Randomized Algorithms for Distributed Nonlinear Optimization Under Sparsity Constraints

Randomized algorithms for distributed nonlinear optimization under sparsity constraints

Chiara Ravazzi, *Member, IEEE*, Sophie M. Fosson, *Member, IEEE*, and Enrico Magli, *Senior Member, IEEE*

***Abstract*—Distributed optimization in multi-agent systems under sparsity constraints has recently received a lot of attention. In this paper, we consider the in-network minimization of a continuously differentiable nonlinear function which is a combination of local agent objective functions subject to sparsity constraints on the variables.**

A crucial issue of in-network optimization is the handling of the communications, which may be expensive. This calls for efficient algorithms, that are able to reduce the number of required communication links and transmitted messages. To this end, we focus on asynchronous and randomized distributed techniques. Based on consensus techniques and iterative hard thresholding methods, we propose three methods that attempt to minimize the given function, promoting sparsity of the solution: asynchronous hard thresholding (AHT), broadcast hard thresholding (BHT), and gossip hard thresholding (GHT). Although similar in many aspects, it is difficult to obtain a unified analysis for the proposed algorithms. Specifically, we theoretically prove the convergence and characterize the limit points of AHT in regular networks under some proper assumptions on the functions to be minimized. For BHT and GHT, instead, we characterize the fixed points of the maps that rule their dynamics in terms of stationary points of original problem. Finally, we illustrate the implementation of our techniques in compressed sensing and present several numerical results on performance and number of transmissions required for convergence.

***Index Terms*—Distributed optimization, nonlinear optimization, sparse signal recovery, randomized algorithms.**

I. Introduction

Distributed optimization has been receiving increasing attention in the last years [1], [2], [3], [4], [5] due to its applications in diverse multi-agent frameworks, ranging from detection and estimation over sensor networks [6], [7] to compressed sensing [8] and medical imaging [9]. Modern networked technologies have demonstrated that distributed systems of interconnected and low-power units can efficiently replace a single, powerful centralized processor for tasks like, *e.g.*, monitoring, tracking, localization, and imaging [9], [10], [11], [12], [13]. In some cases, networked systems are used only to acquire data, and optimization is performed by a single data fusion center. However, more interesting and challenging is the problem of distributed optimization, whose goal is to compute the solution in-network, leveraging local communication and cooperation among agents.

The authors are with the Politecnico di Torino – Department of Electronics and Telecommunications, Italy. e-mail: {name.surname}@polito.it. This work has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement no. 279848.

As explained in [1], [2], [3], [4], the natural mathematical formulation of distributed optimization consists in the minimization of a cost functional that is sum of different terms, each of which associated with an agent. In this paper, we undertake such model imposing a sparsity constraint, that is, we assume that the desired solution is a vector with few non-zero components. In many practical cases, in fact, models are constrained structurally so that only few degrees of freedom compared to their ambient dimension are significant. In the last decades, optimization problems under sparsity constraints have attracted much interest, especially in statistics, signal processing, machine learning, and coding theory. The reader can refer to [14] and references therein for an overview of possible applications.

In the sparse distributed optimization context, considerable effort has focused on sparse linear regression models [15], which lead to classical square residual cost functions. One of the most studied examples is compressed sensing, where optimization starts from linear, compressed measurements (that is, measurements of kind $y = Ax$, with $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$, $A \in \mathbb{R}^{m\times n}$, $m \ll n$) acquired by a networked system; the case of totally decentralized optimization is particularly timely [8], [15], [16], [17], [18], [19]. Less attention instead has been devoted to nonlinear sparse models, even if they are able to better describe a variety of real applications, where measurements might be quantized [20], quadratic ($y_i = (Ax)_i^2$, $i = 1, \dots, m$, [21]), exponential ($y_i = e^{(Ax)_i}$, $i = 1, \dots, m$) or realizations of random variables [22]. Nevertheless, it is well known that nonlinearity is more difficult to analyze theoretically, which explains the lack of literature in particular for what concerns sparse models. New interest in this direction has however raised very recently: in [21], [23], [24] nonlinear sparse optimization has been addressed, while in [8], [17] the networked case is considered.

The goal of this paper is to tackle distributed sparse nonlinear optimization (*e.g.*, as in [8]), devoting particular attention to energy efficiency, in terms of transmissions and memory requirements. In distributed optimization, indeed, a single agent typically has limited memory and processing capability, therefore a cooperation is the key to compensate for this shortcoming and achieve satisfactory performance, as proved in many works, *e.g.*, [8], [18], [19]. Cooperation, however, raises the problem of communication among agents, which can be expensive from different viewpoints. For example, if we consider a territorial monitoring wireless sensor network, the agents (that is, the sensors) may be deployed at large distance or in unfavorable conditions, which makes communication uneconomical.

The drawbacks due to network communications have motivated us to investigate algorithms that reduce the number of necessary information exchanges. In particular, in this paper we consider randomized techniques, in which data transmissions are ruled by suitable probabilistic models [25]. This not only allows us to limit the communication load, but also to overcome synchronization issues. Specifically, structured communication schemes typically require the agents to coordinate in order to activate and transmit at the right moment. This is no more necessary in the randomized setting, in which agents are randomly activated or are activated by other agents.

The distributed algorithms that we propose are based on iterative hard thresholding (IHT, [26], [27]), which combines a gradient descent step and a thresholding step. We show that IHT can be generalized to a distributed and randomized setting, and also that it requires little computational effort at each iteration, providing an extremely appealing solution to distributed sparse nonlinear optimization. We analyze three different protocols, through both mathematical analysis and numerical simulations: asynchronous hard thresholding (AHT), broadcast hard thresholding (BHT), and gossip hard thresholding (GHT). In particular, we theoretically prove the convergence and characterize the limit points of AHT in regular networks under some proper assumptions on the functions to be minimized. Our numerical simulations suggest that these hypotheses on the network regularity, that are useful to prove the convergence of AHT, are not really necessary. In fact, we have tested also non-regular topologies, and convergence has been always achieved (see Section VI). For BHT and GHT, instead, we characterize the fixed points of the maps that rule their dynamics in terms of stationary points of original problem.

In [8], distributed methods based on IHT were developed as well, which however are different from ours in the network communication management. A comprehensive comparison is proposed in Section IV-F.

The example of compressed sensing is considered throughout the paper as application benchmark on which we test the proposed theory. The paper is organized as follows. In Section II we discuss the model and the optimization problem associated with sparse constraints in a distributed setting. In Section III we review the IHT method in a non distributed setting, discussing the related literature and known theoretical results. This is used in Section IV to derive the proposed algorithms. Section V is then devoted to our theoretical results. The most technical parts are postponed to the Appendix. Further discussion on the algorithms' performance is provided in Section VI through a consistent variety of numerical simulations. Finally, we draw our conclusions (Section VII).

We conclude this introduction with some notation used in the sequel. We denote column vectors with small letters, and matrices with capital letters. If $x \in \mathbb{R}^n$ we denote its j-th element with x^j and, given $S \in [n] := \{1, \dots, n\}$, by $x|_S$ the subvector of x corresponding to the indices in S. The support set of x is defined by $\text{supp}(x) = \{i \in [n] : x^i \neq 0\}$ and we use $\|x\|_0 = |\text{supp}(x)|$. We denote with $r(x)$ the non increasing rearrangement of x, *i.e.*,

$$r(x) = (|x^{i_1}|, |x^{i_2}|, \dots, |x^{i_n}|)^T \tag{1}$$

where $|x^{i_\ell}| \geq |x^{i_{\ell+1}}|,\ \forall \ell = 1, \dots, n-1$ and $\{i_1, \dots, i_n\} = [n]$ and we define $r^k(x) = |x^{i_k}|$. Given a matrix X, X^T denotes its transpose and $X_{(v)}$ (or x_v) denotes the v-th column of X. For any matrix $M \in \mathbb{R}^{n \times m}$, the Frobenius norm is defined as $\|M\|_F := \sqrt{\sum_{i=1}^n \sum_{j=1}^m |M_{i,j}|^2}$. For a square matrix $M \in \mathbb{R}^{n \times n}$, we consider the induced norm $\|M\|_2 = \sup_{z \neq 0} \|Mz\|_2 / \|z\|_2$. Finally, the symbol $\|\cdot\|$ with no subscript has always to be intended as $\|\cdot\|_2$.

II. Sparsity constrained optimization in multi-agent systems

As in [8], we consider a network of N agents (that we label as $v \in \mathcal{V} = \{1, \dots, N\}$) which can represent sensors or processing units that collect measurements of a physical variable $x^\star \in \mathbb{R}^n$. The agents seek to estimate the unknown vector $x^\star$ that is k-sparse (*i.e.*, it has at most k nonzero entries), starting from their own sets of measurements $y_v = y_v(x^\star) \in \mathbb{R}^m$, $v \in \mathcal{V}$. A loss function $f(x; y_v) : \mathbb{R}^n \to \mathbb{R}$, denoted with $f_v(x)$ for brevity, is defined for each $v \in \mathcal{V}$, which indicates how much a vector $x \in \mathbb{R}^n$ is consistent with the measurements y_v. In addition to their own y_v's, the agents can leverage local communication in the network to estimate $x^\star$. The corresponding optimization problem can be written as follows:

$$\min f(x) \qquad \text{s.t.} \quad x \in \Sigma_k, \tag{2}$$

with

$$f(x) := \sum_{v \in \mathcal{V}} f_v(x) \quad \text{and} \quad \Sigma_k := \{x \in \mathbb{R}^n : \|x\|_0 \leq k\}.$$

The following assumptions are made throughout the paper.

Assumption 1. The problem (2) admits a unique solution x_{opt}.

Assumption 2. For all $v \in \mathcal{V}$,

a) f_v is lower bounded, *i.e.*, there exists $\kappa_v \in \mathbb{R}$ such that $f_v(x) \geq \kappa_v$ for all $x \in \mathbb{R}^n$;
b) f_v is twice continuously differentiable in $\mathbb{R}^n$;
c) the gradient $\nabla f_v(x)$ is Lipschitz continuous over $\mathbb{R}^n$, that is, there exists $L_v \in \mathbb{R}$ such that

$$\|\nabla f_v(x) - \nabla f_v(z)\|_2 \leq L_v \|x - z\|_2 \qquad \forall x, z \in \mathbb{R}^n.$$

Furthermore, let us consider the following definition [28].

Definition 1. Let g be a twice continuously differentiable function whose Hessian is denoted by $\nabla^2 g(\cdot)$ and let

$$\alpha_k = \inf\{x^T \nabla^2 g(\xi) x : |\text{supp}(x) \cup \text{supp}(\xi)| \leq k, \|x\| = 1\}.$$

We say that g satisfies the left stable restricted Hessian property with constant α_k (α_k-LSRHP for short).

Assumption 3. For all $v \in \mathcal{V}$, f_v satisfies the α_k-LSRHP with $\alpha_k^v \geq 0$ and there exists $\overline{v} \in \mathcal{V}$ such that $\alpha_k^{\overline{v}} > 0$.

A key property of the considered model is that the functions $f : \mathbb{R}^n \to \mathbb{R}$ that satisfy the Assumption 3 are convex over canonical sparse subspaces, but they are not necessarily convex everywhere, as often assumed in literature of distributed

optimization [3], [4], [29], [30]. Some examples that describe non-convex functions that satisfy the α_k-LSRHP can be found in [28, Example 1 and Example 2]. This makes the problem much more general and applicable to diverse scenarios as in compressed sensing [31] and transmission tomography [32]. Moreover, the fact that Σ_k is not a convex set, makes the problem very challenging.

Before presenting an example satisfying Assumptions 1, 2, and 3, we give a preliminary definition introduced in [21] and we recall a necessary condition for optimality.

Definition 2. $z \in \mathbb{R}^n$ is called a basic feasible (BF) vector of (2) when one of the following conditions holds:

(a) if $\|z\|_0 < k$ then $\nabla f(z) = 0$;
(b) if $\|z\|_0 = k$ then $\nabla_i f(z) = 0$ for all $i \in \mathrm{supp}(z)$.

Theorem 2.1 in [21] establishes the fact that any optimal solution of (2) is also a BF vector of (2). We show in the following theorem that, under a suitable assumption, BF vectors are local minima of (2).

Theorem 1. *Suppose that Assumptions 2 and 3 hold. Any BF vector of* (2) *is a strict local minimum for* (2)*.*

Proof. Let z be a BF vector of (2) and $\epsilon \in (0, \min_{i\in\mathrm{supp}(z)} |z_i|)$. We now show that for any $z + h \in \Sigma_k$ and $\|h\|_2 < \epsilon$ we have $f(z+h) - f(z) > 0$. From Assumption 2 we have that

$$f(z+h) - f(z) = \langle \nabla f(z), h\rangle + \frac{1}{2} h^T \nabla^2 f(\xi) h$$

where $\xi = z + \gamma h$ with $\gamma \in (0, 1)$. From Definition 2, if $\|z\|_0 < k$ then $\nabla f(z) = 0$ otherwise, if $\|z\|_0 = k$ then the constraint $z + h \in \Sigma_k$ with $\|h\|_2 < \epsilon$ implies that $\mathrm{supp}(h) \subseteq \mathrm{supp}(z)$. Consequently, $\langle \nabla f(z), h\rangle = 0$, and

$$f(z+h) - f(z) = \sum_{v\in\mathcal{V}} \frac{1}{2} h^T \nabla^2 f_v(\xi) h \geq 0.$$

If there exists $\overline{v} \in \mathcal{V}$ such that $f_{\overline{v}}$ satisfies the α_k-LSRHP with $\alpha_k > 0$ (see Assumption 3) then the last inequality is strict and the assertion is proved. □

Example: Distributed compressed sensing We consider the distributed reconstruction problem in compressed sensing (see [8], [15], [16], [17], [18], [19], [33]). We assume that each agent $v \in \mathcal{V}$ of a network senses a common, k-sparse signal $x^\star \in \mathbb{R}^n$ and acquires $m \ll n$ linear measurements of the form

$$y_v = A_v x^\star + \xi_v, \tag{3}$$

where $y_v \in \mathbb{R}^m$, $A_v \in \mathbb{R}^{m\times n}$, and $\xi_v \in \mathbb{R}^m$ is a Gaussian noise $\mathsf{N}(0, \sigma^2)$. The agents seek to estimate $x^\star$ given the measurements and knowing the sparsity level k. It is thus natural to consider the following optimization problem in order to approximate $x^\star$:

$$\min_{x\in\mathbb{R}^n} \sum_{v\in\mathcal{V}} \frac{1}{2}\|y_v - A_v x\|_2^2 \qquad \text{s.t. } x \in \Sigma_k. \tag{4}$$

In absence of noise (*i.e.*, $\xi_v = 0, \forall v \in \mathcal{V}$), if for every index set $\Gamma \subseteq \{1, \ldots, n\}$ with $|\Gamma| = 2k$ the columns of $A = (A_1^T, \ldots, A_N^T)^T$ associated with Γ are linearly independent, then $x^\star$ is the unique solution to (4) [31].

It is straightforward to verify that $f_v(x) = \frac{1}{2}\|y_v - A_v x\|^2$ satisfies Assumption 2.a)-b). The Hessian matrix has the form $\nabla^2 f_v(x) = A_v^T A_v$, and Assumption 2.c) is guaranteed with constant $L_v = \|A_v\|_2^2$. It is easy to check that if A_v has rank not smaller than k, then f_v satisfies also the α_k-LSRHP with $\alpha_k > 0$.

III. Iterative hard thresholding

Problems of the form (2) have been largely studied in the centralized setting, that is, when all the data are processed by a single fusion center. In this section we review a few elements of the centralized reconstruction techniques, which will be the basis to develop our distributed methods.

An efficient method to tackle the problem (2) in a centralized way is the IHT [23], [21]. A comprehensive overview of IHT is given in [21], where different procedures to solve (2) have been proposed. More precisely, at each iteration step, the classical IHT procedure performs a gradient step and then a best k-term approximation. We denote the best k-term approximation of x with

$$\sigma_k(x) := \operatorname*{argmin}_{z\in\Sigma_k} \|x - z\|_2$$

that sets to zero the $n - k$ components of x with smallest magnitude.

Algorithm 1 IHT

1: Initialization: $x(0) = 0 \in \mathbb{R}^n$, $\tau > 0$
2: **for** $t = 0, 1, \ldots, StopIter$ **do**
3: $\quad x(t+1) = \sigma_k[x(t) - \tau \nabla f(x(t))]$
4: **end for**

Definition 3. [21, Definition 2.3] A vector z is called a τ-stationary point of (2) if it satisfies the following relation

$$z = \sigma_k\left(z - \tau \nabla f(z)\right).$$

Theorem 2.2 in [21] shows that, under Assumption 2.c), τ-stationarity is a necessary condition for optimality for any $\tau \in (0, (\sum_{v\in\mathcal{V}} L_v)^{-1})$. Moreover, the following result holds for IHT.

Theorem 2. *[21, Theorem 3.1] Let $\{x(t)\}_{t\in\mathbb{N}}$ be the sequence generated by IHT with stepsize $\tau \in (0, (\sum_{v\in\mathcal{V}} L_v)^{-1})$, where L_v is the Lipschitz constant defined in the Assumption 2.c). Then any accumulation point $\{x(t)\}_{t\in\mathbb{N}}$ is a τ - stationary point.*

A direct result of Lemma 2.2 in [21] is that any τ-stationary point is a BF vector of (2). Combining this with Theorems 1 and 2 we obtain that any accumulation point of IHT is a local minimum of (2).

Results on the convergence of IHT have been proposed in [27] for compressed sensing, and in in [23] for compressed sensing with nonlinear observations.

Although not considered in this work, we remark that other approaches based on the relaxation of the problem to ℓ_0/ℓ_1 regularized functionals have been widely studied in literature. For these approaches there is also a variety of proposed algorithms, including quadratic programming methods [34],

interior-point methods [34], projected gradient methods [35], and iterative (hard and soft) thresholding algorithms [36], [26], [27].

IV. Distributed hard thresholding

Our goal is to develop distributed techniques to solve (2) in an energy efficient way, that is, reducing the number of communications as much as possible. The presence of the common variable x in (2) imposes a coupling between the agents, and distributed algorithms require collaboration. This entails repeated transmissions, which clearly have a cost in terms of energy to transmit and use of the communication links. The study of algorithms that limit the number of sent messages is then encouraged.

Motivated by this observation, we now introduce our family of distributed, randomized algorithms, built on IHT (Section III). From now on, we consider a connected network and we model it by an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of agents and $\mathcal{E} \in \mathcal{V} \times \mathcal{V}$ represents the set of the available communication links. The set of neighbors of $v \in \mathcal{V}$ is denoted as $\mathcal{N}_v = \{w \in \mathcal{V} : \ (v, w) \in \mathcal{E}\}$. We assume that $(v, v) \in \mathcal{E}$ for all $v \in \mathcal{V}$.

Given this network structure, our algorithms are all based on the idea that each agent performs an IHT reconstruction procedure, but adjusts its own estimate based on knowledge of the estimates of its neighbors. The communication protocols can be of different kind, that give rise to different algorithms. Here, we study three cases, that correspond to the above mentioned AHT, BHT, and GHT. These methods are iterative, hence stopping criteria should be defined. Generally, if the algorithm is analytically proved to converge, we can stop it when numerical convergence is achieved, that is, when the distance between the estimates of two successive iterates is below a fixed threshold; if the algorithm is not guaranteed to converge, we can fix a maximum number of iterations. These criteria will be discussed for AHT, BHT, and GHT in Section VI, after having studied their convergence properties.

We now describe each algorithm in detail.

A. Asynchronous hard thresholding (AHT)

In the AHT algorithm (see Algorithm 2), at each iteration step, one agent (that is, a vertex in the graph) is selected and communicates with its neighbors to receive their estimates. After communication, the selected agent performs the following operations: (a) gradient of its loss function f_v; (b) average of the received neighbors' estimates (included itself; in Algorithm 2, $d_v = |\mathcal{N}_v|$); (c) combined gradient step using (a) and (b); (d) best k-term approximation of (c). The procedure is iterated until a stopping criterion is met. The selection at step 3 is discussed in Section IV-D.

B. Broadcast hard thresholding (BHT)

In the BHT procedure (Algorithm 3), the communication protocol is reversed with respect to AHT.

At each iteration step, one agent $v \in \mathcal{V}$ is selected and sends its estimation to the neighbors $w \in \mathcal{N}_v$. After communication, each agent $w \in \mathcal{N}_v$ updates its status performing the following operations: (a) computation of gradient $\nabla f_w(x_w)$; (b) average between its estimate x_w and the received estimate x_v; (c) combined gradient step using (a) and (b); (d) best k-term approximation of (c). The procedure is iterated until a stopping criterion is met.The selection at step 3 is discussed in Section IV-D.

Algorithm 2 AHT

1: Initialization: $x_v(0) = 0 \in \mathbb{R}^n$, $\tau > 0$
2: **for** $t = 0, 1, \ldots, StopIter$ **do**
3: Selection of $v \in \mathcal{V}$
4: $x_v(t+1) = \sigma_k \left[\frac{1}{d_v} \sum_{w \in \mathcal{N}_v} x_w(t) - \tau \nabla f_v(x_v(t)) \right]$
5: $x_h(t+1) = x_h(t)$ for any $h \neq v$
6: **end for**

Algorithm 3 BHT

1: Initialization: $x_v(0) = 0 \in \mathbb{R}^n$, $\tau > 0$
2: **for** $t = 0, 1, \ldots, StopIter$ **do**
3: Selection of a $v \in \mathcal{V}$
4: $x_w(t+1) = \sigma_k \left[\frac{1}{2}(x_v(t) + x_w(t)) - \tau \nabla f_w(x_w(t)) \right]$ for all $w \in \mathcal{N}_v$
5: $x_h(t+1) = x_h(t)$ for all $h \notin \mathcal{N}_v$
6: **end for**

C. Gossip hard thresholding (GHT)

In GHT, only one communication link is used at each iteration step. One agent is selected and woken up and, in turn, chooses one neighbor, receives its estimate, and performs the update as in Algorithm 4. The procedure is iterated until a stopping criterion is reached. The selection at step 3 is discussed in Section IV-D.

Algorithm 4 GHT

1: Initialization: $x_v(0) = 0 \in \mathbb{R}^n$ for all $v \in \mathcal{V}$, $\tau > 0$
2: **for** $t = 0, 1, \ldots, StopIter$ **do**
3: Selection of $(v, w) \in \mathcal{E}$
4: $x_v(t+1) = \sigma_k \left[\frac{x_v(t) + x_w(t)}{2} - \tau \nabla f_v(x_v(t)) \right]$
5: $x_h(t+1) = x_h(t)$ for all $h \neq v$
6: **end for**

D. Selection models

For the selection of the node in AHT and BHT and of the edge in GHT (see step 3 in Algorithms 2, 3, and 4), we consider two scenarios, which are formally described in the following definitions.

Definition 4. A network of N nodes is said to be uniformly persistent if there exists a positive integer number $T > 0$ such that, for all $t \in \mathbb{N}$, each node makes the update at least once within the iteration-interval $[t, t + T)$.

Definition 5. A network of N nodes is said to be randomly persistent if there exists a N-upla $(p_1, \ldots p_N)$ such that $p_v >$

0, for all $v \in \mathcal{V}$, $\sum_{v\in\mathcal{V}} p_v = 1$, and such that, for all $t \in \mathbb{N}$, $\mathbb{P}[\Omega_{v,t}] = p_v$, where $\Omega_{v,t}$ is the event

$$\Omega_{v,t} = \{\text{node } v \text{ makes the update at iteration } t\}.$$

Basically, the node/edge selection can be done in any way that guarantees that the network is uniformly persistent or randomly persistent (see Theorem 6).

These definitions can be extended also in the case an edge is selected instead of a node.

In Figure 1, an example of iterative step is shown for a 4-agent network, illustrating the difference between AHT, BHT, and GHT. Agent 3 is initially selected for all schemes. The active agents, that is, the ones that update their estimates, are boldfaced, as well as the active links; the communication direction is indicated by the arrows.

E. Computational requirements

We conclude the presentation of the algorithms with some remarks on their computational requirements. For simplicity we assume that all the nodes $v \in \mathcal{V}$ have degree $d_v = d$.

The requirements in terms of memory usage are comparable for the all three algorithms: each agent v has to store the information encoded by f_v and $2k$ real values of the solution estimation (the nonzero components and their positions) at each agent, while the temporary information for computation is of order $O(n)$.

For GHT only one communication link is used at each iteration step, while for AHT and BHT the number depends on the degree d. It is however clear that the total number of usages of communication links depends on the total number of iterations.

The use of best k-term approximation is an advantage for what concerns the transmission of the current estimate at each iteration step, as the number of real values that have to be sent is reduced from n to $2k$ for each communication link [8], [15], [16], [17], [18], [19].

We finally notice that the number of computations per agent is actually the same for each activated agent for all three algorithms. However, the number of updating agents (that is, agents that perform computations to update their own estimations) is d for BHT, and just 1 for both AHT and GHT. The number of agents sending messages is d on average for AHT, and 1 for BHT and GHT.

In Table I, AHT, BHT and GHT are compared in terms of computational effort and communication requirements.

Table I
COMPUTATIONAL REQUIREMENTS AT EACH ITERATION STEP

	AHT	BHT	GHT
Active comm. links	$d-1$	$d-1$	1
Sent values	$2k(d-1)$	$2k(d-1)$	$2k$
Updating agents	1	d	1
Sender agents	$d-1$	1	1

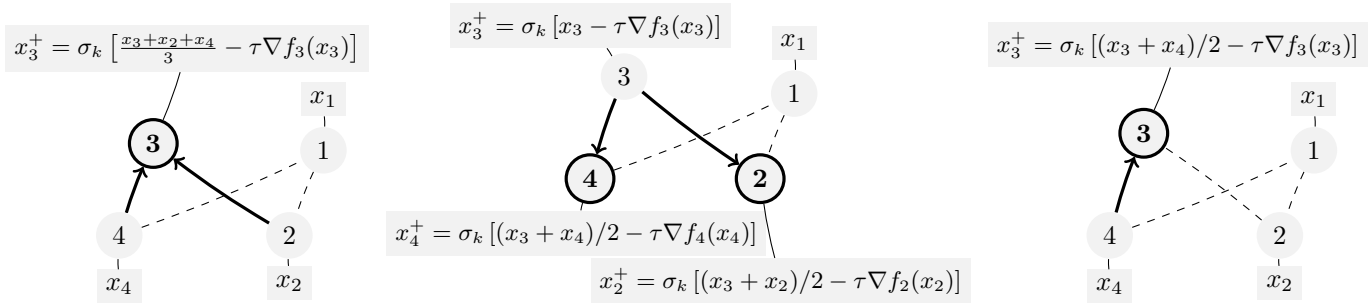
F. Relation to prior literature

Distributed approaches to problems of kind (2) have drawn much attention very recently. Our main reference is [8], in which the authors address (2) over static and time-varying networks proposing two protocols based on IHT.

The first method is distributed iterative hard thresholding (DIHT), that can be outlined as follows. Given a network, one agent r is chosen and a spanning tree is built fixing r as root; then, iterative communication is activated from the root towards the leaves and vice versa so that (a) r broadcasts its estimate of the signal, (b) the other agents receive this estimate and transmit back information to r to update it. More precisely, given the current estimate $x(t)$, each agent $v \neq r$ receives it and computes its own gradient $\nabla f_v(x(t))$; then, starting from the leaves the gradients are sent back and accumulated so that r receives their sum and use it (along with is own $\nabla f_r(x(t))$) to update $x(t)$ to $x(t+1)$ through hard thresholding. In [8], this method has been shown to work efficiently and overcome other distributed methods in terms of convergence times and number of required transmitted values. Its main limitation is in the imposed hierarchy, specifically, an exclusion of agent r (due, for example, to a failure) would seriously disrupt the process, as it is the only agent storing all the necessary information to pursue the recovery.

The necessity of a spanning tree is removed in the second method proposed in [8], known as CB-DIHT and based on diffusive consensus. In CB-DIHT, instead of summing the gradients from the leaves to the root of a spanning tree, local means of the gradients are computed by each agent as in consensus procedures. In order to do this, all the agents should receive from a prescribed agent r the current estimate $x(t)$ through a diffusive procedure. This idea can be used even when the topology of the network is time-varying, provided that some connectivity conditions are respected. Interestingly, CB-DIHT is identical to IHT except that at each iteration the gradient is approximated [8, Proposition 4.4]. Concerning convergence times and number of transmitted values, numerical results show that CB-DIHT is barely worse than DIHT, see [8, Tables II-III].

The current literature offers various other algorithms that address problems similar to (2). In particular, in [17] different constraints are considered: the variables should belong to closed and convex sets. For example, problems in which the sparse constraint is relaxed with the ℓ_1-norm is considered in [16], [17]. The algorithm proposed in these works, known as D-ADMM, is an extension of ADMM to the distributed context, using graph coloring techniques. In terms of required transmitted values, in [8] D-ADMM is proved to be less efficient than DIHT and CB-DIHT, which is mainly due to the local transmission of non sparse information. In DIHT and CB-DIHT, instead, the estimate $x(t)$ diffused by the agent r is k-sparse, which requires less bandwidth usage. Nevertheless, DIHT and CB-DIHT also require to share non sparse information when the information about gradients is conveyed towards r. This aspect is improved by AHT, BHT, and GHT, in which only k-sparse vectors (namely, the signal's estimates produced by each single agent) are shared.



(a) AHT (b) BHT (c) GHT

Figure 1. Example of a network with 4 agents: behaviors of AHT, BHT, and GHT when agent 3 is initially selected. In AHT, agent 3 gets information from its neighbor and updates its estimate; in BHT, agent 3 broadcasts its estimates, and its neighbors 2 and 4 update their own estimate; in GHT, agent 4 is in turn randomly selected among the neighbors of agent 3, which gets the estimate of agent 4 and updates itself.

We finally mention that relaxed versions of distributed problems with constraints can be tackled by methods like distributed subgradient algorithms (DSM, [2], [3], [29]) and deterministic distributed (soft and hard) iterative thresholding (DISTA and DIHTA, [18], [19]).

V. Theoretical results

In this section, we analyze the behaviors of AHT, BHT, and GHT. Specifically, we prove the convergence of AHT and we characterize the fixed points of the maps that rule the dynamics of BHT and GHT.

A. Convergence of AHT

We analyze the convergence of AHT under the following assumption.

Assumption 4. The graph of communication is

1) connected;
2) regular, that is, all nodes $v \in \mathcal{V}$ have degree $d_v = d$.

In order to prove the convergence, we first recast the optimization problem in (2) into a separable form that facilitates distributed implementation. The goal is to split the problem into simpler subtasks executed locally at each node. Let us replace the global variable x in (2) with local variables $\{x_v\}_{v\in\mathcal{V}}$. We rewrite the distributed problem as follows

$$\begin{aligned} &\min_{x_1,\dots,x_N\in\mathbb{R}^n} \sum_{v\in\mathcal{V}} f_v(x_v), \\ &\text{s.t. } x_v \in \Sigma_k \text{ and } x_v = x_w, \quad \forall w \in \mathcal{N}_v, \forall v \in \mathcal{V}. \end{aligned} \tag{5}$$

We now relax the problem (5) and consider the minimization of the functional $F : \mathbb{R}^{n\times N} \longmapsto \mathbb{R}^+$ defined as follows

$$\begin{aligned} &F(X) := \sum_{v\in\mathcal{V}} \left[f_v(x_v) + \frac{1}{4\tau d} \sum_{w\in\mathcal{N}_v} \|x_v - x_w\|^2 \right] \\ &\text{s.t. } x_v \in \Sigma_k, \quad \forall v \in \mathcal{V} \end{aligned} \tag{6}$$

where $X = (x_1, \dots, x_N)$. By minimizing F, each node seeks to estimate the sparse solution to (2) and to enforce agreement with the estimates calculated by other nodes in the network. It should also be noted that $F(x\mathbb{1}^T) = f(x)$.

Inspired by the terminology introduced in [21], we derive necessary optimality conditions of (6) and discuss the relationships with the original problem (2).

Definition 6. $Z = (z_1, \dots, z_N) \in \Sigma_k^N$ is called a basic feasible (BF) point of (6) if it satisfies the following conditions $\forall v \in \mathcal{V}$:

$$\begin{aligned} z_v &= \overline{z}_v - \tau\nabla f_v(z_v) && \text{if } \|z_v\|_0 < k \\ z_v^j &= \overline{z}_v^j - \tau\nabla_j f_v(z_v), \forall j \in \operatorname{supp}(z_v) && \text{if } \|z_v\|_0 = k \end{aligned}$$

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

Proposition 1. *If Z is an optimal solution of* (6)*, then Z is a BF point of* (6).

Proof. If Z is such that $\|z_v\|_0 < k$, then for any $j \in \{1, \dots, n\}$ we have

$$0 \in \operatorname{argmin}\{G(t) = F(Z + te_j e_v^T)\}.$$

By imposing $G'(0) = 0$ and by using the Assumption 4, we obtain

$$\begin{aligned} 0 =& \nabla_j f_v(z_v) + \frac{\sum_{w\in\mathcal{N}_v}(x_v^j - x_w^j)}{2\tau d} + \frac{\sum_{w\in\mathcal{V}:v\in\mathcal{N}_w}(x_v^j - x_w^j)}{2\tau d} \\ =& \nabla_j f_v(z_v) + \frac{1}{\tau}x_v^j - \frac{1}{\tau d}\sum_{w\in\mathcal{N}_v} x_w^j. \end{aligned}$$

If $\|z_v\|_0 = k$, the same condition holds for $j \in \operatorname{supp}(z_v)$. The proof is completed iterating the argument for all $v \in \mathcal{V}$. □

Proposition 1 gives a necessary condition for optimality. However this condition is weak and there are in principle many BF points of(6) that are not optimal. Let us introduce the following definition.

Definition 7. $Z \in \Sigma_k^N$ is called a τ-stationary point of (6) if it satisfies the following condition $\forall v \in \mathcal{V}$:

$$z_v = \sigma_k\left(\overline{z}_v - \tau\nabla f_v(z_v)\right).$$

The following proposition gives another representation of τ-stationary points. The proof is omitted for brevity but can be easily checked from Definition 7.

Proposition 2. *Z is a τ-stationary point of* (6) *if and only if $\forall v \in \mathcal{V}$ the vector $z_v \in \Sigma_k$ and*

$$\begin{cases} \overline{z}_v^j - \tau\nabla_j f_v(z_v) = z_v^j & \text{if } j \in \operatorname{supp}(z_v) \\ |\overline{z}_v^j - \tau\nabla_j f_v(z_v)| \le r^k(z_v) & \text{if } j \notin \operatorname{supp}(z_v). \end{cases} \tag{7}$$

Corollary 1. *If Z is τ-stationary point of* (6)*, then it is a BF point of* (6).

We now show that under an appropriate Lipschitz condition, τ-stationarity is a necessary condition for optimality. Before presenting the optimality conditions we give a preliminary result.

Lemma 1. *Let $X, Y \in \Sigma_k^N$ such that $y_v = \sigma_k(\overline{x}_v - \tau\nabla f_v(x_v))$, then the following relation is true for all $v \in \mathcal{V}$:*

$$\langle \nabla f_v(x_v), y_v - x_v\rangle \leq -\frac{1}{2\tau}\|y_v - x_v\|^2 - \frac{1}{\tau d}\sum_{w\in\mathcal{N}_v}\langle x_v - x_w, y_v - x_v\rangle.$$

Proof. It should be noticed that

$$y_v = \operatorname*{argmin}_{s\in\Sigma_k}\left[\frac{1}{2}\|s - \big(x_v - \tau\nabla f_v(x_v)\big)\|^2 + \frac{1}{d}\sum_{w\in\mathcal{N}_v}\langle x_v - x_w, s\rangle\right].$$

This yields to the following inequality:

$$\frac{1}{2}\|y_v - (x_v - \tau\nabla f_v(x_v))\|^2 + \frac{1}{d}\sum_{w\in\mathcal{N}_v}\langle x_v - x_w, y_v\rangle \leq \frac{1}{2}\|\tau\nabla f_v(x_v))\|^2 + \frac{1}{d}\sum_{w\in\mathcal{N}_v}\langle x_v - x_w, x_v\rangle$$

and, consequently,

$$\frac{1}{2}\|y_v - x_v\|^2 + \tau\langle\nabla f_v(x_v), y_v - x_v\rangle \leq -\frac{1}{d}\sum_{w\in\mathcal{N}_v}\langle x_v - x_w, y_v - x_v\rangle.$$

□

Theorem 3. *Suppose that Assumptions 2 and Assumption 4 hold and let $Z^\star$ be an optimal solution of* (6)*. Then $Z^\star$ is a τ-stationary point for any $\tau < \frac{1}{dL}$ with $L = \max_{v\in\mathcal{V}} L_v$.*

Proof. Let us suppose ad absurdum that $Z^\star \in \Sigma_k^N$ is not a τ-stationary point. This means that there exists $\ell \in \mathcal{V}$ such that

$$y_\ell = \sigma_k(\overline{z}_\ell^\star - \tau\nabla f_\ell(z_\ell^\star)) \neq z_\ell^\star.$$

Let us consider now $\widetilde{Y} = (z_1^\star, z_2^\star, \ldots, y_\ell, \ldots, z_N^\star) \in \Sigma_k^N$. Then

$$\begin{aligned}
F(\widetilde{Y}) - F(Z^\star) &= \sum_{v\in\mathcal{V}}\left[f_v(\widetilde{y}_v) - f_v(z_v^\star)\right] \\
&+ \sum_{v\in\mathcal{V}}\sum_{w\in\mathcal{N}_v}\frac{1}{4\tau d}\left[\|\widetilde{y}_v - \widetilde{y}_w\|^2 - \|z_v^\star - z_w^\star\|^2\right] \\
&= f_\ell(y_\ell) - f_\ell(z_\ell^\star) + \sum_{w\in\mathcal{N}_\ell\setminus\{\ell\}}\frac{1}{2\tau d}\left[\|y_\ell - z_w^\star\|^2 - \|z_\ell^\star - z_w^\star\|^2\right] \\
&= \langle\nabla f_\ell(z_\ell^\star), y_\ell - z_\ell^\star\rangle + \frac{1}{2}(y_\ell - z_\ell^\star)^T\nabla^2 f_\ell(\xi)(y_\ell - z_\ell^\star) \\
&+ \sum_{w\in\mathcal{N}_\ell\setminus\{\ell\}}\frac{1}{2\tau d}\left[\|y_\ell - z_\ell^\star\|^2 + 2\langle y_\ell - z_\ell^\star, z_\ell^\star - z_w^\star\rangle\right]
\end{aligned}$$

where in the first equality the multiplying factor $1/2\tau d$ (instead of $1/4\tau d$) follows from the following observation:

$$\begin{aligned}
&\sum_{w\in\mathcal{N}_\ell\setminus\{\ell\}}\frac{1}{4\tau d}\left[\|y_\ell - z_w^\star\|^2 - \|z_\ell^\star - z_w^\star\|^2\right] \\
&\quad= \sum_{v\in\mathcal{V}\setminus\{\ell\}}\sum_{\ell\in\mathcal{N}_v}\frac{1}{4\tau d}\left[\|y_\ell - z_w^\star\|^2 - \|z_\ell^\star - z_w^\star\|^2\right]
\end{aligned} \tag{8}$$

and $\xi = (1-\gamma)z_\ell^\star + \gamma y_\ell$ for some $\gamma \in (0,1)$. Applying Lemma 1 to $y_\ell = \sigma_k(\overline{z}_\ell^\star - \tau\nabla f_\ell(z_\ell^\star))$ we have

$$\langle\nabla f_\ell(z_\ell^\star), y_\ell - z_\ell^\star\rangle \leq -\frac{1}{2\tau}\|y_\ell - z_\ell^\star\|^2 - \frac{1}{\tau d}\sum_{w\in\mathcal{N}_\ell}\langle z_\ell^\star - z_w^\star, y_\ell - z_\ell^\star\rangle$$

and, together with Assumption 2.c), we conclude

$$\begin{aligned}
&F(\widetilde{Y}) - F(Z^\star) \\
&\quad\leq -\frac{1}{2\tau}\|y_\ell - z_\ell^\star\|^2 + \frac{1}{2}L\|y_\ell - z_\ell^\star\|^2 + \frac{d-1}{2\tau d}\|y_\ell - z_\ell^\star\|^2 \\
&\quad\leq \frac{L\tau d - d + d - 1}{2\tau d}\|y_\ell - z_\ell^\star\|^2
\end{aligned}$$

or, equivalently,

$$F(Z^\star) - F(\widetilde{Y}) \geq \frac{1 - L\tau d}{2\tau d}\|y_\ell - z_\ell^\star\|^2 > 0 \tag{9}$$

contradicting the optimality of $Z^\star$. We conclude that

$$z_v^\star = \sigma_k(\overline{z}_v^\star - \tau\nabla f_v(z_v^\star)),\ \forall v \in \mathcal{V}.$$

□

The following theorem proves that, with the additional Assumption 3, τ-stationary points are local minima of (6). This means that τ-stationarity is necessary, but not sufficient for optimality.

Theorem 4. *Suppose that Assumptions 2, 3, and 4 hold. Any τ-stationary point of* (6) *is a local minimum for* (6)*.*

Proof. Let Z be a τ-stationary point of (6). We now show that for any $Z+H$ with $z_v + h_v \in \Sigma_k$ for all $v \in \mathcal{V}$ and $\|H\|_F < \epsilon$ we have $F(Z+H) - F(Z) > 0$. Fix $\epsilon = \min_{v\in\mathcal{V}} r^k(z_v)$ and $z_v + h_v \in \Sigma_k$ for all $v \in \mathcal{V}$. This fact implies that

$$\begin{aligned}
F(Z+H) - F(Z) &= \sum_{v\in\mathcal{V}}\left[f_v(z_v + h_v) - f_v(z_v)\right] \\
&+ \sum_{v\in\mathcal{V}}\sum_{w\in\mathcal{N}_v}\frac{1}{4\tau d}\left[\|z_v + h_v - z_w - h_w\|^2 - \|z_v - z_w\|^2\right] \\
&= \sum_{v\in\mathcal{V}}\left[\langle\nabla f_v(z_v), h_v\rangle + \frac{1}{2}h_v^T\nabla^2 f_v(\xi_v)h_v\right] \\
&+ \sum_{v\in\mathcal{V}}\sum_{w\in\mathcal{N}_v}\frac{1}{4\tau d}\left[\|h_v - h_w\|^2 + 2\langle z_v - z_w, h_v - h_w\rangle\right]
\end{aligned}$$

where $\xi_v = z_v + \gamma_v h_v$ for some $\gamma_v \in (0,1)$. It can be easily checked that in a regular graph

$$\frac{1}{d}\sum_{v\in\mathcal{V}}\sum_{w\in\mathcal{N}_v}\langle z_w, h_w\rangle = \sum_{v\in\mathcal{V}}\langle z_v, h_v\rangle$$

and

$$\frac{1}{d}\sum_{v\in\mathcal{V}}\sum_{w\in\mathcal{N}_v}\langle z_v,h_w\rangle=\frac{1}{d}\sum_{v\in\mathcal{V}}\sum_{w\in\mathcal{N}_v}\langle z_w,h_v\rangle$$

from which

$$\begin{aligned}&F(Z+H)-F(Z)\\&=\sum_{v\in\mathcal{V}}\left[\sum_{w\in\mathcal{N}_v}\frac{1}{4\tau d}\|h_v-h_w\|^2+\frac{1}{2}h_v^T\nabla^2 f_v(\xi_v)h_v\right]\\&+\frac{1}{\tau}\sum_{v\in\mathcal{V}}\langle z_v-(\overline{z}_v-\tau\nabla f_v(z_v)),h_v\rangle.\end{aligned}\tag{10}$$

Recall that Z is also a τ-stationary point and from Corollary 1 also a BF point of (6). Therefore, if $\|z_v\|_0<k$ then

$$z_v-(\overline{z}_v-\tau\nabla f_v(z_v))=0$$

otherwise, if $\|z_v\|_0=k$ then the constraint $z_v+h_v\in\Sigma_k$ with $\|H\|_F<\epsilon$ implies that $\mathrm{supp}(h_v)\subseteq\mathrm{supp}(z_v)$ and

$$\langle z_v-(\overline{z}_v-\tau\nabla f_v(z_v)),h_v\rangle=0,$$

$$F(Z+H)-F(Z)\geq\sum_{v\in\mathcal{V}}\frac{1}{2}h_v^T\nabla^2 f_v(\xi)h_v\geq 0.$$

If there exists $\overline{v}\in\mathcal{V}$ such that $f_{\overline{v}}$ satisfies the α_k-LSRHP with $\alpha_k>0$ then the last inequality is strict. □

Let us state the relation between the global minimizer of (6) and of (2).

Theorem 5. *Suppose that Assumptions 1, 2, and 3 hold. Let us denote as $\widehat{X}^\tau$ the minimizer of $F(X)$ in* (6)*. If $\mathcal{G}$ is connected, then $\lim_{\tau\to0}\widehat{X}^\tau=x_{opt}\mathbb{1}^T$, where x_{opt} is the minimizer of* (2)*.*

Proof. We prove the assertion by showing

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

$$\lim_{\tau\to0}\|\widehat{x}_v^\tau-\widehat{x}_w^\tau\|=0\quad\forall v,w\in\mathcal{V};$$

ii. the convergence to a common value

$$\forall v\in\mathcal{V}\quad\lim_{\tau\to0}\widehat{x}_v^\tau=x_\star,$$

which is the solution of (2), *i.e.* $f(x_\star)\leq f(x),\forall x\in\Sigma_k$.

We start with point i.. From Assumption 2 and from the definition of global minimizer for (6) we have that there exists $(v,w)\in\mathcal{E}$ such that

$$\sum_{v\in\mathcal{V}}\kappa_v+\frac{1}{4\tau d}\|\widehat{x}_v^\tau-\widehat{x}_w^\tau\|^2\leq F(\widehat{X}^\tau)<F(0)=\sum_{\ell\in\mathcal{V}}f_\ell(0)$$

from which we conclude that $\frac{1}{4\tau d}\|\widehat{x}_v^\tau-\widehat{x}_w^\tau\|^2$ is bounded and $\|\widehat{x}_v^\tau-\widehat{x}_w^\tau\|^2\to0$ as $\tau\to0$. Since the graph is connected we deduce that $\lim_{\tau\to0}\|\widehat{x}_v^\tau-\widehat{x}_w^\tau\|=0,\ \forall v,w\in\mathcal{V}$.

We now prove point ii.: by definition of $\widehat{X}^\tau$ we have for all $X\in\Sigma_k^N$

$$\sum_{v\in\mathcal{V}}f_v(\widehat{x}_v^\tau)\leq F(\widehat{X}^\tau)\leq\sum_{v\in\mathcal{V}}\left[f_v(x_v)+\frac{1}{4\tau d}\sum_{w\in\mathcal{N}_v}\|x_w-x_v\|_2^2\right].$$

and, in particular, by definition of minimum, there exists $C=\sum_{v\in\mathcal{V}}f_v(0)$

$$\sum_{v\in\mathcal{V}}f_v(\widehat{x}_v^\tau)\leq F(\widehat{X}^\tau)\leq C=\sum_{v\in\mathcal{V}}f_v(0).$$

From Assumption 2 and Assumption 3 we have that there exist $\overline{v}\in\mathcal{V}$, $\alpha_k^{\overline{v}}>0$ and $\xi=\gamma\widehat{x}_{\overline{v}}^\tau$ with $\gamma\in(0,1)$ such that

$$\begin{aligned}f_{\overline{v}}(\widehat{x}_{\overline{v}}^\tau)&=f(0)+\langle\nabla f_{\overline{v}}(0),\widehat{x}_{\overline{v}}^\tau\rangle+\frac{1}{2}(\widehat{x}_{\overline{v}}^\tau)^T\nabla^2 f(\xi)\widehat{x}_{\overline{v}}^\tau\\&\geq f(0)+\langle\nabla f_{\overline{v}}(0),\widehat{x}_{\overline{v}}^\tau\rangle+\frac{1}{2}\alpha_k^{\overline{v}}\|\widehat{x}_{\overline{v}}^\tau\|^2\end{aligned}$$

and, consequently,

$$\begin{aligned}C\geq F(\widehat{X}^\tau)&\geq\sum_{v\in\mathcal{V}\setminus\overline{v}}f_v(\widehat{x}_v^\tau)+f_{\overline{v}}(\widehat{x}_{\overline{v}}^\tau)\\&\geq\sum_{v\in\mathcal{V}\setminus\overline{v}}f_v(\widehat{x}_v^\tau)+f(0)+\langle\nabla f_{\overline{v}}(0),\widehat{x}_{\overline{v}}^\tau\rangle+\frac{1}{2}\alpha_k^{\overline{v}}\|\widehat{x}_{\overline{v}}^\tau\|^2\\&=\sum_v\kappa_v+\langle\nabla f_{\overline{v}}(0),\widehat{x}_{\overline{v}}^\tau\rangle+\frac{1}{2}\alpha_k^{\overline{v}}\|\widehat{x}_{\overline{v}}^\tau\|^2.\end{aligned}$$

We deduce that $\widehat{x}_{\overline{v}}^\tau$ is bounded for any τ. Then for any sequence $\{\widehat{x}_{\overline{v}}^{\tau_\ell}\}_{\ell\in\mathbb{N}}$ we can extract a convergent subsequence and from point i. we deduce that $\{\widehat{x}_{\overline{v}}^{\tau_{\ell_s}}\}_{s\in\mathbb{N}}$ such that $\lim_{s\to\infty}\widehat{x}_v^{\tau_{\ell_s}}=\xi$ for all $v\in\mathcal{V}$.

By letting $s\to\infty$ and considering that F is a continuous function, we obtain

$$f(\xi)=\lim_{s\to\infty}F(\widehat{X}^{\tau_{\ell_s}})\leq f(x),\quad\forall x\in\Sigma_k.$$

Repeating the argument for any subsequence and from uniqueness of the global minimizer we conclude that $\lim_{\tau\to0}\widehat{x}_v^\tau=\xi=x_{opt}$.

□

Theorem 5 guarantees that parameter τ can be interpreted as a temperature; as τ decreases, estimates x_v's associated with adjacent nodes become increasingly correlated. This suggests that if τ is sufficiently small, then each vector $\widehat{x}_v^\tau$ can be used as an approximation of the optimal solution x_{opt} of (2).

We are now ready to state the main convergence result, whose proof is postponed to the Appendix.

Theorem 6 (AHT convergence)**.** *Let $X(0)\in\Sigma_k^N$ and $\{X(t)\}_{t\in\mathbb{N}}$ be the sequence generated by AHT. Under Assumption 2, 3, 4, and assuming that the network is uniformly persistent (see Definition 4), if $\tau<1/(dL)$ then the sequence $\{X(t)\}_{t\in\mathbb{N}}$ converges to a τ-stationary point of* (6)*.*

Theorem 6 guarantees that, under proper assumptions, the AHT converges to a limit point. The proof can be extended to randomly persistent networks (see Definition 5) with similar techniques to [37]. In the proposed numerical experiments (see Section VI), nodes are sampled according to uniform distributions: this choice is made for simplicity, but the approach can in principle be extended to any other distribution that guarantees that the network is randomly persistent. Then, the convergence has to be intended almost surely.[1].

Theorem 7. *Let us consider the scenario described in Definition 4 and assume that Assumption 2, 3, 4 hold. Let $\tau<1/(dL)$ and $\widehat{X}^\tau$ be the limit point produced by AHT with initial condition $X(0)=0$. Then $\lim_{\tau\to0}\widehat{X}^\tau=\widetilde{x}\mathbb{1}^T$, where $\widetilde{x}$ is a local minimum of* (2)*.*

[1]We say that the sequence $\{X(t)\}_{t\in\mathbb{N}}$ converges almost surely towards X if events for which $\{X(t)\}_{t\in\mathbb{N}}$ does not converge to X have probability 0.

Proof. It can be proved that for any $\tau < 1/(dL)$ the sequence $\{X(t)\}_{t\in\mathbb{N}}$ generated by AHT is such that $F(X(t+1)) \leq F(X(t))$ for every $t \in \mathbb{N}$ (see Lemma 3 in Appendix). Therefore, the limit point $\widehat{X}^\tau$, that is the output of AHT with initial condition $X(0) = 0$, is such that $F(\widehat{X}^\tau) \leq F(0)$. This fact and the argument used to prove Point i. in the proof of Theorem 6 imply that $\lim_{\tau\to 0} \widehat{x}_v^\tau = \widetilde{x},\ \forall v \in \mathcal{V}$.

We now prove that $\widetilde{x}$ is a BF vector and, consequently, a local minimum of (2) (see Theorem 1). We consider the case when $\|\widetilde{x}\|_0 = k$ (the case with $\|\widetilde{x}\|_0 < k$ can be treated with similar arguments). Let us fix $\epsilon \in (0, \min_{j\in\mathrm{supp}(\widetilde{x})} |\widetilde{x}^j|)$; then there exists τ_0 such that for all $\tau \in (0, \tau_0)$ it holds $|(\widehat{x}_v^\tau)^j - \widetilde{x}^j| < \epsilon$. This implies that if $j \in \mathrm{supp}(\widetilde{x})$, then $j \in \mathrm{supp}(\widehat{x}_v^\tau)$ for any $\tau \in (0, \tau_0)$ and for all $v \in \mathcal{V}$. We deduce that for any $j \in \mathrm{supp}(\widehat{x}_v^\tau)$, for any $\tau \in (0, \tau_0)$, and for all $v \in \mathcal{V}$

$$(\widehat{x}_v^\tau)^j = \frac{1}{d}\sum_{w\in\mathcal{N}_v} (\widehat{x}_w^\tau)^j - \tau\nabla_j f_v(\widehat{x}_v^\tau)$$

from which, computing the average over all possible nodes, we obtain $\frac{1}{N}\sum_{v\in\mathcal{V}} \nabla_j f_v(\widehat{x}_v^\tau) = 0$ for any $j \in \mathrm{supp}(\widehat{x}_v^\tau)$ and for any $\tau \in (0, \tau_0)$. By letting τ go to 0 and by the fact that ∇f_v are Lipschitz continuous we obtain $\sum_{v\in\mathcal{V}} \nabla_j f_v(\widetilde{x}) = \nabla_j f(\widetilde{x}) = 0$ for all $j \in \mathrm{supp}(\widetilde{x})$. From Definition 2 $\widetilde{x}$ is a BF vector. □

B. Fixed points analysis for BHT and GHT

In this section we characterize the fixed points of the maps that rule the GHT dynamics. A similar argument can be used for BHT.

Let us denote for each $(v, w) \in \mathcal{E}$ the map $\phi_{(v,w)} : \mathbb{R}^{n\times N} \to \mathbb{R}^{n\times N}$ which acts on $X = (x_1, \ldots, x_N)$ as

$$(\phi_{(v,w)}(X))_u = \begin{cases} x_u & \text{if } u \neq v \\ \sigma_k(\frac{x_v + x_w}{2} - \tau\nabla f_v(x_v)) & \text{if } u = v. \end{cases}$$

Definition 8. $Z \in \mathbb{R}^n$ is called a fixed point of $\Phi = \{\phi_{(v,w)} : (v, w) \in \mathcal{E}\}$ if

$$Z \in \bigcap_{(v,w)\in\mathcal{E}} \{X \in \mathbb{R}^{n\times N} : \phi_{(v,w)}(X) = X\}.$$

The set of fixed point is denoted with $\mathrm{Fix}(\Phi)$

In the following theorem, we prove that each fixed point is a consensus point and we give its characterization.

Theorem 8. *If $\mathcal{G}$ is connected, then for any $X \in \mathrm{Fix}(\Phi)$, there exists $x \in \mathbb{R}^n$ such that $X = x\mathbb{1}^T$, and $x = \sigma_k\left(x - \frac{\tau}{N}\nabla f(x)\right)$.*

Proof. Let $X \in \mathrm{Fix}(\Phi)$ and $S_v = \mathrm{supp}(x_v)$. By definition, for any $j \in S_v$

$$x_v^j = \frac{x_v^j + x_w^j}{2} - \tau\nabla_j f_v(x_v) \tag{11}$$

for all $w \in \mathcal{N}_v$ (included v itself). We then obtain $\nabla_j f_v(x_v) = 0$ and, consequently, $x_v^j = x_w^j \neq 0$ for all $w \in \mathcal{N}_v$. If $\mathcal{G}$ is connected then there exists a path connecting every pair of vertices. Iterating the argument in (11) for all edges in the aforementioned path we conclude that $x_v^j = x_w^j \neq 0$ for all $v, w \in \mathcal{V}$. We thus have $X = x\mathbb{1}^T$ with $x = \sigma_k(x - \tau\nabla f_v(x))$ for all $v \in \mathcal{V}$.

Let $S = \mathrm{supp}(x)$. By definition of fixed point, we obtain for each $v \in \mathcal{V}$ and $\forall j \in S$ that $\nabla_j f_v(x) = 0$ and, consequently,

$$\sum_v \nabla_j f_v(x) = \nabla_j \sum_v f_v(x) = \nabla_j f(x) = 0. \tag{12}$$

If $j \notin S$ we can write for all $v \in \mathcal{V}$

$$\begin{aligned} |x^j - \tau\nabla_j f_v(x)| &< r^k(x - \tau\nabla_j f_v(x)) \\ &= |x^{i_k} - \tau\nabla_{i_k} f_v(x)| = |x^{i_k}| \end{aligned}$$

where the last equality follows from the fact that $i_k \in S$. Let us assume that $x^{i_k} > 0$ (the case $x^{i_k} < 0$ can be proved in an analogous way), then $x^j - \tau\nabla_j f_v(x) \in (-x^{i_k}, x^{i_k})$ for all $v \in \mathcal{V}, j \notin S$, from which, by multiplying each inclusion by $1/N$ and summing over all possible nodes, we obtain $x^j - \frac{\tau}{N}\nabla_j f(x) \in (-x^{i_k}, x^{i_k})$ and hence

$$\begin{aligned} x^j - \frac{\tau}{N}\nabla_j f(x) &\geq -x^{i_k} - \frac{\tau}{N}\nabla_{i_k} f(x) \\ x^j - \frac{\tau}{N}\nabla_j f(x) &\leq +x^{i_k} - \frac{\tau}{N}\nabla_{i_k} f(x) \end{aligned} \tag{13}$$

where the last inclusion follows from (12), being $i_k \in S$. From (12) and (13), we conclude the proof. □

Theorem 8 guarantees that fixed points of GHT and BHT are stationary points of (2). Finally, we recall the following result.

Theorem 9. *Let $\mathcal{G}$ be connected, let f satisfy the α_{2k}-LSRHP with $\alpha_{2k} > 0$ and $\tau \in (\frac{3}{4\alpha_{2k}}, \frac{1}{\sum_{v\in\mathcal{V}} L_v})$. If $\widetilde{x}$ is a τ stationary point for (2) then*

$$\|\widetilde{x} - x_{opt}\|_2 \leq c(\tau, \alpha_{2k}) f(x_{opt})$$

where x_{opt} is the optimal solution of (2) and $c(\tau, \alpha_{2k}) > 0$ is a function of τ and α_{2k}.

It should be noted that if $f(x^\star) = 0$, Theorem 9 implies that $\widetilde{x} = x^\star$. The proof, omitted for brevity, can be obtained immediately using techniques devised in Theorem 5 in [23].

VI. Numerical results

In this section, we present some results of numerical tests about AHT, BHT, and GHT on compressed sensing.

We focus on noise-free scenarios, we study the behavior of our algorithms in terms recovery accuracy and transmission efficiency. Finally, we report some considerations about how to tackle the noisy cases.

A. Example (continued): recovery accuracy

We start by presenting some numerical results in the noise-free compressed sensing framework (*i.e.*, $\xi_v = 0$ for all $v \in \mathcal{V}$). In all the simulations we have performed, we have observed that convergence to the true signal can be achieved by AHT, BHT, and GHT, provided that the number of measurements is large enough (but keeping the number of measurements per node smaller than the number sufficient for individual reconstruction).

The considered setting is as follows. We fix the parameters $n = 200$ and $k = 10$. The nonzero components' positions are chosen uniformly at random; the amplitude of each nonzero component is drawn from a Gaussian distribution $\mathsf{N}(0, 1)$. The sensing matrices $A_v \in \mathbb{R}^{m\times n}$ are sampled from the Gaussian ensemble: $A_v^{ij} \sim \mathsf{N}(0, 1/m)$, $\forall v \in \mathcal{V}$ which is a popular choice in compressed sensing [31]. The measurements are $y_v = A_v x^\star \in \mathbb{R}^m$.

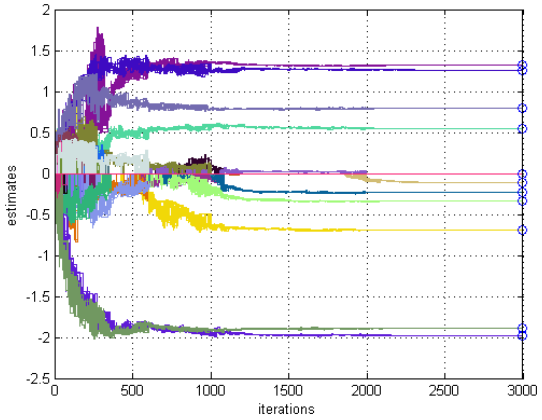


Figure 2. Noise-free compressed sensing, $n = 200$, $k = 10$, $m = 15$, $N = 10$, $\tau = 0.1$, complete graph: GHT converges to $x^\star$ (whose components are marked by blue circles). The same color is assigned to the same estimates' components for each node.

An illustrative single example where perfect recovery is achieved is shown in Figure 2: here, GHT is implemented, over a complete graph, with $m = 15$, $N = 10$, $\tau = 0.1$. Clearly, $m = 15$ is not sufficient for individual recovery, but collaboration among the 10 agents allows to get it, in a reasonable number of iterations.

We assume that each agent can in turn acquire $m = 10, 15, 20$ measurements and we study how the recovery accuracy varies at the increasing of the number of agents. On one hand, adding agents we augment the total number of measurements, which is expected to improve the recovery; on the other hand, a larger network may cause some degradation due to greater decentralization.

In Figure 3, we show that the good effect prevails, that is, increasing the total number of measurements mN (namely, the network size, having fixed m) the performance accuracy improves. More precisely, in Figure 3 we show the results over two different topologies: ring (all the nodes communicates with two neighbors), and random geometric (we assign a uniformly random position to each node in the square $[0, 1] \times [0, 1]$, and we let communication between nodes with distance below a certain radius, in this case 0.75, [38]). The ring topology represents the least connected, regular case, while with the random geometric topology we explore the non-regular framework (for which we do not have theoretical guarantees). The algorithms are stopped at time T such that $\sum_{v\in\mathcal{V}} \left\|x_v^{T-1} - x_v^T\right\|_2^2 < 10^{-15}$ or after $T = 2\times10^5$ iterations, whichever occurred first [8, Section V.C].

The graphs show the rate of success (we declare a success when the accuracy condition $\frac{\sum_{v\in\mathcal{V}}\left\|x_v^T - x^\star\right\|_2^2}{N\|x^\star\|_2^2} < 10^{-4}$ holds) as a function of mN for AHT, BHT, and GHT. All the results are obtained by averaging over 500 different runs. For these experiments, for each node $v \in \mathcal{V}$ we choose a $\tau_v = N^{-1}\left\|A_v\right\|_2^{-2}$ ($N\left\|A_v\right\|_2^2$ turns out to be a good individual approximation for the theoretical bound $d\max_{v\in\mathcal{V}}\left\|A_v\right\|_2^2$). In all the figures, we draw the curve for the (centralized) IHT as a benchmark.

Table II
SPARCO PROBLEMS' SETTING

Problem	n	m	N	k
Sparco 902	1000	4	50	3
Sparco 7	2560	15	40	20

Observing Figure 3, we conclude that BHT tends to work better in the few measurements regime, immediately followed by GHT. AHT is a bit less reliable, in particular for $m = 10$ and ring topology. This can be explained with the presence of many stationary points when the number of measurements is low, among which the search of the true signal is even more difficult over few connected networks that do not support much collaboration. We finally notice that the gap with IHT is not dramatic, and that a 95% of success is achieved by all the algorithms, over the different topologies, for $mN \geq 120$ (except for AHT with $m = 10$ in the ring topology).

B. Example (continued): number of sent values

We now study the efficiency of AHT, BHT, and GHT in terms of number of transmitted values necessary to achieve convergence (by transmitted value we mean a real scalar sent over a communication link in the network). As communications typically are energy expensive, we aim to keep that number as low as possible. To the best of our knowledge, among the possible distributed approaches to problem (2), DIHT [8] is the most efficient in terms of transmitted values: in [8, Section V], some tests are proposed that compare DIHT to CB-DIHT, D-ADMM and subgradient methods, and the outcomes attest its higher performance. Those tests were conducted on a set of sparse problems selected from the Sparco dataset [39]; here, we consider a couple of those Sparco problems, with the same network topologies and parameters, and we show that AHT, BHT and GHT outperform DIHT.

We retrieve the experiments of [8]. More precisely, we consider the following problems: (a) Sparco 7, in which a sign spike signal is compressed through a Gaussian matrix; (b) Sparco 902, in which a signal sparse in the DCT domain has to be recovered (see [39]). Signals' lengths, number of measurements, and network specifications that we assume (see Table II) are taken from [8, Table 1]). We consider Erdos-Rényi (ER) and random geometric graphs (Geo), respectively with connectivity parameters 0.25, 0.75 and 0.5, 0.75. The setting that we consider is then envisaged in [8, Section V], the unique differences being in the different realizations for the random graphs and in the number of instances: our results are averaged over 100 instances, while in [8] 5 runs were performed.

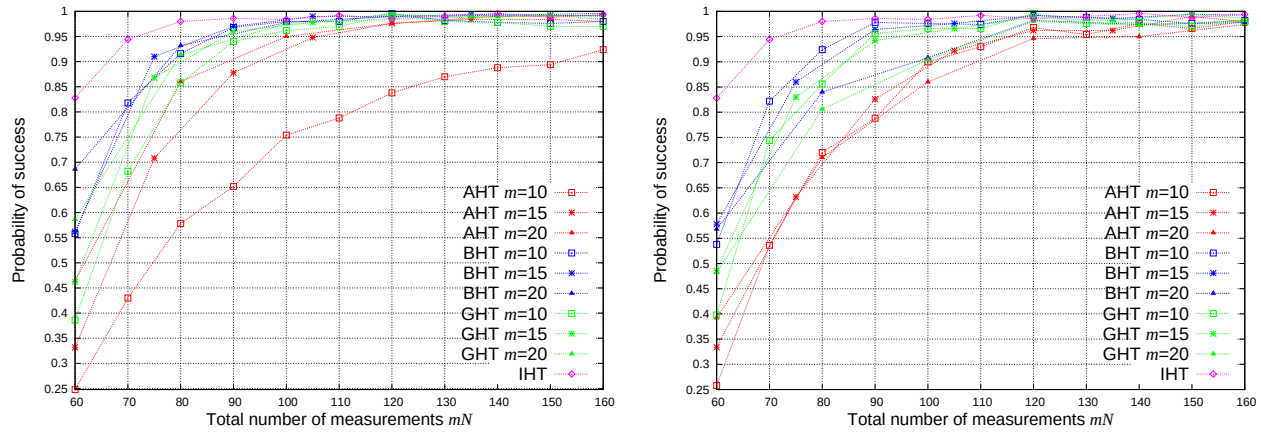


Figure 3. Noise-free compressed sensing: probability of success over a ring (left) and over a random geometric graph with radius 0.75 (right) as a function of mN

Table III
SPARCO TESTS: TOTAL NUMBER OF SENT VALUES TO CONVERGE. BHT AND GHT CONVERGE TO THE TRUE SIGNAL $x^\star$, WHILE AHT CONVERGES TO A STATIONARY POINT (WHICH IS INDICATED BY *). A SUBSTANTIAL REDUCTION OF SENT VALUES IS OBTAINED WITH RESPECT TO DIHT [8]

Accuracy →	10^{-2}				10^{-5}			
	BHT mean	**BHT** min - max	**GHT** mean	**GHT** min - max	**BHT** mean	**BHT** min - max	**GHT** mean	**GHT** min - max
	Sparco 902							
ER $p=0.25$	1.67×10^4	15768 - 18108	1.71×10^4	16110 - 18300	4.27×10^4	41616 - 43362	4.31×10^4	39492 - 45054
ER $p=0.75$	2.72×10^4	25944 - 29652	2.77×10^4	26994 - 28482	5.07×10^4	48414 - 54300	5.24×10^4	50826 - 53580
Geo $d=0.5$	2.30×10^4	19104 - 26070	2.30×10^4	20634 - 24456	4.72×10^4	44364 - 50586	4.87×10^4	45408 - 50424
Geo $d=0.75$	2.63×10^4	24168 - 28998	2.68×10^4	26244 - 27372	4.94×10^4	44676 - 53694	5.14×10^4	50250 - 52398
	Sparco 7							
ER $p=0.25$	1.82×10^5	165360 - 188760	1.57×10^5	133200 - 182440	3.79×10^5	359240 - 391840	3.67×10^5	341040 - 395600
ER $p=0.75$	1.89×10^5	178600 - 203360	1.73×10^5	159960 - 184560	3.74×10^5	362560 - 396040	3.56×10^5	340960 - 372440
Geo $d=0.5$	1.66×10^5	157640 - 173080	1.42×10^5	125040 - 152600	3.57×10^5	347880 - 372840	3.38×10^5	314560 - 352480
Geo $d=0.75$	1.76×10^5	160240 - 188480	1.70×10^5	156640 - 178760	3.62×10^5	334400 - 383760	3.57×10^5	344360 - 366080
	AHT* mean	**AHT*** min - max	**DIHT**		**AHT*** mean	**AHT*** min - max	**DIHT**	
	Sparco 902							
ER $p=0.25$	4.58×10^4	36690 - 58254	2.32×10^6		3.02×10^5	290874 - 316698	5.67×10^6	
ER $p=0.75$	3.71×10^5	337524 - 402828	2.32×10^6		1.23×10^6	1204188 - 1254288	5.67×10^6	
Geo $d=0.5$	1.78×10^5	145758 - 233646	2.32×10^6		7.02×10^5	585984 - 785532	5.67×10^6	
Geo $d=0.75$	3.79×10^5	332460 - 434796	2.32×10^6		1.25×10^6	1122006 - 1333182	5.67×10^6	
	Sparco 7							
ER $p=0.25$	4.98×10^5	375320 - 623280	6.39×10^6		2.44×10^6	2252680 - 2632720	1.39×10^7	
ER $p=0.75$	1.70×10^6	1611440 - 1761160	6.39×10^6		6.80×10^6	6655040 - 6932000	1.39×10^7	
Geo $d=0.5$	8.29×10^5	607840 - 1002960	6.39×10^6		4.43×10^6	3680560 - 4894320	1.39×10^7	
Geo $d=0.75$	1.94×10^6	1655360 - 2444440	6.39×10^6		7.54×10^6	6838160 - 8961760	1.39×10^7	

In Table III we present our experimental results of the implementation of AHT, BHT, GHT, and DIHT on Sparco 7 and Sparco 902. As in [8, Section V.C], (a) algorithms are stopped at the time T such that $\sqrt{\frac{\sum_{v\in\mathcal{V}}\|x_v^T-x^\star\|_2^2}{N\|x^\star\|_2^2}}<\mathsf{tol}$, with $\mathsf{tol}=10^{-2}$ or $\mathsf{tol}=10^{-5}$; (b) $x^\star$ is the true original signal or the stationary point to which the algorithm converges. In the following, we will show that not only AHT, but also BHT and GHT are always convergent in this setting, which makes unnecessary to fix an upper limit of iterations.

As in [8], in the implementation of DIHT we have considered $\tau=\frac{1}{2.01}$. The results that we obtain (in which we neglect the communications necessary to build the spanning tree) are substantially consistent with those presented in the original paper. It should be noted that the number of sent values for DIHT does not depend on the given topology, as communications are performed on the spanning tree which always has $N-1$ edges. The number of transmitted values is then $(N+1)(2k+n)T$, where T is the number of iterations to get convergence, $2k$ is the number of values required to diffuse the current k-sparse estimate from the root to the leaves, while n is the length of the non sparse gradients that are accumulated and sent from the leaves to the root.

Concerning AHT, BHT. and GHT, we have fixed $\tau=0.01$. For the Sparco 7 problem, no particular initialization is required, and the initial estimates are fixed to zero. For Sparco 902, instead, we assume that before starting the iterative procedure, each node v computes $x_v(0)=\sigma_k\left[\sum_{w\in\mathcal{N}_v}\tau A_w^{\mathsf{T}}y_w\right]$ (if connectivity is high, the sum can be reduced over a selection of neighbors). This initialization has been experimentally proved to speed up the convergence. The total number of sent values is evaluated as $2kT(\sum_{t=1}^{T}l(t)+I)$, where T is the total number of iterations to get convergence, $l(t)$ is the number of used links at each time step (this number depends on the nodes' degree for AHT and BHT, while is simply 1 for GHT)

and I is the number of used links in case of initialization. Notice that, as a difference from DIHT, only k-sparse vectors are transmitted (from which the coefficient $2k$), as each node performs hard thresholding before transmitting.

We remind that in principle BHT and GHT may not converge, and in particular may not converge to $x^\star$; however, for Sparco 7 and 902 considered in Table II, we always observe convergence to $x^\star$, no matter which communication topology is assumed. In the table, we show mean, the minimum and the maximum number of sent values we have obtained over 100 runs (for each run, a different topology is generated). We point out that BHT, GHT, and DIHT converge to the true signal, while AHT converges to a stationary point different from $x^\star$ (this fact is highlighted by a * in Table III).

In Table III we can appreciate the gain obtained by AHT, BHT, and GHT with respect to DIHT: in the 100 runs we consider, the number of sent values is always smaller using our methods, which are then expected to outperform also CB-DIHT, D-ADMM and subgradient methods, according to the results in [8].

We finally remark that AHT and BHT require less sent values over less connected topologies. This means that a limited collaboration, which reduces the number of used links at each iteration step, does not necessarily slow down the total algorithms' dynamics.

C. Example (continued): approach to the noisy case

When noise occurs, *i.e.*, $\xi_v \neq 0$, the estimates provided by GHT and BHT may oscillate and not converge in a deterministic sense. This is not surprising, as $x^\star$ is no more a fixed point of the dynamics being $f_v^{\text{ls}}(x^\star) > 0$ for all $v \in \mathcal{V}$. However, simulations show that the oscillations asymptotically concentrate around a mean value that approximates $x^\star$. Therefore, oscillations can be smoothed out performing a time-averaging operation as in [40], which is reported in Algorithm 5 and must be added as last instruction in Algorithms 3 and 4. This inner-loop just requires that each agent individually stores the number of times it has woken up in the variable $\kappa_v(t)$ and uses it to construct a time-averaged estimate $\widetilde{x}_v(t)$; no knowledge of global clocks or any other global variables is needed.

An example is presented in [41, Figure 3], where the smoothing effect can be appreciated: for any v, $\widetilde{x}_v(t)$ converges in a neighborhood of $x^\star$. The analysis of the ergodicity of the dynamics Algorithm 4 in case of noisy measurements and, consequently, the convergence of $\widetilde{x}_v(t)$ is left for future research. We refer to [37] for an overview of ergodic dynamics over networks.

It is also worth mentioning that different implementations have been proposed to smooth the oscillations of a dynamical system and guarantee the almost sure convergence of the system. In [42] this goal is achieved using diminishing step-sizes, which damp the gradient step in the long run, while maintaining it active for a sufficiently long time. A thorough comparison of the latter approach with GHT is an interesting topic for future research.

Algorithm 5 Smoothing procedure

Require: $\theta(t) = (v, w)$, $x_v(t+1)$, $x_w(t+1)$

$\kappa_v(t+1) = \kappa_v(t) + 1$

$\kappa_w(t+1) = \kappa_w(t) + 1$

$\kappa_h(t+1) = \kappa_h(t)$ for any $h \neq v, w$

$\widetilde{x}_v(t+1) = \frac{1}{\kappa_v(t+1)}(\kappa_v(t)\widetilde{x}_v(t) + x_v(t+1))$

$\widetilde{x}_w(t+1) = \frac{1}{\kappa_w(t+1)}(\kappa_w(t)\widetilde{x}_w(t) + x_w(t+1))$

$\widetilde{x}_h(t+1) = \widetilde{x}_h(t)$ for any $h \neq v, w$

VII. Concluding remarks

In this paper, we have presented distributed, randomized algorithms for in-network optimization under sparsity constraints, which dramatically reduce the number of necessary transmissions and overcome synchronization issues. The algorithms are shown to converge almost surely to the right solution under some conditions, which have been investigated in a number of simulations.

References

[1] 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.

[2] A. Nedić, A. Ozdaglar, and P. A. Parrillo, "Constrained consensus and optimization in multi-agent networks," *IEEE Trans. Autom. Control*, vol. 55, pp. 922 – 938, 2010.

[3] S. Sundhar Ram, A. Nedić, and V. V. Veeravalli, "Distributed subgradient projection algorithm for convex optimization," in *Proc. of IEEE ICASSP*, 2009, pp. 3653 – 3656.

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

[5] J. Tsitsiklis, D. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE Trans. Autom. Control*, vol. 31, no. 9, pp. 803 – 812, 1986.

[6] F. Fagnani, S. M. Fosson, and C. Ravazzi, "A distributed classification/estimation algorithm for sensor networks," *SIAM J. Control Optim.*, vol. 52, no. 1, pp. 189 – 218, 2014.

[7] M. Rabbat and R. Nowak, "Distributed optimization in sensor networks," in *Proc. of IPSN*, New York, NY, USA, 2004, pp. 20 – 27.

[8] S. Patterson, Y. Eldar, and I. Keidar, "Distributed compressed sensing for static and time-varying networks," *IEEE Trans. Signal Process.*, vol. 62, no. 19, pp. 4931 – 4946, 2014.

[9] J. Cui, G. Pratx, B. Meng, and C. Levin, "Distributed MLEM: An iterative tomographic image reconstruction algorithm for distributed memory architectures," *IEEE Trans. Med. Imag.*, vol. 32, no. 5, pp. 957 – 967, 2013.

[10] D. Li, K. D. Wong, Y. H. Hu, and A. M. Sayeed, "Detection, classification, and tracking of targets," *IEEE Signal Process. Mag.*, vol. 19, no. 2, pp. 17 – 29, 2002.

[11] C.-Y. Chong and S. Kumar, "Sensor networks: evolution, opportunities, and challenges," *Proc. of the IEEE*, vol. 91, no. 8, pp. 1247 – 1256, 2003.

[12] J. Ke, P. Shankar, and M. A. Neifeld, "Distributed imaging using an array of compressive cameras," *Optics Communications*, vol. 282, pp. 185 – 197, 2009.

[13] S. Nikitaki and P. Tsakalides, "Localization in wireless networks via spatial sparsity," in *Signals, Systems and Computers (ASILOMAR)*, Nov 2010, pp. 236 – 239.

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

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

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

[17] ——, "D-ADMM: a communication-efficient distributed algorithm for separable optimization," *IEEE Trans. Signal Process.*, vol. 61, no. 10, pp. 2718 – –2723, 2013.

[18] C. Ravazzi, S. M. Fosson, and E. Magli, "Distributed soft thresholding for sparse signal recovery," in *Proc. of IEEE GLOBECOM*, 2013, pp. 3429 – 3434.

[19] ——, "Distributed iterative thresholding for l_0/l_1-regularized linear inverse problems," *IEEE Trans. Inf. Theory*, vol. 61, no. 4, pp. 2081 – 2100, 2015.

[20] A. Zymnis, S. Boyd, and E. Candes, "Compressed sensing with quantized measurements," *IEEE Signal Process. Lett.*, vol. 17, no. 2, pp. 149 – 152, 2010.

[21] A. Beck and Y. C. Eldar, "Sparsity constrained nonlinear optimization: Optimality conditions and algorithms." *SIAM J. Optim.*, vol. 23, no. 3, pp. 1480–1509, 2013.

[22] J. M. Bardsley, D. Calvetti, and E. Somersalo, "Hierarchical regularization for edge-preserving reconstruction of PET images," *Inverse Problems*, vol. 26, no. 3, pp. 1 – 16, 2010.

[23] T. Blumensath, "Compressed sensing with nonlinear observations and related nonlinear optimization problems," *IEEE Trans. Inf. Theory*, vol. 59, no. 6, pp. 3466–3474, 2013.

[24] W. Xu, M. Wang, J.-F. Cai, and A. Tang, "Sparse error correction from nonlinear measurements with applications in bad data detection for power networks," *IEEE Trans. Signal Process.*, vol. 61, no. 24, pp. 6175 – 6187, 2013.

[25] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Randomized gossip algorithms," *IEEE Trans. Inf. Theory*, vol. 52, no. 6, pp. 2508 – 2530, 2006.

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

[27] ——, "Iterative hard thresholding for compressed sensing," *Appl. Comput. Harmon. Ana.*, vol. 27, no. 3, pp. 265 – 274, 2009.

[28] S. Bahmani, B. Raj, and P. T. Boufounos, "Greedy sparsity-constrained optimization," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 807–841, 2013.

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

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

[31] 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.

[32] H. Erdogan and J. Fessler, "Monotonic algorithms for transmission tomography," *IEEE Trans. Med. Imag.*, vol. 18, no. 9, pp. 801–814, 1999.

[33] J. Mota, J. Xavier, P. Aguiar, and M. Puschel, "Basis pursuit in sensor networks," in *Proc. of IEEE ICASSP*, 2011, pp. 2916 – 2919.

[34] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.

[35] 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, Dec. 2008.

[36] 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.

[37] C. Ravazzi, P. Frasca, R. Tempo, and H. Ishii, "Ergodic randomized algorithms and dynamics over networks," *IEEE Trans. Control Netw. Syst.*, vol. 2, no. 1, pp. 78 – 87, 2015.

[38] M. Penrose, *Random Geometric Graphs (Oxford Studies in Probability)*. Oxford University Press, USA, Jul. 2003.

[39] E. v. Berg, M. P. Friedlander, G. Hennenfent, F. Herrmann, R. Saab, and Ö. Yılmaz, "Sparco: A testing framework for sparse reconstruction," Dept. Computer Science, University of British Columbia, Vancouver, Tech. Rep. TR-2007-20, 2007.

[40] B. T. Polyak and A. B. Juditsky, "Acceleration of Stochastic Approximation by Averaging," *SIAM J. Control Optim.*, vol. 30, no. 4, pp. 838 – 855, 1992.

[41] C. Ravazzi, S. M. Fosson, and E. Magli, "Energy-saving gossip algorithm for compressed sensing in multi-agent systems," in *Proc. of IEEE ICASSP*, 2014, pp. 5060 – 5064.

[42] A. G. Dimakis, S. Kar, J. M. F. Moura, M. G. Rabbat, and A. Scaglione, "Gossip algorithms for distributed signal processing," *Proceedings of the IEEE*, vol. 98, no. 11, pp. 1847 – 1864, 2010.

[43] T. M. Apostol, *Calculus. Vol. I. , One-variable calculus, with an introduction to linear algebra*. New-York: J. Wiley & Sons, 1967. [Online]. Available: http://opac.inria.fr/record=b1099210

[44] F. Bullo, J. Cortes, and S. Martinez, *Distributed Control of Robotic Networks*, ser. Applied Mathematics Series, 2009.

APPENDIX

The Appendix is devoted to the proof of Theorem 6, which requires some intermediate steps. In what follows, we indicate by $\{\theta(t)\}_{t\in\mathbb{N}}$ the sequence of activated nodes for AHT.

Lemma 2. *Let $X(0) \in \Sigma_k^N$ and $\{X(t)\}_{t\in\mathbb{N}}$ be the sequence generated by AHT and let $\{f_v\}_{v\in\mathcal{V}}$ satisfy Assumption 2. If, at a certain time step t, $\theta(t) = v$, then the AHT iterate satisfies the following relation:*

$$\langle \nabla f_v(x_v(t)), x_v(t+1) - x_v(t)\rangle \leq -\frac{1}{2\tau}\|x_v(t+1) - x_v(t)\|^2 - \frac{1}{\tau d}\sum_{w\in\mathcal{N}_v}\langle x_v(t) - x_w(t), x_v(t+1) - x_v(t)\rangle.$$

Proof. Since $\theta(t) = v$, we have $x_\ell(t+1) = x_\ell(t), \forall \ell \neq v$ and $x_v(t+1) = \sigma_k(\overline{x}_v(t) - \tau\nabla f_v(x_v(t)))$. The assertion is obtained applying Lemma 1. □

Lemma 3. *Let Assumption 4 hold, and let $X(0) \in \Sigma_k^N$ and $\{X(t)\}_{t\in\mathbb{N}}$ be the sequence generated by AHT and let $\{f_v\}_{v\in\mathcal{V}}$ satisfy Assumption 2. Then, given $L = \max_v L_v$, for any $t \in \mathbb{N}$*

$$F(X(t)) - F(X(t+1)) \geq \frac{(1/d - \tau L)}{2\tau}\|X(t+1) - X(t)\|_F^2.$$

Proof. Let us suppose that $\theta(t) = \ell$. Then the regularity of the graph implies that

$$\begin{aligned}
F(X(t+1)) &= \sum_{v\in\mathcal{V}\setminus\{\ell\}}\Bigg[f_v(x_v(t+1)) \\
&\quad + \frac{1}{4\tau d}\sum_{w\in\mathcal{N}_v\setminus\{\ell\}}\|x_v(t+1) - x_w(t+1)\|^2\Bigg] \\
&\quad + f_\ell(x_\ell(t+1)) + \frac{1}{2\tau d}\sum_{w\in\mathcal{N}_\ell\setminus\{\ell\}}\|x_\ell(t+1) - x_w(t+1)\|^2 \\
&= \sum_{v\in\mathcal{V}\setminus\{\ell\}}\left[f_v(x_v(t)) + \frac{1}{4\tau d}\sum_{w\in\mathcal{N}_v\setminus\{\ell\}}\|x_v(t) - x_w(t)\|^2\right] \\
&\quad + f_\ell(x_\ell(t+1)) + \frac{1}{2\tau d}\sum_{w\in\mathcal{N}_\ell\setminus\{\ell\}}\|x_\ell(t+1) - x_w(t)\|^2
\end{aligned}$$

where the multiplying factor $1/(2\tau d)$ in the last equation follows from (8). From Assumption 2 and the definition of

L we obtain

$$\begin{aligned}
F(X(t+1)) \leq \\
&\leq \sum_{v\in\mathcal{V}\setminus\{\ell\}}\left[f_v(x_v(t)) + \frac{1}{4\tau d}\sum_{w\in\mathcal{N}_v\setminus\{\ell\}}\|x_v(t)-x_w(t)\|^2\right]\\
&+ f_\ell(x_\ell(t)) + \langle\nabla f_\ell(x_\ell(t)), x_\ell(t+1)-x_\ell(t)\rangle\\
&+ \frac{L}{2}\|x_\ell(t+1)-x_\ell(t)\|^2 + \frac{1}{2\tau d}\sum_{w\in\mathcal{N}_\ell\setminus\{\ell\}}\|x_\ell(t)-x_w(t)\|^2\\
&+ \frac{1}{\tau d}\sum_{w\in\mathcal{N}_\ell\setminus\{\ell\}}\langle x_\ell(t)-x_w(t), x_\ell(t+1)-x_\ell(t)\rangle\\
&+ \frac{1}{2\tau}\frac{d-1}{d}\|x_\ell(t+1)-x_\ell(t)\|^2\\
&= F(X(t)) + \langle\nabla f_\ell(x_\ell(t)), x_\ell(t+1)-x_\ell(t)\rangle\\
&+ \frac{L}{2}\|x_\ell(t+1)-x_\ell(t)\|^2\\
&+ \frac{1}{\tau d}\sum_{w\in\mathcal{N}_\ell\setminus\{\ell\}}\langle x_\ell(t)-x_w(t), x_\ell(t+1)-x_\ell(t)\rangle\\
&+ \frac{1}{2\tau}\frac{d-1}{d}\|x_\ell(t+1)-x_\ell(t)\|^2.
\end{aligned}$$

Finally, applying Lemma 2, we get

$$\begin{aligned}
F(X(t+1)) &\leq F(X(t)) + \frac{L\tau-1}{2\tau}\|x_\ell(t+1)-x_\ell(t)\|^2\\
&+ \frac{1}{2\tau}\frac{d-1}{d}\|x_\ell(t+1)-x_\ell(t)\|^2\\
&= F(X(t)) + \frac{L\tau-1/d}{2\tau}\|x_\ell(t+1)-x_\ell(t)\|^2.
\end{aligned}$$

We conclude that

$$F(X(t+1)) \leq F(X(t)) + \frac{L\tau-1/d}{2\tau}\|X(t+1)-X(t)\|_F^2$$

being $x_v(t+1) = x_v(t)$, $\forall v \neq \ell$. □

Using this result, we can establish that two successive iterations of AHT become closer and closer. If AHT is executed with finite precision, it will converge numerically (when the distance between two consecutive estimates is below the machine epsilon, the algorithm reads the same value).

Proposition 3. *Let $X(0) \in \Sigma_k^N$ and $\{X(t)\}_{t\in\mathbb{N}}$ be the sequence generated by AHT, and let $\{f_v\}_{v\in\mathcal{V}}$ satisfy Assumption 2. If $\tau < 1/(dL)$ with $L = \max_{v\in\mathcal{V}} L_v$, then*

1) *there exists $0 < \alpha < \infty$ such that for all $T \geq 0$ $\sum_{t=0}^{T}\|X(t+1)-X(t)\|_F^2 \leq \alpha$;*
2) *$\lim_{t\to\infty}\|X(t+1)-X(t)\|_F^2 = 0$.*

Proof. Let us consider the sum over time

$$\begin{aligned}
&\sum_{t=0}^{T}(F(X(t)) - F(X(t+1)))\\
&\quad = F(X(0)) - F(X(T+1)) \leq F(X(0)) - c
\end{aligned}$$

where the last inequality follows from the fact F is lower bounded by a constant $c = \sum_v \kappa_v$. This bound holds for all $T \geq 0$. From Lemma 3 we have, for all $T \geq 0$,

$$\begin{aligned}
&\frac{1/d-\tau L}{2\tau}\sum_{t=0}^{T}\|X(t+1)-X(t)\|_F^2\\
&\quad\leq \sum_{t=0}^{T}(F(X(t)) - F(X(t+1)))\\
&\quad\leq F(X(0)) - c.
\end{aligned}$$

Since $F(X(0))$ is finite, this proves 1) with $\alpha = 2\tau(F(X(0)) - c)/(1/d - \tau L)$. It should be noted that $\left\{\sum_{t=0}^{T}\|X(t+1)-X(t)\|_F^2\right\}_{T\in\mathbb{N}}$ is monotonic increasing and upper bounded. We evince that it admits limit for $T \to +\infty$. □

Proof of Theorem 6

1) (Accumulation points – τ-stationary points) Let $\widetilde{X}$ be an accumulation point of the sequence $\{X(t)\}_{t\in\mathbb{N}}$, we want to prove that $\widetilde{X}$ is a τ-stationary point for (6). From definition of accumulation point, there exists a subsequence $\{X(t_s)\}_{s\in\mathbb{N}}$ converging to $\widetilde{X}$ [43]. Since from Proposition 3

$$\lim_{s\to\infty} F(X(t_s)) - F(X(t_s+1)) = 0$$

then, due to the continuity of the function F, $\lim_{s\to+\infty} X(t_s+1) = \widetilde{X}$. If the network is uniformly persistent (see Definition 4), all the nodes are activated infinitely many times. Let $\{t_\ell\}_{\ell\in\mathbb{N}}$ be the sequence, for which v has been updated . If we consider $j \in \operatorname{supp}(\widetilde{x}_v)$ then $\forall\epsilon \in (0, |\widetilde{x}_v^j|)$ there exists ℓ_0 such that for all $\ell \geq \ell_0$ it holds $|x_v^j(t_\ell) - \widetilde{x}_v^j| < \epsilon$. This implies that $j \in \operatorname{supp}(x_v(t_\ell))$ and $j \in \operatorname{supp}(x_v(t_\ell+1))$ for all $\ell \geq \ell_0$. Therefore, from line 4 of Algorithm 2, we have $\forall\ell > \ell_0$

$$x_v^j(t_\ell+1) = \frac{1}{d}\sum_{w\in\mathcal{N}_v} x_w^j(t_\ell) - \tau\nabla_j f_v(x_v(t_\ell)).$$

Taking ℓ go to infinity, the fact that $\{\nabla f_v(x)\}_{v\in\mathcal{V}}$ are Lipschitz-continuous implies that $\forall v \in \mathcal{V}$

$$0 \leq \|\nabla f_v(x_v(t_\ell)) - \tau\nabla f_v(\widetilde{x}_v)\| \leq L_v\|x_v(t_\ell) - \widetilde{x}_v\| \to 0,$$

and, consequently, we obtain for all $j \in \operatorname{supp}(\widetilde{x}_v)$

$$\widetilde{x}_v^j = \frac{1}{d}\sum_{w\in\mathcal{N}_v}\widetilde{x}_w^j - \tau\nabla_j f_v(\widetilde{x}_v^j). \tag{14}$$

Let us consider now the zero elements of $\widetilde{x}_v$, *i.e.* $j \notin \operatorname{supp}(\widetilde{x}_v)$. We want to prove that if $j \notin \operatorname{supp}(\widetilde{x}_v)$

$$\frac{1}{d}\sum_{w\in\mathcal{N}_v}\widetilde{x}_w^j - \tau\nabla_j f_v(\widetilde{x}_v^j) < r^k(\widetilde{x}_v)$$

(see (1) for the definition of r). Two scenarios have to be considered: (a) there exists a subsequence $\{t_{\ell_s}\}_{s\in\mathbb{N}}$ of $\{t_\ell\}_{\ell\in\mathbb{N}}$ such that $j \in \operatorname{supp}(x_v(t_{\ell_s}+1))$ (b) there exists ℓ_1 such that $j \notin \operatorname{supp}(x_v(t_\ell+1))$ for all $\ell > \ell_1$. In the first scenario we have

$$x_v^j(t_{\ell_s}+1) = \frac{1}{d}\sum_{w\in\mathcal{N}_v} x_w^j(t_{\ell_s}) - \tau\nabla_j f_v(x_v(t_{\ell_s})).$$

Moreover, if we let s go to infinity, using the fact that $\{\nabla f_v(x)\}_{v\in\mathcal{V}}$ are Lipschitz-continuous, we obtain

$$\widetilde{x}_v^j = \frac{1}{d}\sum_{w\in\mathcal{N}_v} \widetilde{x}_w^j - \tau\nabla_j f_v(\widetilde{x}_v).$$

Since $j \notin \mathrm{supp}(\widetilde{x}_v)$, then

$$\frac{1}{d}\sum_{w\in\mathcal{N}_v} \widetilde{x}_w^j - \tau\nabla_j f_v(\widetilde{x}_v^j) = 0 < r^k(\widetilde{x}_v).$$

In the second scenario, *i.e.* if there exists ℓ_1 such that $j \notin \mathrm{supp}(x_v(t_\ell+1))$ for all $\ell > \ell_1$ then

$$\left|\frac{1}{d}\sum_{w\in\mathcal{N}_v} x_w^j(t_\ell) - \tau\nabla_j f_v(x_v(t_\ell))\right| < r^k(x_v(t_\ell+1))$$

and, letting ℓ go to infinity, we finally obtain

$$\left|\frac{1}{d}\sum_{w\in\mathcal{N}_v} \widetilde{x}_w^j - \nabla_j f_v(\widetilde{x}_v^j)\right| < r^k(\widetilde{x}_v).$$

From Proposition 2 we conclude that $\widetilde{X}$ is a τ-stationary point of (6).

2) (Convergence) Since $\{f_v\}_{v\in\mathcal{V}}$ satisfy Assumption 3, we obtain that there exists $\overline{v} \in \mathcal{V}$ such that $f_{\overline{v}}$ satisfies the $\alpha_k^{\overline{v}}$-LSRHP with $\alpha_k^{\overline{v}} > 0$. From Lemma 3 and from Assumption 2 we have, for all $t \in \mathbb{N}$,

$$F(X(0)) \geq F(X(t)) \geq \sum_{v\in\mathcal{V}\setminus\{\overline{v}\}} \kappa_v + f_{\overline{v}}(x_{\overline{v}}(t))$$

then there exists $c = F(X(0)) - \sum_{v\in\mathcal{V}\setminus\{\overline{v}\}}\kappa_v$ such that for all $t \in \mathbb{N}$

$$\begin{aligned} c &\geq f_{\overline{v}}(x_{\overline{v}}(t)) \geq f_{\overline{v}}(0) + \langle\nabla f_{\overline{v}}(0), x_{\overline{v}}(t)\rangle + \frac{1}{2}\alpha_k^{\overline{v}}\|x_{\overline{v}}(t)\|^2 \\ &\geq f_{\overline{v}}(0) - \|\nabla f_{\overline{v}}(0)\|\|x_{\overline{v}}(t)\| + \frac{1}{2}\alpha_k^{\overline{v}}\|x_{\overline{v}}(t)\|^2. \end{aligned}$$

We deduce that the sequence $\{x_{\overline{v}}(t)\}_{t\in\mathbb{N}}$ is bounded.

In analogous way, for all $t \in \mathbb{N}$,

$$F(X(0)) - \sum_{v\in\mathcal{V}}\kappa_v \geq \frac{1}{2\tau d}\|x_{\overline{v}}(t) - x_w(t)\|, \quad \forall w \in \mathcal{N}_{\overline{v}}$$

hence also the sequences $\{x_w(t)\}_{t\in\mathbb{N}}$, $w \in \mathcal{N}_{\overline{v}}$, are bounded. Iterating the argument and by connectivity of the graph we conclude that $\{x_v(t)\}_{t\in\mathbb{N}}$ is bounded for all $v \in \mathcal{V}$. Finally, by LaSalle invariance principle [44], we conclude that $\{X(t)\}_{t\in\mathbb{N}}$ converges to a τ-stationary point.

Chiara Ravazzi (M'13) 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 (M'12) 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. Her main research interests are in the field of sparse signal processing, compressed sensing, and distributed systems.

Enrico Magli (S'97–M'01–SM'07) received the M.Sc. and Ph.D. degrees from Politecnico di Torino, Torino, Italy, in 1997 and 2001, respectively. He is currently an Associate Professor with Politecnico di Torino, Torino, Italy. His research interests are in the field of compressive sensing, image and video coding, and vision. He is an associate editor of the IEEE Transactions on circuits and systems for video technology, the IEEE Transactions on multimedia, and the EURASIP Journal on image and video processing, an IEEE Distinguished Lecturer for 2015-2016, and a corecipient of the IEEE Geoscience and Remote Sensing Society 2011 Transactions Prize Paper Award.