{supp}(u_v^\star)$ and $I_{0,v}=\{1,\dots,n\}\setminus I_{1,v}$ we obtain:

$$\begin{aligned}
&\mathcal{F}(U^\star+H)-\mathcal{F}(U^\star)\\
&=\sum_{v\in\mathcal{V}}\left\{q\left\|A_vh_v\right\|_2^2+\frac{1-q}{d\tau}\sum_{w\in\mathcal{N}_v}\left\|h_v-\overline{h}_w\right\|_2^2+\right.\\
&+\sum_{j\in I_{0,v}}\left[\frac{2\alpha}{\tau}\left|h_{jv}\right|^0-2\frac{1-q}{\tau}h_{jv}\overline{\overline{u^\star}}_v\right.\\
&\left.+2qh_{jv}A_v^{\mathsf{T}}(A_vu_v^\star-y_v)\right)_j\Big]\\
&+\sum_{j\in I_{1,v}}\left[\frac{2\alpha}{\tau}\left|u_{jv}+h_{jv}\right|^0-\frac{2\alpha}{\tau}\left|u_{jv}\right|^0+\right.\\
&\left.\left.+2h_{jv}\left(\frac{1-q}{\tau}(u_v^\star-\overline{\overline{u^\star}}_v)+qA_v^{\mathsf{T}}(A_vu_v^\star-y_v)\right)_j\right]\right\}\\
&\geq\sum_{v\in\mathcal{V}}\left\{q\left\|A_vh_v\right\|_2^2+\frac{1-q}{d\tau}\sum_{w\in\mathcal{N}_v}\left\|h_v-\overline{h}_w\right\|_2^2\right.\\
&+\sum_{j\in I_{0,v}}\left[\frac{2\alpha}{\tau}\left|h_{jv}\right|^0-2|h_{jv}|\frac{\sqrt{2\alpha}}{\tau}\right]\\
&\left.+\sum_{j\in I_{1,v}}\left[\frac{2\alpha}{\tau}\left|u_{jv}+h_{jv}\right|^0-\frac{2\alpha}{\tau}\left|u_{jv}\right|^0\right]\right\}.
\end{aligned}\tag{30}$$

If each $|h_{jv}|\leq\min_{j\in I_{1,v}}\left\{\frac{\sqrt{2\alpha}}{2},|u_{jv}|\right\}$

$$\begin{aligned}
&\mathcal{F}(U^\star+H)-\mathcal{F}(U^\star)\\
&\qquad\geq\sum_{v\in\mathcal{V}}\left[q\left\|A_vh_v\right\|_2^2+\frac{1-q}{d\tau}\sum_{w\in\mathcal{N}_v}\left\|h_v-\overline{h}_w\right\|_2^2\right].
\end{aligned}$$

Since $A^{\mathsf{T}}A=\sum_{v\in\mathcal{V}}A_v^{\mathsf{T}}A_v$ is positive definite then $\sum_{v\in\mathcal{V}}\left[q\left\|A_vh_v\right\|_2^2+\frac{1-q}{d\tau}\sum_{w\in\mathcal{N}_v}\left\|h_v-\overline{h}_w\right\|_2^2\right]=0$ if and only if $H=0$. In fact, we should otherwise have $h_v=h$ for each $v\in\mathcal{V}$ in order to obtain $\sum_{v\in\mathcal{V}}\left[\frac{1-q}{d\tau}\sum_{w\in\mathcal{N}_v}\left\|h_v-\overline{h}_w\right\|_2^2\right]=0$, but $\sum_{v\in\mathcal{V}}\left\|A_vh\right\|_2^2=0$ has only the zero solution. We conclude that if H is sufficiently small then $\mathcal{F}(U^\star+H)-\mathcal{F}(U^\star)>0$. □

Chiara Ravazzi received the B.Sc. and M.Sc., and Ph.D. degrees in applied mathematics from Politecnico di Torino, Italy, in 2005 and 2007, and 2011 respectively. She is currently a Postdoctoral Associate at the Department of Electronics and Telecommunications (DET), Politecnico di Torino, Italy. During 2010, she was a visiting student at the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge. Specific themes of current interest include the mathematics of control and information theory, coding theory, and signal processing.

Sophie M. Fosson received the B.Sc. and M.Sc. degrees in applied mathematics from Politecnico di Torino, Italy, in 2002 and 2005, respectively. She received the Ph.D. degree in mathematics for the industrial technologies from Scuola Normale Superiore di Pisa, Italy, in 2011. She is currently a Postdoctoral Associate at the Department of Electronics and Telecommunications (DET), Politecnico di Torino, Italy. Her research interests include control and information theory, network science, and digital signal processing.

Enrico Magli received the Ph.D. degree in Electrical Engineering in 2001, from Politecnico di Torino, Italy. He is currently an Associate Professor at the same university, where he leads the Image Processing Lab. His research interests are in the field of compression of satellite images, multimedia signal processing and networking, compressive sensing, distributed source coding, image and video security. He is an associate editor of the IEEE Transactions on Circuits and Systems for Video Technology, of the IEEE Transactions on Multimedia, and of the EURASIP Journal on Image and Video Processing. He is general co-chair of IEEE ICME 2015 and IEEE MMSP 2013, and has been TPC co-chair of ICME 2012, VCIP 2012, VCIP 2014, MMSP 2011 and IMAP 2007. He has published about 50 papers in refereed international journals, 4 book chapters, and over 130 conference papers. He is a co-recipient of the IEEE Geoscience and Remote Sensing Society 2011 Transactions Prize Paper Award. He has received the 2010 Best Reviewer Award of IEEE Journal of Selected Topics in Applied Earth Observation and Remote Sensing, and the Best Associate Editor Award of IEEE Transactions on Circuits and Systems for Video Technology in 2012 and 2014.