

Gossip Algorithms for Convex Consensus Optimization over Networks*

Jie Lu, Choon Yik Tang, and Paul R. Regier[†]
 School of Electrical and Computer Engineering
 University of Oklahoma
 Norman, OK 73019, USA
 {jie.lu-1,cytang,paulregier}@ou.edu

Travis D. Bow[†]
 Department of Mechanical Engineering
 Stanford University
 Stanford, CA 94305, USA
 tbow@stanford.edu

November 1, 2018

Abstract

In many applications, nodes in a network desire not only a consensus, but an optimal one. To date, a family of subgradient algorithms have been proposed to solve this problem under general convexity assumptions. This paper shows that, for the scalar case and by assuming a bit more, novel non-gradient-based algorithms with appealing features can be constructed. Specifically, we develop *Pairwise Equalizing* (PE) and *Pairwise Bisectioning* (PB), two gossip algorithms that solve unconstrained, separable, convex consensus optimization problems over undirected networks with time-varying topologies, where each local function is strictly convex, continuously differentiable, and has a minimizer. We show that PE and PB are easy to implement, bypass limitations of the subgradient algorithms, and produce switched, nonlinear, networked dynamical systems that admit a common Lyapunov function and asymptotically converge. Moreover, PE generalizes the well-known Pairwise Averaging and Randomized Gossip Algorithm, while PB relaxes a requirement of PE, allowing nodes to never share their local functions.

1 Introduction

Consider an N -node multi-hop network, where each node i observes a convex function f_i , and all the N nodes wish to determine an optimal consensus x^* , which minimizes the sum of the f_i 's:

$$x^* \in \arg \min_x \sum_{i=1}^N f_i(x). \quad (1)$$

*This work was supported by the National Science Foundation under grant CMMI-0900806.

[†]P. R. Regier and T. D. Bow were supported by the National Science Foundation Research Experiences for Undergraduates program under grant EEC-0755011.

Since each node i knows only its own f_i , the nodes cannot individually compute the optimal consensus x^* and, thus, must collaborate to do so. This problem of achieving unconstrained, separable, convex consensus optimization has many applications in multi-agent systems and wired/wireless/social networks, some examples of which can be found in [1, 2].

The current literature offers a large body of work on distributed consensus (see [3] for a survey), including a line of research that focuses on solving problem (1) for an optimal consensus x^* [1, 2, 4–17]. This line of work has resulted in a family of discrete-time subgradient algorithms, including the *incremental* subgradient algorithms [1, 2, 4–8, 10, 15], whereby an estimate of x^* is passed around the network, and the *non-incremental* ones [9, 11–14, 16, 17], whereby each node maintains an estimate of x^* and updates it iteratively by exchanging information with neighbors.

Although the aforementioned subgradient algorithms are capable of solving problem (1) under fairly weak assumptions, they suffer from one or more of the following limitations:

- L1. *Stepsizes*: The algorithms require selection of stepsizes, which may be constant, diminishing, or dynamic. In general, constant stepsizes ensure only convergence to neighborhoods of x^* , rather than to x^* itself. Moreover, they present an inevitable trade-off: larger stepsizes tend to yield larger convergence neighborhoods, while smaller ones tend to yield slower convergence. In contrast, diminishing stepsizes typically ensure asymptotic convergence. However, the convergence may be very slow, since the stepsizes may diminish too quickly. Finally, dynamic stepsizes allow shaping of the convergence behavior [4, 6]. Unfortunately, their dynamics depend on global information that is often costly to obtain. Hence, selecting appropriate stepsizes is not a trivial task, and inappropriate choices can cause poor performance.
- L2. *Hamiltonian cycle*: Many incremental subgradient algorithms [1, 2, 4–7, 10, 15] require the nodes to construct and maintain a Hamiltonian cycle (i.e., a closed path that visits every node exactly once) or a pseudo one (i.e., that allows multiple visits), which may be very difficult to carry out, especially in a decentralized, leaderless fashion.
- L3. *Multi-hop transmissions*: Some incremental subgradient algorithms [4–6] require the node that has the latest estimate of x^* to pass it on to a randomly and equiprobably chosen node in the network. This implies that every node must be aware of all the nodes in the network, and the algorithms must run alongside a routing protocol that enables such passing, which may not always be the case. The fact that the chosen node is typically multiple hops away also implies that these algorithms are communication inefficient, requiring plenty of transmissions (up to the network diameter) just to complete a single iteration.
- L4. *Lack of asymptotic convergence*: A variety of convergence properties have been established for the subgradient algorithms in [1, 2, 4–17], including error bounds, convergence in expectations, convergence in limit inferiors, convergence rates, etc. In contrast, relatively few asymptotic convergence results have been reported, except for the subgradient algorithms with diminishing or dynamic stepsizes in [4–6, 10, 15–17].

Limitations L1–L4 facing the subgradient algorithms raise the question of whether it is possible to devise algorithms, which require neither the notion of a stepsize, the construction of a (pseudo-)Hamiltonian cycle, nor the use of a routing protocol for multi-hop transmissions, and yet guarantee asymptotic convergence, bypassing L1–L4. In this paper, we show that, for the *one-dimensional* case and with a few mild assumptions, such algorithms can be constructed. Specifically, instead of letting the network be directed, we assume that it is undirected, with possibly a time-varying topology unknown to any of the nodes. In addition, instead of letting each f_i in (1) be convex but not necessarily differentiable, we assume that it is strictly convex, continuously differentiable, and has a minimizer. Based on these assumptions, we develop two gossip-style, distributed asynchronous iterative algorithms, referred to as *Pairwise Equalizing* (PE) and *Pairwise Bisectioning* (PB), which not only solve problem (1) and circumvent limitations L1–L4, but also are rather easy to implement—although computationally they are more demanding than the subgradient algorithms.

As will be shown in the paper, PE and PB exhibit a number of notable features. First, they produce switched, nonlinear, networked dynamical systems whose state evolves along an invariant manifold whenever nodes gossip with each other. The switched systems are proved, using Lyapunov stability theory, to be asymptotically convergent, as long as the gossiping pattern is sufficiently rich. In particular, we show that the first-order convexity condition can be used to form a common Lyapunov function, as well as to characterize drops in its value after every gossip. Second, PE and PB do not belong to the family of subgradient algorithms as they utilize fundamentally different, non-gradient-based update rules that involve no stepsize. These update rules are synthesized from two simple ideas—*conservation* and *dissipation*—which are somewhat similar to how Pairwise Averaging [18] was conceived back in the 1980s. Indeed, we show that PE reduces to Pairwise Averaging [18] and Randomized Gossip Algorithm [19] when problem (1) specializes to an averaging problem. Finally, PE requires one-time sharing of the f_i 's between gossiping nodes, which may be costly or impermissible in some applications. This requirement is eliminated by PB at the expense of more communications per iteration.

2 Problem Formulation

Consider a multi-hop network consisting of $N \geq 2$ nodes, connected by bidirectional links in a time-varying topology. The network is modeled as an undirected graph $\mathcal{G}(k) = (\mathcal{V}, \mathcal{E}(k))$, where $k \in \mathbb{N} = \{0, 1, 2, \dots\}$ denotes time, $\mathcal{V} = \{1, 2, \dots, N\}$ represents the set of N nodes, and $\mathcal{E}(k) \subset \{\{i, j\} : i, j \in \mathcal{V}, i \neq j\}$ represents the nonempty set of links at time k . Any two nodes $i, j \in \mathcal{V}$ are one-hop neighbors and can communicate at time $k \in \mathbb{N}$ if and only if $\{i, j\} \in \mathcal{E}(k)$.

Suppose, at time $k = 0$, each node $i \in \mathcal{V}$ observes a function $f_i : \mathcal{X} \rightarrow \mathbb{R}$, which maps a nonempty open interval $\mathcal{X} \subset \mathbb{R}$ to \mathbb{R} , and which satisfies the following assumption:

Assumption 1. For each $i \in \mathcal{V}$, the function f_i is strictly convex, continuously differentiable, and

has a minimizer $x_i^* \in \mathcal{X}$.

Suppose, upon observing the f_i 's, all the N nodes wish to solve the following unconstrained, separable, convex optimization problem:

$$\min_{x \in \mathcal{X}} F(x), \quad (2)$$

where the function $F : \mathcal{X} \rightarrow \mathbb{R}$ is defined as $F(x) = \sum_{i \in \mathcal{V}} f_i(x)$. Clearly, F is strictly convex and continuously differentiable. To show that F has a unique minimizer in \mathcal{X} so that problem (2) is well-posed, let $f'_i : \mathcal{X} \rightarrow \mathbb{R}$ and $F' : \mathcal{X} \rightarrow \mathbb{R}$ denote the derivatives of f_i and F , respectively, and consider the following lemma and proposition:

Lemma 1. *Let $g_i : \mathcal{X} \rightarrow \mathbb{R}$ be a strictly increasing and continuous function and $z_i \in \mathcal{X}$ for $i = 1, 2, \dots, n$. Then, there exists a unique $z \in \mathcal{X}$ such that $\sum_{i=1}^n g_i(z) = \sum_{i=1}^n g_i(z_i)$. Moreover, $z \in [\min_{i \in \{1, 2, \dots, n\}} z_i, \max_{i \in \{1, 2, \dots, n\}} z_i]$.*

Proof. Since g_i is strictly increasing and continuous $\forall i \in \{1, 2, \dots, n\}$, so is $\sum_{i=1}^n g_i : \mathcal{X} \rightarrow \mathbb{R}$. Thus, $\sum_{i=1}^n g_i(\min_{j \in \{1, 2, \dots, n\}} z_j) \leq \sum_{i=1}^n g_i(z_i) \leq \sum_{i=1}^n g_i(\max_{j \in \{1, 2, \dots, n\}} z_j)$. It follows from the Intermediate Value Theorem that there exists a unique $z \in \mathcal{X}$ such that $\sum_{i=1}^n g_i(z) = \sum_{i=1}^n g_i(z_i)$, and that $z \in [\min_{i \in \{1, 2, \dots, n\}} z_i, \max_{i \in \{1, 2, \dots, n\}} z_i]$. \square

Proposition 1. *With Assumption 1, there exists a unique $x^* \in \mathcal{X}$, which satisfies $F'(x^*) = 0$, minimizes F over \mathcal{X} , and solves problem (2), i.e., $x^* = \arg \min_{x \in \mathcal{X}} F(x)$.*

Proof. By Assumption 1, for every $i \in \mathcal{V}$, f'_i is strictly increasing and continuous. By Lemma 1, there exists a unique $x^* \in \mathcal{X}$ such that $\sum_{i \in \mathcal{V}} f'_i(x^*) = \sum_{i \in \mathcal{V}} f'_i(x_i^*)$. Since $F' = \sum_{i \in \mathcal{V}} f'_i$ and $f'_i(x_i^*) = 0 \forall i \in \mathcal{V}$, $F'(x^*) = 0$. Since F is strictly convex, x^* minimizes F over \mathcal{X} , solving (2). \square

Given the above, the goal is to construct a distributed asynchronous iterative algorithm free of limitations L1–L4, with which each node can asymptotically determine the unknown optimizer x^* .

3 Pairwise Equalizing

In this section, we develop a gossip algorithm having the aforementioned features.

Suppose, at time $k = 0$, each node $i \in \mathcal{V}$ creates a state variable $\hat{x}_i \in \mathcal{X}$ in its local memory, which represents its estimate of x^* . Also suppose, at each subsequent time $k \in \mathbb{P} = \{1, 2, \dots\}$, an iteration, called *iteration* k , takes place. Let $\hat{x}_i(0)$ represent the initial value of \hat{x}_i , and $\hat{x}_i(k)$ its value upon completing each iteration $k \in \mathbb{P}$. With this setup, the goal may be stated as

$$\lim_{k \rightarrow \infty} \hat{x}_i(k) = x^*, \quad \forall i \in \mathcal{V}. \quad (3)$$

To design an algorithm that guarantees (3), consider a *conservation condition*

$$\sum_{i \in \mathcal{V}} f'_i(\hat{x}_i(k)) = 0, \quad \forall k \in \mathbb{N}, \quad (4)$$

which says that the $\hat{x}_i(k)$'s evolve in a way that the sum of the derivatives f'_i 's, evaluated at the $\hat{x}_i(k)$'s, is always conserved at zero. Moreover, consider a *dissipation condition*

$$\lim_{k \rightarrow \infty} \hat{x}_i(k) = \tilde{x}, \quad \forall i \in \mathcal{V}, \text{ for some } \tilde{x} \in \mathcal{X}, \quad (5)$$

which says that the $\hat{x}_i(k)$'s gradually dissipate their differences and asymptotically achieve some arbitrary consensus $\tilde{x} \in \mathcal{X}$. Note that if (4) is met, then $\lim_{k \rightarrow \infty} \sum_{i \in \mathcal{V}} f'_i(\hat{x}_i(k)) = \lim_{k \rightarrow \infty} 0 = 0$. If, in addition, (5) is met, then due to the continuity of every f'_i , $\sum_{i \in \mathcal{V}} \lim_{k \rightarrow \infty} f'_i(\hat{x}_i(k)) = \sum_{i \in \mathcal{V}} f'_i(\lim_{k \rightarrow \infty} \hat{x}_i(k)) = \sum_{i \in \mathcal{V}} f'_i(\tilde{x}) = F'(\tilde{x})$. Because $\lim_{k \rightarrow \infty} f'_i(\hat{x}_i(k))$ exists for every $i \in \mathcal{V}$, $\lim_{k \rightarrow \infty} \sum_{i \in \mathcal{V}} f'_i(\hat{x}_i(k)) = \sum_{i \in \mathcal{V}} \lim_{k \rightarrow \infty} f'_i(\hat{x}_i(k))$. Combining the above, we obtain $F'(\tilde{x}) = 0$. From Proposition 1, we see that the arbitrary consensus \tilde{x} must be the unknown optimizer x^* , i.e., $\tilde{x} = x^*$, so that (3) holds. Therefore, to design an algorithm that ensures (3)—where x^* explicitly appears, it suffices to make the algorithm satisfy both the conservation and dissipation conditions (4) and (5)—where x^* is implicitly encoded.

To this end, observe that (4) holds if and only if the $\hat{x}_i(0)$'s are such that $\sum_{i \in \mathcal{V}} f'_i(\hat{x}_i(0)) = 0$, and the $\hat{x}_i(k)$'s are related to the $\hat{x}_i(k-1)$'s through

$$\sum_{i \in \mathcal{V}} f'_i(\hat{x}_i(k)) = \sum_{i \in \mathcal{V}} f'_i(\hat{x}_i(k-1)), \quad \forall k \in \mathbb{P}. \quad (6)$$

To satisfy $\sum_{i \in \mathcal{V}} f'_i(\hat{x}_i(0)) = 0$, it suffices that each node $i \in \mathcal{V}$ computes x_i^* on its own and sets

$$\hat{x}_i(0) = x_i^*, \quad \forall i \in \mathcal{V}, \quad (7)$$

since $f'_i(x_i^*) = 0$. To satisfy (6), consider a gossip algorithm, whereby at each iteration $k \in \mathbb{P}$, a pair $u(k) = \{u_1(k), u_2(k)\} \in \mathcal{E}(k)$ of one-hop neighbors $u_1(k)$ and $u_2(k)$ gossip and update their $\hat{x}_{u_1(k)}(k)$ and $\hat{x}_{u_2(k)}(k)$, while the rest of the N nodes stay idle, i.e.,

$$\hat{x}_i(k) = \hat{x}_i(k-1), \quad \forall k \in \mathbb{P}, \forall i \in \mathcal{V} - u(k). \quad (8)$$

With (8), equation (6) simplifies to

$$f'_{u_1(k)}(\hat{x}_{u_1(k)}(k)) + f'_{u_2(k)}(\hat{x}_{u_2(k)}(k)) = f'_{u_1(k)}(\hat{x}_{u_1(k)}(k-1)) + f'_{u_2(k)}(\hat{x}_{u_2(k)}(k-1)), \quad \forall k \in \mathbb{P}. \quad (9)$$

Hence, all that is needed for (6) to hold is a gossip between nodes $u_1(k)$ and $u_2(k)$ to share their $f_{u_1(k)}$, $f_{u_2(k)}$, $\hat{x}_{u_1(k)}(k-1)$, and/or $\hat{x}_{u_2(k)}(k-1)$, followed by a joint update of their $\hat{x}_{u_1(k)}(k)$ and $\hat{x}_{u_2(k)}(k)$, which ensures (9).

Obviously, (9) alone does not uniquely determine $\hat{x}_{u_1(k)}(k)$ and $\hat{x}_{u_2(k)}(k)$. This suggests that the available degree of freedom may be used to account for the dissipation condition (5). Unlike

the conservation condition (4), however, (5) is about where the $\hat{x}_i(k)$'s should approach as $k \rightarrow \infty$, which nodes $u_1(k)$ and $u_2(k)$ cannot guarantee themselves since they are only responsible for two of the N $\hat{x}_i(k)$'s. Nevertheless, given that all the N $\hat{x}_i(k)$'s should approach the *same* limit, nodes $u_1(k)$ and $u_2(k)$ can help make this happen by imposing an *equalizing condition*

$$\hat{x}_{u_1(k)}(k) = \hat{x}_{u_2(k)}(k), \quad \forall k \in \mathbb{P}. \quad (10)$$

With (10) added, there are now two equations with two variables, providing nodes $u_1(k)$ and $u_2(k)$ a chance to uniquely determine $\hat{x}_{u_1(k)}(k)$ and $\hat{x}_{u_2(k)}(k)$ from (9) and (10).

The following proposition asserts that (9) and (10) always have a unique solution, so that the evolution of the $\hat{x}_i(k)$'s is well-defined:

Proposition 2. *With Assumption 1 and (7)–(10), $\hat{x}_i(k) \forall k \in \mathbb{N} \forall i \in \mathcal{V}$ are well-defined, i.e., unambiguous and in \mathcal{X} . Moreover, $[\min_{i \in \mathcal{V}} \hat{x}_i(k), \max_{i \in \mathcal{V}} \hat{x}_i(k)] \subset [\min_{i \in \mathcal{V}} \hat{x}_i(k-1), \max_{i \in \mathcal{V}} \hat{x}_i(k-1)] \forall k \in \mathbb{P}$.*

Proof. By induction on $k \in \mathbb{N}$. By Assumption 1 and (7), $\hat{x}_i(0) \forall i \in \mathcal{V}$ are unambiguous and in \mathcal{X} . Next, let $k \in \mathbb{P}$ and suppose $\hat{x}_i(k-1) \forall i \in \mathcal{V}$ are unambiguous and in \mathcal{X} . We show that so are $\hat{x}_i(k) \forall i \in \mathcal{V}$. From (8), $\hat{x}_i(k) \forall i \in \mathcal{V} - u(k)$ are unambiguous and in \mathcal{X} . To show that so are $\hat{x}_{u_1(k)}(k)$ and $\hat{x}_{u_2(k)}(k)$, we show that (9) and (10) have a unique solution $(\hat{x}_{u_1(k)}(k), \hat{x}_{u_2(k)}(k)) \in \mathcal{X}^2$. By Lemma 1, there is a unique $z \in \mathcal{X}$ such that

$$f'_{u_1(k)}(z) + f'_{u_2(k)}(z) = f'_{u_1(k)}(\hat{x}_{u_1(k)}(k-1)) + f'_{u_2(k)}(\hat{x}_{u_2(k)}(k-1)), \quad (11)$$

which satisfies $z \in [\min_{i \in u(k)} \hat{x}_i(k-1), \max_{i \in u(k)} \hat{x}_i(k-1)]$. Setting $\hat{x}_{u_1(k)}(k) = \hat{x}_{u_2(k)}(k) = z$, we see that $(\hat{x}_{u_1(k)}(k), \hat{x}_{u_2(k)}(k))$ is a solution to (9) and (10), confirming the existence. Now let $(a_1, a_2) \in \mathcal{X}^2$ and $(b_1, b_2) \in \mathcal{X}^2$ be two solutions of (9) and (10). Then, due to (10), (9), and Lemma 1, we have $a_1 = a_2 = b_1 = b_2$, confirming the uniqueness. Therefore, $\hat{x}_i(k) \forall i \in \mathcal{V}$ are well-defined as desired. Finally, the second statement follows from (8) and the fact that $\hat{x}_{u_1(k)}(k) = \hat{x}_{u_2(k)}(k) \in [\min_{i \in u(k)} \hat{x}_i(k-1), \max_{i \in u(k)} \hat{x}_i(k-1)] \forall k \in \mathbb{P}$. \square

Proposition 2 calls for a few remarks. First, the interval $[\min_{i \in \mathcal{V}} \hat{x}_i(k), \max_{i \in \mathcal{V}} \hat{x}_i(k)]$ can only shrink or remain unchanged over time k . While this does not guarantee the dissipation condition (5), it shows that the $\hat{x}_i(k)$'s are “trying” to converge and are, at the very least, bounded even if \mathcal{X} is not. Second, the proofs of Proposition 2 and Lemma 1 suggest a simple, practical procedure for nodes $u_1(k)$ and $u_2(k)$ to solve (9) and (10) for $(\hat{x}_{u_1(k)}(k), \hat{x}_{u_2(k)}(k))$: apply a numerical *root-finding method*, such as the *bisection method* with initial bracket $[\min_{i \in u(k)} \hat{x}_i(k-1), \max_{i \in u(k)} \hat{x}_i(k-1)]$, to solve (11) for the unique z and then set $\hat{x}_{u_1(k)}(k) = \hat{x}_{u_2(k)}(k) = z$. Finally, since (11) always has a unique solution z , we can eliminate z and write

$$\hat{x}_{u_1(k)}(k) = \hat{x}_{u_2(k)}(k) = (f'_{u_1(k)} + f'_{u_2(k)})^{-1}(f'_{u_1(k)}(\hat{x}_{u_1(k)}(k-1)) + f'_{u_2(k)}(\hat{x}_{u_2(k)}(k-1))), \quad \forall k \in \mathbb{P}, \quad (12)$$

where $(f'_i + f'_j)^{-1} : (f'_i + f'_j)(\mathcal{X}) \rightarrow \mathcal{X}$ denotes the inverse of the injective function $f'_i + f'_j$ with its codomain restricted to its range.

Expressions (7), (8), and (12) collectively define a gossip-style, distributed asynchronous iterative algorithm that yields a switched, nonlinear, networked dynamical system

$$\hat{x}_i(k) = \begin{cases} (\sum_{j \in u(k)} f'_j)^{-1}(\sum_{j \in u(k)} f'_j(\hat{x}_j(k-1))), & \text{if } i \in u(k), \\ \hat{x}_i(k-1), & \text{otherwise,} \end{cases} \quad \forall k \in \mathbb{P}, \forall i \in \mathcal{V}, \quad (13)$$

with initial condition (7), and with $(u(k))_{k=1}^{\infty}$ representing the sequence of gossiping nodes that trigger the switchings. As this algorithm ensures the conservation condition (4), the state trajectory $(\hat{x}_1(k), \hat{x}_2(k), \dots, \hat{x}_N(k))$ must remain on an $(N-1)$ -dimensional manifold $\mathcal{M} = \{(x_1, x_2, \dots, x_N) \in \mathcal{X}^N : \sum_{i \in \mathcal{V}} f'_i(x_i) = 0\} \subset \mathcal{X}^N \subset \mathbb{R}^N$, making \mathcal{M} an invariant set. Given that the algorithm involves repeated, pairwise equalizing of the $\hat{x}_i(k)$'s, we refer to it as *Pairwise Equalizing* (PE). PE may be expressed in a compact algorithmic form as follows:

Algorithm 1 (Pairwise Equalizing).

Initialization:

1. Each node $i \in \mathcal{V}$ computes $x_i^* \in \mathcal{X}$, creates a variable $\hat{x}_i \in \mathcal{X}$, and sets $\hat{x}_i \leftarrow x_i^*$.

Operation: At each iteration:

2. A node with one or more one-hop neighbors, say, node i , initiates the iteration and selects a one-hop neighbor, say, node j , to gossip. Nodes i and j select one of two ways to gossip by labeling themselves as either nodes a and b , or nodes b and a , respectively, where $\{a, b\} = \{i, j\}$. If node b does not know f_a , node a transmits f_a to node b . Node a transmits \hat{x}_a to node b . Node b sets $\hat{x}_b \leftarrow (f'_a + f'_b)^{-1}(f'_a(\hat{x}_a) + f'_b(\hat{x}_b))$ and transmits \hat{x}_b to node a . Node a sets $\hat{x}_a \leftarrow \hat{x}_b$. ■

Due to space limitations, we omit remarks concerning the execution of Algorithm 1 and refer the reader to an earlier, conference version of this paper [20].

Notice that PE does not rely on a stepsize parameter to execute, nor does it require the construction of a (pseudo-)Hamiltonian cycle, as well as the concurrent use of a routing protocol for multi-hop transmissions. Indeed, all it essentially needs is that every node is capable of applying a root-finding method, maintaining a list of its one-hop neighbors, and remembering the functions it learns along the way. Therefore, PE overcomes limitations L1–L3, while being rather easy to implement—although computationally it is more demanding than the subgradient algorithms.

To show that PE asymptotically converges and, thus, circumvents L4, let $\mathbf{x}^* = (x^*, x^*, \dots, x^*)$ and $\mathbf{x}(k) = (\hat{x}_1(k), \hat{x}_2(k), \dots, \hat{x}_N(k))$. Then, from Propositions 1 and 2, $\mathbf{x}^* \in \mathcal{X}^N$ and $\mathbf{x}(k) \in \mathcal{X}^N \forall k \in \mathbb{N}$. In addition, due to (13), if $\mathbf{x}(k) = \mathbf{x}^*$ for some $k \in \mathbb{N}$, then $\mathbf{x}(\ell) = \mathbf{x}^* \forall \ell > k$. Hence, \mathbf{x}^* is an equilibrium point of the system (13). To show that $\lim_{k \rightarrow \infty} \mathbf{x}(k) = \mathbf{x}^*$, i.e., (3) holds, we seek to construct a Lyapunov function. To this end, recall that for any strictly convex and differentiable

function $f : \mathcal{X} \rightarrow \mathbb{R}$, the first-order convexity condition says that

$$f(y) \geq f(x) + f'(x)(y - x), \quad \forall x, y \in \mathcal{X}, \quad (14)$$

where the equality holds if and only if $x = y$. This suggests the following Lyapunov function candidate $V : \mathcal{X}^N \subset \mathbb{R}^N \rightarrow \mathbb{R}$, which exploits the convexity of the f_i 's:

$$V(\mathbf{x}(k)) = \sum_{i \in \mathcal{V}} f_i(x^*) - f_i(\hat{x}_i(k)) - f'_i(\hat{x}_i(k))(x^* - \hat{x}_i(k)). \quad (15)$$

Notice that V in (15) is well-defined. Moreover, due to Assumption 1 and (14), V is continuous and positive definite with respect to \mathbf{x}^* , i.e., $V(\mathbf{x}(k)) \geq 0 \forall \mathbf{x}(k) \in \mathcal{X}^N$, where the equality holds if and only if $\mathbf{x}(k) = \mathbf{x}^*$. Therefore, to prove (3), it suffices to show that

$$\lim_{k \rightarrow \infty} V(\mathbf{x}(k)) = 0. \quad (16)$$

The following lemma represents the first step toward establishing (16):

Lemma 2. *Consider the use of PE described in Algorithm 1. Suppose Assumption 1 holds. Then, for any given $(u(k))_{k=1}^{\infty}$, $(V(\mathbf{x}(k)))_{k=0}^{\infty}$ is non-increasing and satisfies*

$$V(\mathbf{x}(k)) - V(\mathbf{x}(k-1)) = - \sum_{i \in u(k)} f_i(\hat{x}_i(k)) - f_i(\hat{x}_i(k-1)) - f'_i(\hat{x}_i(k-1))(\hat{x}_i(k) - \hat{x}_i(k-1)), \quad \forall k \in \mathbb{P}. \quad (17)$$

Proof. Let $(u(k))_{k=1}^{\infty}$ be given. Then, from (15) and (13), we have $V(\mathbf{x}(k)) - V(\mathbf{x}(k-1)) = - \sum_{i \in u(k)} f_i(\hat{x}_i(k)) - f_i(\hat{x}_i(k-1)) + f'_i(\hat{x}_i(k))x^* - f'_i(\hat{x}_i(k-1))x^* - f'_i(\hat{x}_i(k))\hat{x}_i(k) + f'_i(\hat{x}_i(k-1))\hat{x}_i(k-1) \forall k \in \mathbb{P}$. Due to (13), $- \sum_{i \in u(k)} f'_i(\hat{x}_i(k))x^*$ cancels $\sum_{i \in u(k)} f'_i(\hat{x}_i(k-1))x^*$, while $\sum_{i \in u(k)} f'_i(\hat{x}_i(k))\hat{x}_i(k)$ becomes $\sum_{i \in u(k)} f'_i(\hat{x}_i(k-1))\hat{x}_i(k)$. This proves (17). Note that the right-hand side of (17) is nonpositive due to (14). Hence, $(V(\mathbf{x}(k)))_{k=0}^{\infty}$ is non-increasing. \square

Lemma 2 has several implications. First, upon completing each iteration $k \in \mathbb{P}$ by *any* two nodes $u_1(k)$ and $u_2(k)$, the value of V must either decrease or, at worst, stay the same, where the latter occurs if and only if $\hat{x}_{u_1(k)}(k-1) = \hat{x}_{u_2(k)}(k-1)$. Second, since $(V(\mathbf{x}(k)))_{k=0}^{\infty}$ is non-increasing irrespective of $(u(k))_{k=1}^{\infty}$, V in (15) may be regarded as a *common* Lyapunov function for the nonlinear switched system (13), which has as many as $\frac{N(N-1)}{2}$ different dynamics, corresponding to the $\frac{N(N-1)}{2}$ possible gossiping pairs. Finally, the first-order convexity condition (14) can be used not only to form the common Lyapunov function V , but also to characterize drops in its value in (17) after every gossip. This is akin to how quadratic functions may be used to form a common Lyapunov function $V(k) = x^T(k)Px(k)$ for a linear switched system $x(k+1) = A(k)x(k)$, $A(k) \in \{A_1, A_2, \dots, A_M\}$, as well as to characterize drops in $V(k)$ via $V(k+1) - V(k) = x^T(k)(A_i^T P A_i - P)x(k) = -x^T(k)Q_i x(k)$. Indeed, as we will show later, when problem (2) specializes to an averaging problem, where the nonlinear switched system (13) becomes linear, both V and its drop become quadratic functions.

As $(V(\mathbf{x}(k)))_{k=0}^{\infty}$ is nonnegative and non-increasing, $\lim_{k \rightarrow \infty} V(\mathbf{x}(k))$ exists and is nonnegative. This, however, is insufficient for us to conclude that $\lim_{k \rightarrow \infty} V(\mathbf{x}(k)) = 0$, since, for some pathological gossiping patterns, $\lim_{k \rightarrow \infty} V(\mathbf{x}(k))$ can be positive (see [20] for examples). Thus, some restrictions must be imposed on the gossiping pattern, in order to establish (16). To this end, let $\mathcal{E}_{\infty} = \{\{i, j\} : u(k) = \{i, j\} \text{ for infinitely many } k \in \mathbb{P}\}$, so that a link $\{i, j\}$ is in \mathcal{E}_{∞} if and only if nodes i and j gossip with each other infinitely often. Then, we may state the following restriction on the gossiping pattern, which was first adopted in [18] and is not difficult to satisfy in practice [20]:

Assumption 2. The sequence $(u(k))_{k=1}^{\infty}$ is such that the graph $(\mathcal{V}, \mathcal{E}_{\infty})$ is connected.

The following theorem says that, under Assumption 2 on the gossiping pattern, PE ensures asymptotic convergence of all the $\hat{x}_i(k)$'s to x^* , circumventing limitation L4:

Theorem 1. *Consider the use of PE described in Algorithm 1. Suppose Assumptions 1 and 2 hold. Then, (16) and (3) hold.*

Proof. See Appendix A.1. □

Finally, we point out that the above results may be viewed as a natural generalization of some known results in distributed averaging. Consider a special case where each node $i \in \mathcal{V}$ observes not an arbitrary function f_i , but a quadratic one of the form $f_i(x) = \frac{1}{2}(x - y_i)^2 + c_i$ with domain $\mathcal{X} = \mathbb{R}$ and parameters $y_i, c_i \in \mathbb{R}$. In this case, finding the unknown optimizer x^* amounts to calculating the network-wide average $\frac{1}{N} \sum_{i \in \mathcal{V}} y_i$ of the node “observations” y_i 's, so that the convex optimization problem (2) becomes an averaging problem. In addition, initializing the node estimates $\hat{x}_i(0)$'s simply means setting them to the y_i 's, and equalizing $\hat{x}_{u_1(k)}(k)$ and $\hat{x}_{u_2(k)}(k)$ simply means averaging them, so that PE reduces to Pairwise Averaging [18] and Randomized Gossip Algorithm [19]. Moreover, the invariant manifold \mathcal{M} becomes the invariant hyperplane $\mathcal{M} = \{(x_1, x_2, \dots, x_N) \in \mathbb{R}^N : \sum_{i \in \mathcal{V}} x_i = \sum_{i \in \mathcal{V}} y_i\}$ in distributed averaging. Furthermore, both the common Lyapunov function V in (15) and its drop in (17) take a quadratic form: $V(\mathbf{x}(k)) = \frac{1}{2}(\mathbf{x}(k) - \mathbf{x}^*)^T(\mathbf{x}(k) - \mathbf{x}^*)$ and $V(\mathbf{x}(k)) - V(\mathbf{x}(k-1)) = -\frac{1}{2}\mathbf{x}^T(k-1)Q_{u(k)}\mathbf{x}(k-1) \forall k \in \mathbb{P}$, where $Q_{\{i,j\}} \in \mathbb{R}^{N \times N}$ is a symmetric positive semidefinite matrix whose ii and jj entries are $\frac{1}{2}$, ij and ji entries are $-\frac{1}{2}$, and all other entries are zero. Therefore, the first-order-convexity-condition-based Lyapunov function (15) generalizes the quadratic Lyapunov function in distributed averaging.

4 Pairwise Bisectioning

Although PE solves problem (2) and bypasses L1–L4, it requires one-time, one-way sharing of the f_i 's between gossiping nodes, which may be costly for certain f_i 's, or impermissible for security and privacy reasons. In this section, we develop another gossip algorithm that eliminates this requirement at the expense of more real-number transmissions per iteration.

Note that PE can be traced back to four defining equations (7)–(10), and that its drawback of having to share the f_i 's stems from having to solve (9) and (10). To overcome this drawback, consider a gossip algorithm satisfying (7)–(9) and a new condition but not (10). Assuming, without loss of generality, that $\hat{x}_{u_1(k)}(k-1) \leq \hat{x}_{u_2(k)}(k-1) \forall k \in \mathbb{P}$, this new condition can be stated as

$$\hat{x}_{u_1(k)}(k-1) \leq \hat{x}_{u_1(k)}(k) \leq \hat{x}_{u_2(k)}(k) \leq \hat{x}_{u_2(k)}(k-1), \quad \forall k \in \mathbb{P}. \quad (18)$$

Termed as the *approaching condition*, (18) says that at each iteration $k \in \mathbb{P}$, nodes $u_1(k)$ and $u_2(k)$ force $\hat{x}_{u_1(k)}(k)$ and $\hat{x}_{u_2(k)}(k)$ to approach each other while preserving their order. Observe that the approaching condition (18) includes the equalizing condition (10) as a special case. Furthermore, unlike (9) and (10), (9) and (18) do not uniquely determine $\hat{x}_{u_1(k)}(k)$ and $\hat{x}_{u_2(k)}(k)$. Rather, they allow $\hat{x}_{u_1(k)}(k)$ and $\hat{x}_{u_2(k)}(k)$ to increase gradually from $\hat{x}_{u_1(k)}(k-1)$ and decrease accordingly from $\hat{x}_{u_2(k)}(k-1)$, respectively, until the two become equal.

The following lemma characterizes the impact of the non-uniqueness on the value of V :

Lemma 3. *Consider (7)–(9) and (18). Suppose Assumption 1 holds. Then, for any given $(u(k))_{k=1}^\infty$, $(V(\mathbf{x}(k)))_{k=0}^\infty$ is non-increasing. Moreover, for any given $k \in \mathbb{P}$ and $\mathbf{x}(k-1) \in \mathcal{X}^N$, $V(\mathbf{x}(k))$ strictly increases with $\hat{x}_{u_2(k)}(k) - \hat{x}_{u_1(k)}(k)$ over $[0, \hat{x}_{u_2(k)}(k-1) - \hat{x}_{u_1(k)}(k-1)]$.*

Proof. Let $(u(k))_{k=1}^\infty$ be given. Then, from (15), (8), and (9), we have $V(\mathbf{x}(k)) - V(\mathbf{x}(k-1)) = -\sum_{i \in u(k)} f_i(\hat{x}_i(k)) - f_i(\hat{x}_i(k-1)) - f'_i(\hat{x}_i(k-1))(\hat{x}_i(k) - \hat{x}_i(k-1)) + (f'_i(\hat{x}_i(k-1)) - f'_i(\hat{x}_i(k)))\hat{x}_i(k) \forall k \in \mathbb{P}$. Due to (9) and (18), $\sum_{i \in u(k)} (f'_i(\hat{x}_i(k-1)) - f'_i(\hat{x}_i(k)))\hat{x}_i(k) = (f'_{u_1(k)}(\hat{x}_{u_1(k)}(k-1)) - f'_{u_1(k)}(\hat{x}_{u_1(k)}(k)))(\hat{x}_{u_1(k)}(k) - \hat{x}_{u_2(k)}(k)) \geq 0$. This, along with (14), implies $V(\mathbf{x}(k)) - V(\mathbf{x}(k-1)) \leq 0 \forall k \in \mathbb{P}$. Now let $k \in \mathbb{P}$ and $\mathbf{x}(k-1) \in \mathcal{X}^N$ be given. By Lemma 1, there exists a unique $x_{\text{eq}} \in \mathcal{X}$ such that $\sum_{i \in u(k)} f'_i(x_{\text{eq}}) = \sum_{i \in u(k)} f'_i(\hat{x}_i(k))$. Also, $x_{\text{eq}} \in [\hat{x}_{u_1(k)}(k), \hat{x}_{u_2(k)}(k)]$. Let $\mathbf{x}_{\text{eq}} \in \mathcal{X}^N$ be such that its i th entry is x_{eq} if $i \in u(k)$ and $\hat{x}_i(k-1)$ otherwise. Then, it follows from (15), (8), and (14) that $V(\mathbf{x}(k)) - V(\mathbf{x}_{\text{eq}}) = \sum_{i \in u(k)} f_i(x_{\text{eq}}) - f_i(\hat{x}_i(k)) - f'_i(\hat{x}_i(k))(x_{\text{eq}} - \hat{x}_i(k)) \geq 0$. Because $f_i(y) - f_i(x) - f'_i(x)(y-x)$ strictly increases with $|y-x|$ for each fixed $y \in \mathcal{X} \forall i \in \mathcal{V}$ and because of (9) and (18), the second claim is true. \square

Lemma 3 says that the value of V can never increase. In addition, the closer $\hat{x}_{u_1(k)}(k)$ and $\hat{x}_{u_2(k)}(k)$ get, the larger the value of V drops, and the drop is maximized when $\hat{x}_{u_1(k)}(k)$ and $\hat{x}_{u_2(k)}(k)$ are equalized. These observations suggest that perhaps it is possible to design an algorithm that only forces $\hat{x}_{u_1(k)}(k)$ and $\hat{x}_{u_2(k)}(k)$ to approach each other (as opposed to becoming equal) to the detriment of a smaller drop in the value of V , but at the benefit of not having to share the f_i 's. The following algorithm, referred to as *Pairwise Bisectioning* (PB), shows that this is indeed the case and utilizes a bisection step that allows $\hat{x}_{u_1(k)}(k)$ and $\hat{x}_{u_2(k)}(k)$ to get arbitrarily close:

Algorithm 2 (Pairwise Bisectioning).

Initialization:

1. Each node $i \in \mathcal{V}$ computes $x_i^* \in \mathcal{X}$, creates variables $\hat{x}_i, a_i, b_i \in \mathcal{X}$, and sets $\hat{x}_i \leftarrow x_i^*$.

Operation: At each iteration:

2. A node with one or more one-hop neighbors, say, node i , initiates the iteration and selects a one-hop neighbor, say, node j , to gossip. Node i transmits \hat{x}_i to node j . Node j sets $a_j \leftarrow \min\{\hat{x}_i, \hat{x}_j\}$ and $b_j \leftarrow \max\{\hat{x}_i, \hat{x}_j\}$ and transmits \hat{x}_j to node i . Node i sets $a_i \leftarrow \min\{\hat{x}_i, \hat{x}_j\}$ and $b_i \leftarrow \max\{\hat{x}_i, \hat{x}_j\}$. Nodes i and j select the number of bisection rounds $R \in \mathbb{P}$.
3. Repeat the following R times: Node j transmits $f'_j(\frac{a_j+b_j}{2}) - f'_j(\hat{x}_j)$ to node i . Node i tests if $f'_j(\frac{a_j+b_j}{2}) - f'_j(\hat{x}_j) + f'_i(\frac{a_i+b_i}{2}) - f'_i(\hat{x}_i) \geq 0$. If so, node i sets $b_i \leftarrow \frac{a_i+b_i}{2}$ and transmits LEFT to node j , and node j sets $b_j \leftarrow \frac{a_j+b_j}{2}$. Otherwise, node i sets $a_i \leftarrow \frac{a_i+b_i}{2}$ and transmits RIGHT to node j , and node j sets $a_j \leftarrow \frac{a_j+b_j}{2}$. End repeat.
4. Node j transmits $f'_j(c_j) - f'_j(\hat{x}_j)$ to node i , where $c_j = \begin{cases} a_j & \text{if } \hat{x}_j \leq a_j \\ b_j & \text{if } \hat{x}_j \geq b_j \end{cases}$. Node i tests if $(f'_j(c_j) - f'_j(\hat{x}_j) + f'_i(c_i) - f'_i(\hat{x}_i))(\hat{x}_i - \frac{a_i+b_i}{2}) \geq 0$, where $c_i = \begin{cases} a_i & \text{if } \hat{x}_i \leq a_i \\ b_i & \text{if } \hat{x}_i \geq b_i \end{cases}$. If so, node i sets $\hat{x}_i \leftarrow (f'_i)^{-1}(f'_i(\hat{x}_i) - f'_j(c_j) + f'_j(\hat{x}_j))$ and node j sets $\hat{x}_j \leftarrow c_j$. Otherwise, node i transmits $f'_i(c_i) - f'_i(\hat{x}_i)$ to node j and sets $\hat{x}_i \leftarrow c_i$, and node j sets $\hat{x}_j \leftarrow (f'_j)^{-1}(f'_j(\hat{x}_j) - f'_i(c_i) + f'_i(\hat{x}_i))$. ■

Notice that Step 1 of PB is identical to that of PE except that each node $i \in \mathcal{V}$ creates two additional variables, a_i and b_i , which are used in Step 2 to represent the initial bracket $[a_i, b_i] = [a_j, b_j] = [\min\{\hat{x}_i, \hat{x}_j\}, \max\{\hat{x}_i, \hat{x}_j\}]$ for bisection purposes. Step 3 describes execution of the bisection method, where $R \in \mathbb{P}$ denotes the number of bisection rounds, which may be different for each iteration (e.g., a large R may be advisable when \hat{x}_i and \hat{x}_j are very different). Observe that upon completing Step 3, $x_{\text{eq}} \in [a_i, b_i] = [a_j, b_j] \subset [\min\{\hat{x}_i, \hat{x}_j\}, \max\{\hat{x}_i, \hat{x}_j\}]$ and $b_i - a_i = b_j - a_j = \frac{1}{2^R}|\hat{x}_j - \hat{x}_i|$, where x_{eq} denotes the equalized value of \hat{x}_i and \hat{x}_j if PE were used. Moreover, upon completing Step 4, $x_{\text{eq}} \in [\min\{\hat{x}_i, \hat{x}_j\}, \max\{\hat{x}_i, \hat{x}_j\}] \subset [a_i, b_i] = [a_j, b_j]$, where \hat{x}_i and \hat{x}_j here represent new values. Therefore, upon completing each iteration $k \in \mathbb{P}$,

$$|\hat{x}_{u_1(k)}(k) - \hat{x}_{u_2(k)}(k)| \leq \frac{1}{2^R} |\hat{x}_{u_1(k)}(k-1) - \hat{x}_{u_2(k)}(k-1)|, \quad \forall k \in \mathbb{P}. \quad (19)$$

Finally, note that unlike PE which requires two real-number transmissions per iteration, PB requires as many as $3 + R$ or $4 + R$. However, it allows the nodes to never share their f_i 's.

The following theorem establishes the asymptotic convergence of PB under Assumption 2:

Theorem 2. *Consider the use of PB described in Algorithm 2. Suppose Assumptions 1 and 2 hold. Then, (16) and (3) hold.*

Proof. See Appendix A.2. □

As it follows from the above, PB represents an alternative to PE, which is useful when nodes are either unable, or unwilling, to share their f_i 's. Although not pursued here, it is straightforward to see that PE and PB may be combined, so that equalizing is used when one of the gossiping nodes can send the other its f_i , and approaching is used when none of them can.

5 Conclusion

In this paper, based on the ideas of conservation and dissipation, we have developed PE and PB, two non-gradient-based gossip algorithms that enable nodes to cooperatively solve a class of convex optimization problems over networks. Using Lyapunov stability theory and the convexity structure, we have shown that PE and PB are asymptotically convergent, provided that the gossiping pattern is sufficiently rich. We have also discussed several salient features of PE and PB, including their comparison with the subgradient algorithms and their connection with distributed averaging.

A Appendix

A.1 Proof of Theorem 1

Suppose Assumption 1 holds and let $(u(k))_{k=1}^{\infty}$ satisfying Assumption 2 be given. Consider the following lemmas:

Lemma 4. *Suppose Assumption 1 holds. Then, $\forall [a, b] \subset \mathcal{X}$, there exists a continuous and strictly increasing function $\gamma : [0, \infty) \rightarrow [0, \infty)$ satisfying $\gamma(0) = 0$ and $\lim_{d \rightarrow \infty} \gamma(d) = \infty$, such that $\forall \eta > 0$, $\forall i \in \mathcal{V}$, $\forall (x, y) \in [a, b]^2$, $f_i(y) - f_i(x) - f'_i(x)(y - x) \leq \eta$ implies $|y - x| \leq \gamma^{-1}(\eta)$.*

Proof. Let $[a, b] \subset \mathcal{X}$. For each $i \in \mathcal{V}$, define $g_i : [a, b]^2 \rightarrow \mathbb{R}$ as $g_i(x, y) = f_i(y) - f_i(x) - f'_i(x)(y - x)$. Due to Assumption 1 and (14), $g_i(x, y) \geq 0 \forall (x, y) \in [a, b]^2$, where the equality holds if and only if $x = y$. Moreover, since f'_i is strictly increasing and $g_i(x, y)$ can be written as $g_i(x, y) = \int_x^y (f'_i(t) - f'_i(x)) dt$, $g_i(x, y)$ is strictly increasing with $|y - x|$ for each fixed $x \in [a, b]$. Furthermore, because f_i and f'_i are continuous, g_i is continuous. Next, for each $d \in [0, b - a]$, let $\mathcal{K}(d) = \{(x, y) \in [a, b]^2 : |y - x| = d\}$. Also, for each $i \in \mathcal{V}$, define $\gamma_i : [0, b - a] \rightarrow \mathbb{R}$ as $\gamma_i(d) = \min_{(x, y) \in \mathcal{K}(d)} g_i(x, y)$. Due to the compactness of $\mathcal{K}(d) \forall d \in [0, b - a]$ and the continuity of g_i , γ_i is well-defined and continuous. In addition, since $g_i(x, y) = 0 \forall (x, y) \in \mathcal{K}(0)$, $\gamma_i(0) = 0$. Now pick any d_1 and d_2 such that $0 \leq d_1 < d_2 \leq b - a$. Let $(x_2, y_2) \in \mathcal{K}(d_2)$ be such that $\gamma_i(d_2) = g_i(x_2, y_2)$. If $y_2 > x_2$, then $y_2 - x_2 = d_2$. In this case, $\exists y_1 \in [x_2, y_2)$ such that $y_1 - x_2 = d_1$. Since $g_i(x_2, y)$ is strictly increasing with y for $y \geq x_2$, we have $\gamma_i(d_1) \leq g_i(x_2, y_1) < g_i(x_2, y_2) = \gamma_i(d_2)$. Similarly, if $y_2 < x_2$, we also have $\gamma_i(d_1) < \gamma_i(d_2)$. Hence, γ_i is strictly increasing. Finally, define $\gamma : [0, \infty) \rightarrow [0, \infty)$ as $\gamma(d) = \begin{cases} \min_{i \in \mathcal{V}} \gamma_i(d) & \text{if } d \in [0, b - a] \\ \min_{i \in \mathcal{V}} \gamma_i(b - a) + d - (b - a) & \text{if } d \in (b - a, \infty) \end{cases}$. Note that $\gamma(0) = 0$ since $\gamma_i(0) = 0 \forall i \in \mathcal{V}$, and that $\lim_{d \rightarrow \infty} \gamma(d) = \infty$. Moreover, since γ_i is continuous and strictly increasing $\forall i \in \mathcal{V}$, so is γ on $[0, b - a]$. Also, observe that γ is continuous and strictly increasing on $[b - a, \infty)$. Thus, γ is continuous and strictly increasing. Now let $\eta > 0$, $i \in \mathcal{V}$, and $(x, y) \in [a, b]^2$. Suppose $g_i(x, y) \leq \eta$. If $\eta \leq \gamma(b - a)$, then $|y - x| \leq \gamma^{-1}(\eta)$ because $\gamma(|y - x|) \leq \gamma_i(|y - x|) \leq g_i(x, y) \leq \eta$. If $\eta > \gamma(b - a)$, then $|y - x| \leq b - a < \gamma^{-1}(\eta)$. \square

Lemma 5. *Suppose Assumption 1 holds. Then, $\forall [a, b] \subset \mathcal{X}$, $\exists \beta \in (0, \infty)$ such that $\forall i \in \mathcal{V}$, $\forall (x, y) \in [a, b]^2$, $f_i(y) - f_i(x) - f'_i(x)(y - x) \leq \beta|y - x|$.*

Proof. Let $[a, b] \subset \mathcal{X}$ and $\beta = 1 + 2 \max_{j \in \mathcal{V}} |f'_j(b)|$. Obviously, $\beta > 0$, and by Assumption 1, $\beta < \infty$. Let $i \in \mathcal{V}$ and $(x, y) \in [a, b]^2$. Since f_i is continuously differentiable, by the Mean Value Theorem, $\exists c$ between x and y such that $f_i(y) - f_i(x) = f'_i(c)(y - x)$. This, along with the triangle inequality and the fact that f'_i is strictly increasing, implies that $f_i(y) - f_i(x) - f'_i(x)(y - x) = (f'_i(c) - f'_i(x))(y - x) \leq |f'_i(c) - f'_i(x)| \cdot |y - x| \leq (|f'_i(c)| + |f'_i(x)|)|y - x| \leq 2|f'_i(b)| \cdot |y - x| \leq \beta|y - x|$. \square

Let $a = \min_{i \in \mathcal{V}} \hat{x}_i(0)$ and $b = \max_{i \in \mathcal{V}} \hat{x}_i(0)$. Then, it follows from Proposition 2 that $\hat{x}_i(k) \in [a, b] \subset \mathcal{X} \forall k \in \mathbb{N} \forall i \in \mathcal{V}$ and from (4) and Lemma 1 that $x^* \in [a, b]$. By Lemma 4, there exists a continuous and strictly increasing function $\gamma : [0, \infty) \rightarrow [0, \infty)$ satisfying $\gamma(0) = 0$ and $\lim_{d \rightarrow \infty} \gamma(d) = \infty$, such that $\forall \eta > 0$, $\forall i \in \mathcal{V}$, $\forall (x, y) \in [a, b]^2$, $f_i(y) - f_i(x) - f'_i(x)(y - x) \leq \eta$ implies $|y - x| \leq \gamma^{-1}(\eta)$. Also, by Lemma 5, $\exists \beta \in (0, \infty)$ such that $\forall i \in \mathcal{V}$, $\forall (x, y) \in [a, b]^2$, $f_i(y) - f_i(x) - f'_i(x)(y - x) \leq \beta|y - x|$. From Lemma 2, $(V(\mathbf{x}(k)))_{k=0}^\infty$ is nonnegative and non-increasing. Thus, $\exists c \geq 0$ such that $\lim_{k \rightarrow \infty} V(\mathbf{x}(k)) = c$. To show that c must be zero, assume, to the contrary, that $c > 0$. Let $\epsilon > 0$ be given by $\epsilon = \gamma(\frac{c}{4\beta N^2})$. Then, $\exists k_1 \in \mathbb{N}$ such that

$$c \leq V(\mathbf{x}(k)) < c + \epsilon, \quad \forall k \geq k_1. \quad (20)$$

Due to (20), $V(\mathbf{x}(k-1)) - V(\mathbf{x}(k)) < \epsilon \forall k \geq k_1 + 1$. Hence, from (14) and (17), $f_i(\hat{x}_i(k)) - f_i(\hat{x}_i(k-1)) - f'_i(\hat{x}_i(k-1))(\hat{x}_i(k) - \hat{x}_i(k-1)) < \epsilon \forall k \geq k_1 + 1 \forall i \in u(k)$. As a result, $|\hat{x}_i(k) - \hat{x}_i(k-1)| \leq \gamma^{-1}(\epsilon) \forall k \geq k_1 + 1 \forall i \in u(k)$. Because of this and (10),

$$|\hat{x}_i(k) - \hat{x}_j(k)| \leq 2\gamma^{-1}(\epsilon), \quad \forall k \geq k_1, \forall i, j \in u(k+1). \quad (21)$$

Now suppose $\max_{i \in \mathcal{V}} \hat{x}_i(k_1) - \min_{i \in \mathcal{V}} \hat{x}_i(k_1) > 2(N-1)\gamma^{-1}(\epsilon)$. Then, $\exists p, q \in \mathcal{V}$ such that $\hat{x}_q(k_1) - \hat{x}_p(k_1) > 2\gamma^{-1}(\epsilon)$ and $\mathcal{C}_1 \cup \mathcal{C}_2 = \mathcal{V}$, where $\mathcal{C}_1 = \{i \in \mathcal{V} : \hat{x}_i(k_1) \leq \hat{x}_p(k_1)\}$ and $\mathcal{C}_2 = \{i \in \mathcal{V} : \hat{x}_i(k_1) \geq \hat{x}_q(k_1)\}$. Next, we show by induction that $\forall k \geq k_1$, $\hat{x}_i(k) \leq \hat{x}_p(k_1) \forall i \in \mathcal{C}_1$ and $\hat{x}_i(k) \geq \hat{x}_q(k_1) \forall i \in \mathcal{C}_2$. Clearly, the statement is true for $k = k_1$. For $k \geq k_1 + 1$, suppose $\hat{x}_i(k-1) \leq \hat{x}_p(k_1) \forall i \in \mathcal{C}_1$ and $\hat{x}_i(k-1) \geq \hat{x}_q(k_1) \forall i \in \mathcal{C}_2$. Then, due to (21), $\forall i \in \mathcal{C}_1, \forall j \in \mathcal{C}_2, \{i, j\} \neq u(k)$, i.e., $u(k) \subset \mathcal{C}_1$ or $u(k) \subset \mathcal{C}_2$. It follows from (13) and Lemma 1 that $\hat{x}_i(k) \leq \hat{x}_p(k_1) \forall i \in \mathcal{C}_1$ and $\hat{x}_i(k) \geq \hat{x}_q(k_1) \forall i \in \mathcal{C}_2$, completing the induction. Due again to (21), we have $\forall i \in \mathcal{C}_1, \forall j \in \mathcal{C}_2, \{i, j\} \neq u(k) \forall k \geq k_1 + 1$, which violates Assumption 2. Consequently, $\max_{i \in \mathcal{V}} \hat{x}_i(k_1) - \min_{i \in \mathcal{V}} \hat{x}_i(k_1) \leq 2(N-1)\gamma^{-1}(\epsilon)$. It follows from (4) and Lemma 1 that $|x^* - \hat{x}_i(k_1)| \leq \max_{j \in \mathcal{V}} \hat{x}_j(k_1) - \min_{j \in \mathcal{V}} \hat{x}_j(k_1) \leq 2(N-1)\gamma^{-1}(\epsilon) \forall i \in \mathcal{V}$. Hence, $V(\mathbf{x}(k_1)) \leq \beta \sum_{i \in \mathcal{V}} |x^* - \hat{x}_i(k_1)| \leq \beta \cdot N \cdot 2(N-1)\gamma^{-1}(\epsilon) < c$, which contradicts (20). Therefore, $c = 0$, i.e., (16) holds, implying that (3) is satisfied.

A.2 Proof of Theorem 2

The proof is similar to that of Theorem 1. Let a, b, γ , and β be as defined in Appendix A.1. Then, due to (8), (18), (4), and Lemma 1, we have $\hat{x}_i(k) \in [a, b] \forall k \in \mathbb{N} \forall i \in \mathcal{V}$ and $x^* \in [a, b]$. From

Lemma 3, $\lim_{k \rightarrow \infty} V(\mathbf{x}(k)) = c$ for some $c \geq 0$. To show that $c = 0$, assume to the contrary that $c > 0$ and let ϵ be as defined in A.1. Then, (20) holds for some $k_1 \in \mathbb{N}$. It follows from the proof of Lemma 3 that $f_i(\hat{x}_i(k)) - f_i(\hat{x}_i(k-1)) - f'_i(\hat{x}_i(k-1))(\hat{x}_i(k) - \hat{x}_i(k-1)) \leq V(\mathbf{x}(k-1)) - V(\mathbf{x}(k)) < \epsilon \forall k \geq k_1 + 1 \forall i \in u(k)$. Thus, $|\hat{x}_i(k) - \hat{x}_i(k-1)| \leq \gamma^{-1}(\epsilon) \forall k \geq k_1 + 1 \forall i \in u(k)$. This, along with (19) and the fact that $R \in \mathbb{P}$, implies $|\hat{x}_i(k) - \hat{x}_j(k)| \leq \frac{2\gamma^{-1}(\epsilon)}{1 - \frac{1}{2R}} \leq 4\gamma^{-1}(\epsilon) \forall k \geq k_1 \forall i, j \in u(k+1)$. Then, using the same idea as in A.1, it can be shown that $\max_{i \in \mathcal{V}} \hat{x}_i(k_1) - \min_{i \in \mathcal{V}} \hat{x}_i(k_1) \leq 4(N-1)\gamma^{-1}(\epsilon)$. This leads to $V(\mathbf{x}(k_1)) < c$, which contradicts (20). Therefore, (16) and (3) hold.

References

- [1] S.-H. Son, M. Chiang, S. R. Kulkarni, and S. C. Schwartz, "The value of clustering in distributed estimation for sensor networks," in *Proc. International Conference on Wireless Networks, Communications and Mobile Computing*, Maui, HI, 2005, pp. 969–974.
- [2] M. G. Rabbat and R. D. Nowak, "Distributed optimization in sensor networks," in *Proc. International Symposium on Information Processing in Sensor Networks*, Berkeley, CA, 2004, pp. 20–27.
- [3] R. Olfati-Saber, J. A. Fax, and R. M. Murray, "Consensus and cooperation in networked multi-agent systems," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 215–233, 2007.
- [4] A. Nedić and D. P. Bertsekas, "Incremental subgradient methods for nondifferentiable optimization," *SIAM Journal on Optimization*, vol. 12, no. 1, pp. 109–138, 2001.
- [5] A. Nedić, D. P. Bertsekas, and V. S. Borkar, "Distributed asynchronous incremental subgradient methods," in *Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications*, D. Butnariu, Y. Censor, and S. Reich, Eds. Amsterdam, Holland: Elsevier, 2001, pp. 381–407.
- [6] A. Nedić and D. P. Bertsekas, "Convergence rate of incremental subgradient algorithms," in *Stochastic Optimization: Algorithms and Applications*, S. P. Uryasev and P. M. Pardalos, Eds. Norwell, MA: Kluwer Academic Publishers, 2001, pp. 223–264.
- [7] M. G. Rabbat and R. D. Nowak, "Quantized incremental algorithms for distributed optimization," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 4, pp. 798–808, 2005.
- [8] B. Johansson, M. Rabi, and M. Johansson, "A simple peer-to-peer algorithm for distributed optimization in sensor networks," in *Proc. IEEE Conference on Decision and Control*, New Orleans, LA, 2007, pp. 4705–4710.

- [9] A. Nedić and A. Ozdaglar, “On the rate of convergence of distributed subgradient methods for multi-agent optimization,” in *Proc. IEEE Conference on Decision and Control*, New Orleans, LA, 2007, pp. 4711–4716.
- [10] S. S. Ram, A. Nedić, and V. V. Veeravalli, “Stochastic incremental gradient descent for estimation in sensor networks,” in *Proc. Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, 2007, pp. 582–586.
- [11] B. Johansson, T. Keviczky, M. Johansson, and K. H. Johansson, “Subgradient methods and consensus algorithms for solving convex optimization problems,” in *Proc. IEEE Conference on Decision and Control*, Cancun, Mexico, 2008, pp. 4185–4190.
- [12] I. Lobel and A. Ozdaglar, “Convergence analysis of distributed subgradient methods over random networks,” in *Proc. Allerton Conference on Communication, Control, and Computing*, Monticello, IL, 2008, pp. 353–360.
- [13] A. Nedić, A. Olshevsky, A. Ozdaglar, and J. N. Tsitsiklis, “Distributed subgradient methods and quantization effects,” in *Proc. IEEE Conference on Decision and Control*, Cancun, Mexico, 2008, pp. 4177–4184.
- [14] A. Nedić and A. Ozdaglar, “Distributed subgradient methods for multi-agent optimization,” *IEEE Transactions on Automatic Control*, vol. 54, no. 1, pp. 48–61, 2009.
- [15] S. S. Ram, A. Nedić, and V. V. Veeravalli, “Incremental stochastic subgradient algorithms for convex optimization,” *SIAM Journal on Optimization*, vol. 20, no. 2, pp. 691–717, 2009.
- [16] —, “Asynchronous gossip algorithms for stochastic optimization,” in *Proc. IEEE Conference on Decision and Control*, Shanghai, China, 2009, pp. 3581–3586.
- [17] —, “Distributed stochastic subgradient projection algorithms for convex optimization,” *Journal of Optimization Theory and Applications*, vol. 147, no. 3, pp. 516–545, 2010.
- [18] J. N. Tsitsiklis, “Problems in decentralized decision making and computation,” Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, 1984.
- [19] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, “Randomized gossip algorithms,” *IEEE Transactions on Information Theory*, vol. 52, no. 6, pp. 2508–2530, 2006.
- [20] J. Lu, C. Y. Tang, P. R. Regier, and T. D. Bow, “A gossip algorithm for convex consensus optimization over networks,” in *Proc. American Control Conference*, Baltimore, MD, 2010, pp. 301–308.