A Second-Order Proximal Algorithm for Consensus Optimization

Xuyang Wu, Zhihai Qu and Jie Lu

***Abstract*—We develop a distributed second-order proximal algorithm, referred to as SoPro, for addressing in-network consensus optimization. The proposed SoPro algorithm is shown to be linearly convergent to the exact optimal solution, provided that the global cost function is locally restricted strongly convex. This relaxes the standard global strong convexity condition required by the existing distributed optimization algorithms to establish linear convergence. In addition, we demonstrate that SoPro is computation- and communication-efficient in comparison with the state-of-the-art distributed second-order methods. Finally, competitive convergence performance and scalability of SoPro are illustrated via numerical results.**

***Index Terms*—distributed optimization, consensus optimization, second-order method, proximal algorithm.**

I. Introduction

This paper considers in-network consensus optimization problems, where nodes in a network cooperatively find a global decision that minimizes the sum of their private cost functions. This problem is prevalent in many engineering applications, such as in-network robust estimation [1], network resource allocation [2], and distributed machine learning [3].

A considerable number of (discrete-time) distributed algorithms for convex consensus optimization have been proposed in the last decade, which allow the nodes to reach an optimal consensus solely through local interactions with neighbors. Most of the existing algorithms are first-order (i.e., gradient/subgradient) methods. Typical first-order methods include consensus-based subgradient algorithms [4]–[6], distributed accelerated gradient methods [7]–[16], and dual/primal-dual gradient methods [17]–[20].

Recently, distributed second-order methods [21]–[26] have been brought to attention. Such methods rely on second-order information (i.e., Hessian matrices) to evolve, which are characterized by more accurate approximations of certain global objectives and thus often converge faster than first-order methods. Specifically, the Decentralized Broyden-Fletcher-Goldfarb-Shanno (D-BFGS) method [21], the Distributed Quasi-Newton (DQN) method [22], and the Network Newton(NN) method [23] relax the consensus constraint by adding a penalty to the objective function and then approximate the Hessian inverse of the penalized objective in various ways to enable distributed computations. These methods, due to the discrepancy between the original problem and the penalized one, can only guarantee convergence to a suboptimal solution with an error bound provided. The Newton-Raphson Consensus (NRC) method [24], the Exact Second-Order Method (ESOM) [25], and the Decentralized Quadratically Approximated ADMM (DQM) are capable of reaching exact optimality, where NRC combines a consensus scheme with the Newton-Raphson method, ESOM employs second-order approximations of an augmented Lagrangian function, and DQM introduces quadratic approximations to a decentralized ADMM.

In this paper, we propose a distributed second-order proximal algorithm, referred to as SoPro, for solving convex consensus optimization problems over undirected networks. To develop SoPro, we first design a novel quadratic proximal term and add it to an augmented Lagrangian function. We then incorporate such proximal augmented Lagrangian into the Method of Multipliers [27]. Further, we replace the private cost functions with their second-order approximations when minimizing the proximal augmented Lagrangian, so that computational efforts can be significantly reduced. We show that the resulting SoPro can be fully decentralized and provide an explicit parameter condition to guarantee convergence. The advantages of SoPro are highlighted as follows:

1) SoPro achieves a linear rate of convergence to the exact optimum, under the assumption that the *sum* of the private cost functions is *locally restricted* strongly convex.
2) The convergence result of SoPro relaxes the *global* strong convexity condition required by all the aforementioned second-order methods [21]–[26] and the linearly convergent first-order methods [8], [10]–[16], [19], [20].
3) SoPro is more efficient than most existing distributed second-order methods for consensus optimization and is commensurate with the rest, in the sense of computation and communication complexities per iteration.
4) Simulations reflect that SoPro converges faster and is more scalable with respect to the network size than the existing distributed second-order methods.

The outline of the paper is as follows: Section II formulates the problem. Section III describes the algorithm and Section IV presents the convergence result. Section V illustrates simulation results and Section VI concludes the paper.

Notations: For any differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, its gradient at $x \in \mathbb{R}^d$ is denoted by $\nabla f(x)$. If f is twice-differentiable, its Hessian matrix at x is denoted by $\nabla^2 f(x)$. We use $\mathbf{0}_d$, $\mathbf{1}_d$, $\mathbf{O}_d$, I_d to denote the d-dimensional all-zero vector, d-dimensional all-one vector, $d \times d$ zero matrix, and $d \times d$ identity matrix, respectively. We also use $\mathrm{diag}(A_1, \ldots, A_n)$ to represent a block diagonal matrix, where $A_1, \ldots, A_n$ are square matrices. Moreover, $\otimes$ represents the Kronecker product, $\langle \cdot, \cdot \rangle$ the inner product, and $\| \cdot \|$ the ℓ_2 norm. Moreover, given $A \in \mathbb{R}^{d \times d}$ and $\mathbf{x} \in \mathbb{R}^d$, $\|\mathbf{x}\|_A^2 = \mathbf{x}^T A \mathbf{x}$. We use $\lambda_{\max}(A)$ to denote A's largest real eigenvalue, $\lambda_{\min}(A)$ to denote A's smallest real eigenvalue,

This work has been supported by the National Natural Science Foundation of China under grant 61603254 and the Natural Science Foundation of Shanghai under grant 16ZR1422500.

X. Wu, Z. Qu and J. Lu are with the School of Information Science and Technology, ShanghaiTech University, 201210 Shanghai, China. Email: {wuxy, quzhh1, lujie}@shanghaitech.edu.cn.

and $A^\dagger$ to denote A's pseudo-inverse. Finally, for any set $C \subseteq \mathbb{R}^d$, $P_C(x)$ denotes the projection of $x \in \mathbb{R}^d$ onto C and $\mathrm{diam}(C) = \max_{x,y\in C} \|x - y\|$ is the diameter of C.

II. Problem Formulation

In this section, we formulate the distributed optimization problem and impose assumptions on the problem.

A. Preliminaries

For any $C \subseteq \mathbb{R}^d$, a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be *strongly convex on* C if there is $m_C > 0$ such that

$$(\nabla f(x) - \nabla f(y))^T(x - y) \geq m_C\|x - y\|^2, \quad \forall x, y \in C,$$

where m_C is called the convexity parameter. We say f is *globally strongly convex* if $C = \mathbb{R}^d$. In addition, f is *locally strongly convex* if it is strongly convex on every compact subset of $\mathbb{R}^d$.

Let $\tilde{x} \in \mathbb{R}^d$ and $C \subseteq \mathbb{R}^d$ containing $\tilde{x}$. We say f is *restricted strongly convex with respect to* $\tilde{x}$ *on* C if there exists a convexity parameter $m_C > 0$ such that

$$(\nabla f(x) - \nabla f(\tilde{x}))^T(x - \tilde{x}) \geq m_C\|x - \tilde{x}\|^2, \quad \forall x \in C.$$

Likewise, f is *globally restricted strongly convex with respect to* $\tilde{x}$ if $C = \mathbb{R}^d$, and f is *locally restricted strongly convex with respect to* $\tilde{x}$ if it is restricted strongly convex with respect to $\tilde{x}$ on all compact subsets of $\mathbb{R}^d$ containing $\tilde{x}$.

Given $M \geq 0$, f is said to be M*-smooth* if ∇f satisfies the Lipschitz condition

$$\|\nabla f(x) - \nabla f(y)\| \leq M\|x - y\|, \ \forall x, y \in \mathbb{R}^d.$$

B. Distributed Optimization

Suppose a set $\mathcal{V} = \{1, \ldots, N\}$ of agents form a network, which can be modeled as a connected, undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the vertex set and $\mathcal{E} \subseteq \{\{i, j\} \subseteq \mathcal{V}\times\mathcal{V} \,|i \neq j\}$ is the edge set. For each $i \in \mathcal{V}$, $\mathcal{N}_i = \{j \in \mathcal{V} \,|\{i, j\} \in \mathcal{E}\}$ denotes the set of node i's neighbors. Suppose each node $i \in \mathcal{V}$ observes a private cost function $f_i : \mathbb{R}^d \to \mathbb{R}$, and all the N nodes aim at collectively solving

$$\underset{x\in\mathbb{R}^d}{\text{minimize}} \quad \sum_{i\in\mathcal{V}} f_i(x), \tag{1}$$

which satisfies the assumption below:

Assumption 1. *Problem* (1) *satisfies the following:*

(a) For each $i \in \mathcal{V}$, f_i is convex, twice-differentiable, and M_i-smooth, where $M_i \geq 0$.

(b) There exists an optimal solution $x^\star \in \mathbb{R}^d$ of problem (1)*.*

(c) $\sum_{i\in\mathcal{V}} f_i$ is locally restricted strongly convex with respect to $x^\star$.

Note that Assumption 1(c) guarantees that the optimal solution $x^\star$ of problem (1) is unique. Problems satisfying Assumption 1 are more general than those solved by the existing distributed second-order methods [22]–[26]. Specifically, Assumption 1(a)(b) are also required by [22]–[26], while we eliminate some smoothness conditions in these existing works, such as the Lipschitz continuity of $\nabla^2 f$ in [23]–[26] and the twice continuous differentiability of f_i in [22], [24]. Moreover, Assumption 1(c) is less restrictive than the global strong convexity of each individual f_i required by [22], [23], [25], [26] and the global strong convexity of $\sum_{i\in\mathcal{V}} f_i$ required by [24]. D-BFGS [21] is the only existing second-order method that does not require any strong convexity. It needs no Hessian evaluation yet can only achieve a suboptimal solution.

Below are two examples satisfying Assumption 1, in which the objective functions are not globally strongly convex:

Example 1. *The logistic regression problem [28] arising in machine learning is in the form of* (1)*, where*

$$f_i(x) = \sum_{j=1}^{p_i} \log(1 + \exp(-y_{ij}\langle x, s_{ij}\rangle)), \quad \forall i \in \mathcal{V}.$$

Here, p_i is node i's sample number, s_{ij} is its jth feature, and y_{ij} is the label corresponding to s_{ij}. Observe that each f_i is locally but not globally strongly convex.

Example 2. *Let each $f_i : \mathbb{R} \to \mathbb{R}$, $i \in \mathcal{V}$ be given by*

$$f_1(x) = \begin{cases} -2x/3 + 1/4, & x \leq -1, \\ x^2/2 - x^4/12, & -1 < x < 1, \\ 2x/3 - 1/4, & x \geq 1, \end{cases}$$
$$f_i(x) = c_i x + d_i, \ \ c_i, d_i \in \mathbb{R}, \ \ i = \{2, \ldots, N\}.$$

Observe that the restricted strong convexity of f_1 with respect to $x^\star$ holds on every compact set containing $x^\star$ but fails to hold on any noncompact set. Also note that $f_2, \ldots, f_N$ are affine functions and in fact can be any convex functions with Lipschitz gradients, whose Hessian matrices do not have to be positive definite.

III. SoPro: Second-Order Proximal Algorithm

In this section, we develop a distributed second-order proximal algorithm for solving problem (1) under Assumption 1.

To this end, let $x_i \in \mathbb{R}^d$ be node i's local copy of the global decision variable x and let $\mathbf{x} = (x_1^T, \ldots, x_N^T)^T$. Then, we equivalently transform problem (1) into

$$\begin{aligned} \underset{\mathbf{x}\in\mathbb{R}^{Nd}}{\text{minimize}} \quad & f(\mathbf{x}) = \sum_{i\in\mathcal{V}} f_i(x_i) \\ \text{subject to} \quad & \mathbf{x} \in S, \end{aligned} \tag{2}$$

where $S = \{\mathbf{x} \in \mathbb{R}^{Nd} | \ x_1 = \ldots = x_N\}$ is the set of vectors consisting of N identical d-dimensional blocks. Note that the unique optimal solution of problem (2) is $\mathbf{x}^\star = ((x^\star)^T, \ldots, (x^\star)^T)^T \in S$.

Let $P = P^T \in \mathbb{R}^{N\times N}$ be given by

$$[P]_{ij} = \begin{cases} \sum_{s\in\mathcal{N}_i} p_{is}, & i = j, \\ -p_{ij}, & j \in \mathcal{N}_i, \\ 0, & \text{otherwise}, \end{cases} \qquad \forall i, j \in \mathcal{V},$$

where $p_{ij} = p_{ji} > 0 \ \forall\{i, j\} \in \mathcal{E}$. This, together with the connectivity of $\mathcal{G}$, implies that $\mathrm{Null}(P) = \mathrm{span}\{\mathbf{1}_N\}$ and $P \succeq \mathbf{O}_N$. Let $W = P \otimes I_d \succeq \mathbf{O}_{Nd}$. It can be shown that

$$\mathrm{Range}(W) = \mathrm{Range}(W^{\frac{1}{2}}) = \mathrm{Range}(W^\dagger)$$

$$= \text{Range}((W^\dagger)^{\frac{1}{2}}) = S^\perp, \tag{3}$$

where $S^\perp = \{\mathbf{x} \in \mathbb{R}^{Nd} |\, x_1 + \ldots + x_N = \mathbf{0}_d\}$ is the orthogonal complement of S. Hence, the consensus constraint $x_1 = \ldots = x_N$ can be substituted with $W^{\frac{1}{2}}\mathbf{x} = \mathbf{0}_{Nd}$, so that problem (2) is equivalent to

$$\begin{aligned} \underset{\mathbf{x}\in\mathbb{R}^{Nd}}{\text{minimize}} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & W^{\frac{1}{2}}\mathbf{x} = \mathbf{0}_{Nd}. \end{aligned} \tag{4}$$

Let $\mathbf{v} \in \mathbb{R}^{Nd}$ be the dual variable associated with the constraint $W^{\frac{1}{2}}\mathbf{x} = \mathbf{0}_{Nd}$ in (4). Consider the augmented Lagrangian function $L_\rho : \mathbb{R}^{Nd} \times \mathbb{R}^{Nd} \to \mathbb{R}$ given by

$$L_\rho(\mathbf{x}, \mathbf{v}) = f(\mathbf{x}) + \mathbf{v}^T W^{\frac{1}{2}}\mathbf{x} + \frac{\rho}{2}\mathbf{x}^T W \mathbf{x},$$

where $\rho > 0$ is the penalty parameter. The application of the Method of Multipliers [27] to problem (4) gives

$$\mathbf{x}^{k+1} = \arg\ \min_{\mathbf{x}\in\mathbb{R}^{Nd}} L_\rho(\mathbf{x}, \mathbf{v}^k), \tag{5}$$

$$\mathbf{v}^{k+1} = \mathbf{v}^k + \rho W^{\frac{1}{2}}\mathbf{x}^{k+1}. \tag{6}$$

However, (5) cannot be implemented over the graph $\mathcal{G}$ in a distributed manner due to its coupling structure. To overcome this issue, we add the proximal term $(\mathbf{x}-\mathbf{x}^k)^T(D-\rho W)(\mathbf{x}-\mathbf{x}^k)/2$ to $L_\rho(\mathbf{x}, \mathbf{v}^k)$ in (5), where $D = \text{diag}(D_1, \ldots, D_N)$ is a symmetric block diagonal matrix with each $D_i \in \mathbb{R}^{d\times d}$, and then substitute $f(\mathbf{x})$ in $L_\rho(\mathbf{x}, \mathbf{v}^k)$ with its second-order approximation about $\mathbf{x}^k$. This yields

$$\begin{aligned} \mathbf{x}^{k+1} = \arg\ \min_{\mathbf{x}\in\mathbb{R}^{Nd}} & f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x} - \mathbf{x}^k \rangle \\ & + \frac{(\mathbf{x}-\mathbf{x}^k)^T \nabla^2 f(\mathbf{x}^k)(\mathbf{x}-\mathbf{x}^k)}{2} + (\mathbf{v}^k)^T W^{\frac{1}{2}}\mathbf{x} + \frac{\rho}{2}\mathbf{x}^T W \mathbf{x} \\ & + \frac{(\mathbf{x}-\mathbf{x}^k)^T(D-\rho W)(\mathbf{x}-\mathbf{x}^k)}{2}. \end{aligned}$$

Note that unlike the conventional proximal terms in ℓ_2 norm, here $D - \rho W$ in the quadratic proximal term does not have to be positive semidefinite. Also, in the above equation, $\mathbf{x}^{k+1}$ is well-defined only when $\nabla^2 f(\mathbf{x}^k) + D \succ \mathbf{O}_{Nd}$. We will provide a condition to guarantee this later in Section IV. Then, by the first-order optimality condition, the above update is exactly the following:

$$\begin{aligned} \mathbf{x}^{k+1} = \mathbf{x}^k & \\ - (\nabla^2 f(\mathbf{x}^k) + D)^{-1}(\nabla f(\mathbf{x}^k) + \rho W \mathbf{x}^k + W^{\frac{1}{2}}\mathbf{v}^k). & \end{aligned} \tag{7}$$

Observe that (7) reduces the computation load in updating $\mathbf{x}^{k+1}$ to a great extent compared with (5) owing to the second-order approximation of $f(\mathbf{x})$ at $\mathbf{x}^k$.

Further, we let $\mathbf{q}^k = W^{\frac{1}{2}}\mathbf{v}^k$, so that the above algorithm can be described by

$$\mathbf{x}^{k+1} = \mathbf{x}^k - (\nabla^2 f(\mathbf{x}^k) + D)^{-1}(\nabla f(\mathbf{x}^k) + \rho W \mathbf{x}^k + \mathbf{q}^k), \tag{8}$$

$$\mathbf{q}^{k+1} = \mathbf{q}^k + \rho W \mathbf{x}^{k+1}. \tag{9}$$

In addition, since $\text{Range}(W^{\frac{1}{2}}) = S^\perp$, we require $\mathbf{q}^k \in S^\perp$ $\forall k \geq 0$, which, due to (9), can be satisfied by letting

$$\mathbf{q}^0 \in S^\perp. \tag{10}$$

Indeed, we may simply set $\mathbf{q}^0 = \mathbf{0}_{Nd}$. As the proposed algorithm (8)–(10) employs the Hessian matrices of f and a quadratic proximal term, we refer to it as Second-Order Proximal (SoPro) algorithm.

The SoPro algorithm (8)–(10) can be implemented over the network $\mathcal{G}$ in a fully distributed fashion, i.e., each node $i \in \mathcal{V}$ only interacts with its neighbors. To see this, we associate x_i^k and q_i^k with each node $i \in \mathcal{V}$, which are the ith block of $\mathbf{x}^k$ and $\mathbf{q}^k$, respectively. For better presentation of each node's behaviors, we also let every node $i \in \mathcal{V}$ maintain an auxiliary variable y_i^k such that $\mathbf{y}^k = ((y_1^k)^T, \ldots, (y_N^k)^T)^T = W\mathbf{x}^k$. The detailed actions taken by the nodes are described below in Algorithm 1. Legitimate selections of the algorithm parameters will be discussed in Section IV.

Algorithm 1 Second-Order Proximal Algorithm (SoPro)

1: **Initialization:**
2: Each node $i \in \mathcal{V}$ arbitrarily sets $x_i^0 \in \mathbb{R}^d$, selects $q_i^0 \in \mathbb{R}^d$ so that $\sum_{i\in\mathcal{V}} q_i^0 = \mathbf{0}_d$ (or simply sets $q_i^0 = \mathbf{0}_d$), and sends x_i^0 to every neighbor $j \in \mathcal{N}_i$.
3: Upon receiving x_j^0 $\forall j \in \mathcal{N}_i$, each node $i \in \mathcal{V}$ sets $y_i^0 = \sum_{j\in\mathcal{N}_i} p_{ij}(x_i^0 - x_j^0)$.
4: **for** $k = 0, 1, \ldots$ **do**
5: Each node $i \in \mathcal{V}$ updates

$$x_i^{k+1} = x_i^k - (\nabla^2 f_i(x_i^k) + D_i)^{-1}(\nabla f_i(x_i^k) + \rho y_i^k + q_i^k).$$

6: Each node $i \in \mathcal{V}$ sends x_i^{k+1} to every neighbor $j \in \mathcal{N}_i$.
7: Upon receiving x_j^{k+1} $\forall j \in \mathcal{N}_i$, each node $i \in \mathcal{V}$ updates

$$y_i^{k+1} = \sum_{j\in\mathcal{N}_i} p_{ij}(x_i^{k+1} - x_j^{k+1}),$$

$$q_i^{k+1} = q_i^k + \rho y_i^{k+1}.$$

8: **end for**

The computational complexity and the communication cost (i.e., real-number transmissions) per node at each iteration of the existing distributed second-order methods [21]–[26] and our proposed SoPro are tabulated in Table I. Note that the update of y_i^{k+1} in SoPro can be rewritten as $y_i^{k+1} = (1 - [P]_{ii})x_i^{k+1} - \sum_{j\in\mathcal{N}_i} p_{ij}x_j^{k+1}$, which leads to fewer elementary operations. Similar operations to reduce complexity are carried out for DQN-0, DQN-1, DQN-2, and DQM. Observe that SoPro in the worst case is more computationally efficient than NN-K, K ≥ 1, DQN-1, DQN-2, D-BFGS, NRC, and ESOM-K, K ≥ 1, and has similar complexity as NN-0 and DQN-0. In the special case that each D_i is diagonal and $p_{ij} = p_{ji} = 1$ $\forall\{i, j\} \in \mathcal{E}$, the computational complexity of SoPro is better than that of NN-0 and DQN-0, and is very close to that of ESOM-0 and DQM. Regarding communication cost, SoPro is as efficient as NN-0, DQN-0, ESOM-0, and DQM, and outperforms the remaining algorithms.

IV. Convergence Analysis

This section establishes a linear convergence rate of SoPro and discusses the parameter selections to guarantee the rate.

Algorithm	Computational Complexity		Communication Cost
	Matrix Inverse	Elementary Operation	
NN-K (K ≥ 0)	$d \times d$	$(2K+3)d^2 + (2(K+1)\lvert\mathcal{N}_i\rvert + K + 5)d$	$(K+1)\lvert\mathcal{N}_i\rvert d$
DQN-0	$d \times d$	$3d^2 + (2\lvert\mathcal{N}_i\rvert + 5)d$	$\lvert\mathcal{N}_i\rvert d$
DQN-1	$d \times d$	$3d^2 + (4\lvert\mathcal{N}_i\rvert + 8)d$	$2\lvert\mathcal{N}_i\rvert d$
DQN-2	$d \times d$	$6d^2 + 6\lvert\mathcal{N}_i\rvert d + 10d$	$3\lvert\mathcal{N}_i\rvert d$
D-BFGS	$\lvert\mathcal{N}_i\rvert d \times \lvert\mathcal{N}_i\rvert d$	$8(\lvert\mathcal{N}_i\rvert + 1)^2 d^2 + (12\lvert\mathcal{N}_i\rvert + 13)d - 2$	$3\lvert\mathcal{N}_i\rvert d$
NRC	$d \times d$	$(4\lvert\mathcal{N}_i\rvert + 7)d^2 + (4\lvert\mathcal{N}_i\rvert + 5)d$	$\lvert\mathcal{N}_i\rvert(d^2+3d)$
ESOM-K (K ≥ 0)	$d \times d$	$2(K+1)d^2 + (2\lvert\mathcal{N}_i\rvert(K+1) + K + 6)d$	$(K+1)\lvert\mathcal{N}_i\rvert d$
DQM	$d \times d$	$2d^2 + (\lvert\mathcal{N}_i\rvert + 6)d$	$\lvert\mathcal{N}_i\rvert d$
SoPro (worst case)	$d \times d$	$3d^2 + (2\lvert\mathcal{N}_i\rvert + 6)d$	$\lvert\mathcal{N}_i\rvert d$
SoPro ($p_{ij} = 1$, D diagonal)	$d \times d$	$2d^2 + (\lvert\mathcal{N}_i\rvert + 7)d$	$\lvert\mathcal{N}_i\rvert d$

TABLE I
COMPUTATIONAL COMPLEXITY AND COMMUNICATION COST

A. Convergence Rate

To derive the convergence rate, we first define the following: For any compact $C \subset \mathbb{R}^{Nd}$ containing $\mathbf{x}^\star$, define

$$\bar{C} = \{x \in \mathbb{R}^d |\ x = \sum_{i\in\mathcal{V}} y_i/N,\ (y_1^T, \ldots, y_N^T)^T \in C\} \subset \mathbb{R}^d.$$

Clearly, $\bar{C}$ is a compact subset of $\mathbb{R}^d$ that contains $x^\star$. Due to Assumption 1, there exists $m_{\bar{C}} \in (0,\infty)$ such that $\forall x \in \bar{C}$,

$$\langle \sum_{i\in\mathcal{V}} \nabla f_i(x) - \sum_{i\in\mathcal{V}} \nabla f_i(x^\star), x - x^\star \rangle \geq m_{\bar{C}} \|x - x^\star\|^2. \quad (11)$$

Then, we obtain the following lemma:

Lemma 1. *Suppose Assumption 1 holds. Let $C \subset \mathbb{R}^{Nd}$ be any compact set containing $\mathbf{x}^\star$ and $\zeta_C : (0,\infty) \to \mathbb{R}$ be defined as $\zeta_C(\gamma) = \min\{m_{\bar{C}}/N - 2M\gamma, \rho\frac{\lambda_W}{2(1+1/\gamma^2)}\}$, where $m_{\bar{C}} \in (0,\infty)$ satisfies (11), $M := \max_{i\in\mathcal{V}} M_i \in [0,\infty)$, and $\lambda_W \in (0,\infty)$ is the smallest nonzero eigenvalue of W. Also let $f_a(\mathbf{x}) = f(\mathbf{x}) + \frac{\rho}{4}\|\mathbf{x}\|_W^2$. Then, for any $\gamma > 0$ and $\mathbf{x} \in C$,*

$$\langle \nabla f_a(\mathbf{x}) - \nabla f_a(\mathbf{x}^\star), \mathbf{x} - \mathbf{x}^\star \rangle \geq \zeta_C(\gamma)\|\mathbf{x} - \mathbf{x}^\star\|^2. \quad (12)$$

Proof. See Appendix A. □

Note from Lemma 1 that $\zeta_C(\gamma) > 0$ if and only if $\gamma \in (0, \frac{m_{\bar{C}}}{2MN})$. Therefore, f_a is locally restricted strongly convex with respect to $\mathbf{x}^\star$, and its convexity parameter on any compact C containing $\mathbf{x}^\star$ can be any positive $\zeta_C(\gamma)$. This generalizes the result in [8], which says that $f_a(\mathbf{x})$ is globally restricted strongly convex with respect to $\mathbf{x}^\star$ if $\sum_{i\in\mathcal{V}} f_i(x)$ is globally restricted strongly convex with respect to $x^\star$.

Next, we construct a compact set such that the primal iterates $\mathbf{x}^k\ \forall k \geq 0$ generated by SoPro with proper parameters stay in that set. To this end, note that any $\mathbf{v} \in \mathbb{R}^{Nd}$ satisfying $\nabla f(\mathbf{x}^\star) = -W^{1/2}\mathbf{v}$ is a dual optimum to problem (4). Here, we consider a particular dual optimum

$$\mathbf{v}^\star = -(W^\dagger)^{1/2}\nabla f(\mathbf{x}^\star). \quad (13)$$

Throughout this section, we let $\mathbf{v}^k = (W^\dagger)^{1/2}\mathbf{q}^k$, which satisfies (6), (7) and is a particular choice of the $\mathbf{v}^k$ in Section III. From (3), we know that

$$\mathbf{v}^k,\ \mathbf{v}^\star,\ \mathbf{v}^k - \mathbf{v}^\star \in S^\perp. \quad (14)$$

Further, due to the convexity of each f_i, $\nabla^2 f_i(x) \succeq m_i I_d$ $\forall x \in \mathbb{R}^d$ for some $m_i \geq 0$. Let $\Lambda_m = \mathrm{diag}(m_1, \ldots, m_N) \otimes I_d \succeq \mathbf{O}_{Nd}$, $\Lambda_M = \mathrm{diag}(M_1, \ldots, M_N) \otimes I_d \succeq \mathbf{O}_{Nd}$, and $R = (\Lambda_m + \Lambda_M)/2 + D - \rho W$. We assume $R \succ \mathbf{O}_{Nd}$. Then, define $Q = \begin{pmatrix} \rho R & \mathbf{O}_{Nd} \\ \mathbf{O}_{Nd} & \mathbf{I}_{Nd} \end{pmatrix} \succ \mathbf{O}_{2Nd}$. Also let $\mathbf{z}^k = ((\mathbf{x}^k)^T, (\mathbf{v}^k)^T)^T$ and $\mathbf{z}^\star = ((\mathbf{x}^\star)^T, (\mathbf{v}^\star)^T)^T$, where $\mathbf{x}^k$ is the primal iterate of SoPro and $\mathbf{v}^k, \mathbf{v}^\star$ are defined above. Then, let $S_0 = \{\mathbf{x} |\ \|\mathbf{x} - \mathbf{x}^\star\|_R^2 + \rho\|\mathbf{x}\|_W^2 \leq \frac{1}{\rho}\|\mathbf{z}^0 - \mathbf{z}^\star\|_Q^2 + \rho\|\mathbf{x}^0\|_W^2\}$, which is compact and convex. Recall from Lemma 1 that f_a is restricted strongly convex with respect to $\mathbf{x}^\star$ on S_0 and we denote the corresponding convexity parameter by $m_\rho > 0$.

Based on the above, we show in the lemma below that $(\mathbf{x}^k)_{k=0}^\infty \subseteq S_0$ with proper algorithm parameters.

Lemma 2. *Suppose Assumption 1 holds. Also suppose*

$$\frac{\Lambda_M}{2(1-\eta)} + \frac{(\Lambda_M - \Lambda_m)^2}{8\eta m_\rho} + \Lambda_M - \Lambda_m \prec R \quad (15)$$

for some $\eta \in (0,1)$. Then, $\mathbf{x}^k \in S_0\ \forall k \geq 0$.

Proof. See Appendix B. □

Since $\Lambda_M \succeq \Lambda_m$, (15) implies $R \succ \mathbf{O}_{Nd}$. Moreover, with (15), it can be shown that $\nabla^2 f(\mathbf{x}^k) + D \succ \mathbf{O}_{Nd}\ \forall k \geq 0$, so that the updates of the x_i^k's in SoPro are well-posed. The boundedness of $(\mathbf{x}^k)_{k=0}^\infty$ in Lemma 2 further leads to the following result on the convergence rate of SoPro:

Theorem 1. *Suppose Assumption 1 holds. Also suppose (15) holds for some $\eta \in (0,1)$. Then, there exists $\delta \in (0,1)$ such that for each $k \geq 0$,*

$$\begin{aligned} &\|\mathbf{z}^{k+1} - \mathbf{z}^\star\|_Q^2 + \rho^2\|\mathbf{x}^{k+1}\|_W^2 \\ &\leq (1-\delta)(\|\mathbf{z}^k - \mathbf{z}^\star\|_Q^2 + \rho^2\|\mathbf{x}^k\|_W^2). \end{aligned} \quad (16)$$

In particular,

$$\delta = \sup_{c_1,c_2>0} \min\left\{ \frac{\rho\lambda_W\kappa_{\beta,\eta}}{(1+c_1)\|\Lambda_M + D\|^2}, \frac{1-\eta}{(1+1/c_1)(1+c_2)+1}, \frac{2\eta m_\rho - \beta}{\lambda_{\max}(R + (1+1/c_1)(1+1/c_2)\Lambda_M^2/(\rho\lambda_W))} \right\}, \quad (17)$$

where λ_W is defined in Lemma 1 and $\beta \in (0, 2\eta m_\rho)$ is such that $\kappa_{\beta,\eta} := \lambda_{\min}(R - \Lambda_M/(2(1-\eta)) - (\Lambda_M - \Lambda_m)^2/(4\beta) - \Lambda_M + \Lambda_m) > 0$.

Proof. See Appendix C. □

Theorem 1 shows the linear convergence of $\|\mathbf{z}^k - \mathbf{z}^\star\|_Q^2 + \rho^2\|\mathbf{x}^k\|_W^2$ to zero as well as the R-linear convergence of $\mathbf{x}^k$ to the primal optimum $\mathbf{x}^\star$. Note that $\|\mathbf{z}^k - \mathbf{z}^\star\|_Q^2$ is a metric of both primal and dual optimality and $\rho^2\|\mathbf{x}^k\|_W^2$ quantifies the consensus error of the x_i^k's.

There have been a number of distributed first-order and second-order algorithms [8], [10]–[16], [19], [20], [24]–[26] that also achieve a linear rate of convergence to the exact

optimum or an inexact solution of problem (1). Nevertheless, to establish such linear rates, [10]–[16], [19], [20], [24]–[26] require global strong convexity of the objective functions and [8] needs global restricted strong convexity. SoPro relaxes such conditions to local restricted strong convexity and still ensures a linear rate of convergence to the exact optimum.

B. Parameter Selection

To guarantee the convergence rate in Theorem 1, we need to select appropriate parameters ρ, P and D such that (15) holds. To this end, we let ρ be an arbitrary positive scalar predetermined by all the nodes. Also, let each pair of neighbors $i, j \in \mathcal{V}$ agree on any positive value of $p_{ij} = p_{ji}$. Then, the block diagonal matrix D may be determined as follows:

Note from $R = \frac{\Lambda_m+\Lambda_M}{2} + D - \rho W$ that (15) holds if

$$D \succ \Big(\frac{(M-m)^2}{8\eta m_\rho} + \frac{M}{2(1-\eta)}\Big) I_{Nd} + \rho\Lambda_W - \frac{3}{2}\Lambda_m + \frac{\Lambda_M}{2}, \quad (18)$$

where $M = \max_{i\in\mathcal{V}} M_i$, $m = \min_{i\in\mathcal{V}} m_i$, and $\Lambda_W \succeq W$ is a block diagonal matrix. Since both sides of (18) are block diagonal, each node $i \in \mathcal{V}$ can determine its D_i if it knows m, M, η, the ith block of Λ_W, and a lower bound on m_ρ.

Since each node i knows M_i and m_i, they can collectively find M and m at low cost and even in a decentralized fashion (e.g., [29]). Also, η can be any number in $(0, 1)$. In fact, the optimal choice of η that minimizes $\frac{(M-m)^2}{8\eta m_\rho} + \frac{M}{2(1-\eta)}$ is $\eta^\star = 1 - \frac{1}{1+(M-m)/(2\sqrt{m_\rho M})}$. To find $\Lambda_W \succeq W$, one option is $\Lambda_W = \mathrm{diag}(2[P]_{11}, \ldots, 2[P]_{NN}) \otimes I_d$, for which $\Lambda_W - W$ is diagonally dominant with positive diagonal elements. Another option is $\Lambda_W = \max_{\{i,j\}\in\mathcal{E}} p_{ij} N I_{Nd}$, since $\lambda_{\max}(W) = \lambda_{\max}(P) \le \max_{\{i,j\}\in\mathcal{E}} p_{ij} N$.

Now it remains to determine a lower bound on m_ρ, which can be explicitly computed in the following three cases:

1) Each $f_i(x)$, $i \in \mathcal{V}$ is globally restricted strongly convex with respect to $x^\star$: In this case, $f(\mathbf{x})$ is globally restricted strongly convex with respect to $\mathbf{x}^\star$ and the convexity parameter is $\min_{i\in\mathcal{V}} \theta_i$, where θ_i is the convexity parameter of f_i. Thus, we can simply choose $m_\rho = \min_{i\in\mathcal{V}} \theta_i$.

2) $\sum_{i\in\mathcal{V}} f_i(x)$ is globally restricted strongly convex with respect to $x^\star$: From Lemma 1 with $C = S_0$, m_ρ can be $\zeta_{S_0}(\gamma) > 0$ for any $\gamma \in (0, m_{\bar{S}_0}/(2MN))$, where $m_{\bar{S}_0}$ is the convexity parameter of $\sum_{i\in\mathcal{V}} f_i(x)$ on $\bar{S}_0$. Also, the maximum value of $\zeta_{S_0}(\gamma)$ can be attained when

$$4M(\gamma)^3 + (\rho\lambda_W - \frac{2m_{\bar{S}_0}}{N})(\gamma)^2 + 4M\gamma - \frac{2m_{\bar{S}_0}}{N} = 0. \quad (19)$$

Note that there is exactly one solution to (19) in $(0, \frac{m_{\bar{S}_0}}{2MN})$. Thus, the evaluation of a lower bound on m_ρ requires the knowledge of lower bounds on λ_W and $m_{\bar{S}_0}$. To bound λ_W, note that the smallest nonzero eigenvalue of the graph Laplacian $L_\mathcal{G}$ is no less than $\frac{4}{N\,\mathrm{diam}(\mathcal{G})}$ [30] where $\mathrm{diam}(\mathcal{G}) \le N-1$ is the diameter of $\mathcal{G}$ and that $W \succeq \min_{\{i,j\}\in\mathcal{E}} p_{ij}(L_\mathcal{G} \otimes I_d)$. Hence, $\lambda_W \ge \frac{4\min_{\{i,j\}\in\mathcal{E}} p_{ij}}{N\,\mathrm{diam}(\mathcal{G})} \ge \frac{4\min_{\{i,j\}\in\mathcal{E}} p_{ij}}{N(N-1)}$. Besides, $m_{\bar{S}_0}$ can be taken as the global convexity parameter of $\sum_{i\in\mathcal{V}} f_i(x)$.

3) There exists $p \in \mathcal{V}$ such that $f_p(x)$ is locally restricted strongly convex with respect to $x^\star$: From the previous case, the only issue for this case is to find a lower bound on $m_{\bar{S}_0}$. To this end, we let

$$D = \Big(\frac{(M-m)^2}{8\eta m_\rho} + \frac{M}{2(1-\eta)}\Big) I_{Nd} + \rho\Lambda_W - \frac{3}{2}\Lambda_m + \frac{\Lambda_M}{2} + \tilde{\Lambda},$$

where $\tilde{\Lambda} = \mathrm{diag}(\tilde{\Lambda}_1, \ldots, \tilde{\Lambda}_N)$ and each $\tilde{\Lambda}_i \in \mathbb{R}^{d\times d}$ is a symmetric positive definite matrix predetermined by node i. Clearly, the above D satisfies (18).

Next, we construct a compact set containing S_0. For simplicity, let $\alpha = \frac{(M-m)^2}{8\eta m_\rho} + \frac{M}{2(1-\eta)} > \frac{M}{2}$ and $D_D = \rho\Lambda_W - \Lambda_m + \Lambda_M + \tilde{\Lambda} \succ \mathbf{O}_{Nd}$. Then, $R \preceq \alpha I_{Nd} + D_D$. For any $\mathbf{x} \in S_0$,

$$\|\mathbf{x} - \mathbf{x}^\star\|^2 \le \|\mathbf{x}^0 - \mathbf{x}^\star\|^2_{I_{Nd} + \frac{2D_D}{M}} + \frac{2}{\rho M}\|\mathbf{v}^0 - \mathbf{v}^\star\|^2 + \frac{2\rho}{M}\|\mathbf{x}^0\|^2_W.$$

Due to (13), $\|\mathbf{v}^0 - \mathbf{v}^\star\|^2 = \|\mathbf{v}^0 + (W^\dagger)^{1/2}\nabla f(\mathbf{x}^0) - (W^\dagger)^{1/2}(\nabla f(\mathbf{x}^0) - \nabla f(\mathbf{x}^\star))\|^2 \le 2\|(W^\dagger)^{1/2}(\mathbf{q}^0 + \nabla f(\mathbf{x}^0))\|^2 + 2\|\nabla f(\mathbf{x}^0) - \nabla f(\mathbf{x}^\star)\|^2/\lambda_W \le 2\|\mathbf{q}^0 + \nabla f(\mathbf{x}^0)\|^2/\lambda_W + 2M^2\|\mathbf{x}^0 - \mathbf{x}^\star\|^2/\lambda_W$. Therefore, $S_0 \subseteq S_0' = \{\mathbf{x} |\ \|\mathbf{x} - \mathbf{x}^\star\|^2 \le \Gamma\}$, where $\Gamma \ge \|\mathbf{x}^0 - \mathbf{x}^\star\|^2_{(4M/(\rho\lambda_W)+1)I_{Nd}+2D_D/M} + 4\|\mathbf{q}^0 + \nabla f(\mathbf{x}^0)\|^2/(\rho M\lambda_W) + 2\rho\|\mathbf{x}^0\|^2_W/M$. It can be seen that as long as node p can estimate an upper bound on $\|\mathbf{x}^0 - \mathbf{x}^\star\|$, it is able to evaluate Γ and thus S_0' and $S_0' \supseteq \bar{S}_0$, since all the other global quantities used to determine Γ can be easily found in a decentralized way. Then, the convexity parameter of f_p on S_0', which can be figured out by node p, is a lower bound on $m_{\bar{S}_0}$. Finally, the following proposition may be used to bound $\|\mathbf{x}^0 - \mathbf{x}^\star\|^2 = \sum_{j\in\mathcal{V}} \|x_j^0 - x^\star\|^2$:

Proposition 1. *Suppose $f_i^\star = \inf_{x\in\mathbb{R}^d} f_i(x) > -\infty$ $\forall i \in \mathcal{V}$. Then, $\|x_j^0 - x^\star\| \le \mathrm{diam}\{x \in \mathbb{R}^d : f_i(x) \le c_j - |f_i^\star|\}$ $\forall i, j \in \mathcal{V}$, where $c_j = \sum_{i\in\mathcal{V}}(f_i(x_j^0) + |f_i^\star|)$.*

Proof. See Appendix D. □

According to Proposition 1, node p can find a finite upper bound on each $\|x_j^0 - x^\star\|$, $j \in \mathcal{V}$, provided that the locally restricted strongly convex f_p has bounded level sets and all the f_i's are bounded from below.

V. Numerical Example

Below we compare the convergence performance of SoPro with that of the existing distributed second-order methods NN-K, DQN-K, ESOM-K, NRC, DQM and D-BFGS in [21]–[26] via numerical examples. For NN-K, DQN-K, and ESOM-K, we choose K = 0 so that their computational and communication complexities are commensurate with SoPro. Also note that NN-0 is a special case of DQN-0 and therefore is omitted.

Consider the logistic regression problem with l_2 regularization that often arises in machine learning [27]: Each node i observes p_i training samples $(\mathbf{u}_{ij}, v_{ij}) \in \mathbb{R}^{d-1} \times \{-1, 1\}$, $j = 1, \ldots, p_i$, where $\mathbf{u}_{ij}$ is the feature vector and v_{ij} is the label. All the nodes need to collectively minimize the total logistic loss given by

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\ \frac{\lambda}{2}\|x\|^2 + \sum_{i\in\mathcal{V}}\sum_{j=1}^{p_i} \ln(1+\exp(-(\mathbf{u}_{ij}^T x_1 + x_0)v_{ij})),$$

where $x = (x_1^T, x_0)^T$, $x_1 \in \mathbb{R}^{d-1}$, $x_0 \in \mathbb{R}$, and $\lambda > 0$. By letting $f_i(x) = \frac{\lambda}{2N}\|x\|^2 + \sum_{j=1}^{p_i} \ln(1+\exp(-(\mathbf{u}_{ij}^T x_1 + x_0)v_{ij}))$ $\forall i \in \mathcal{V}$, the above logistic regression problem is in the form of problem (1).

In the simulations, we set $p_i = 10$ and $d = 3$, i.e., each node i holds 10 samples of dimension 3. We also let the number of positive and negative labels be equal. In addition, each element of $\mathbf{u}_{ij} \in \mathbb{R}^{d-1}$ is drawn from the standard normal distribution with mean value $v_{ij} = \pm 1$ and with variance 0.2. The networks are formed by undirected, connected random geometric graphs. Moreover, to illustrate the effect of network size N, connectivity ratios r (i.e., the ratio of the number of links to the number of all possible links) and convexity parameter λ, we set (N, r, λ) to $(50, 0.2, 1)$, $(50, 0.6, 1)$, $(50, 0.2, 10)$, and $(200, 0.2, 1)$, respectively. Also, we let the weight matrix W in SoPro be such that $p_{ij} = p_{ji} = \frac{1}{\max\{|\mathcal{N}_i|,|\mathcal{N}_j|\}+2}$ $\forall \{i, j\} \in \mathcal{V}$ and let $I_{Nd} - W$ be the consensus matrix in other algorithms. Additionally, we let every D_i in SoPro be a diagonal matrix with identical diagonal entries. All the algorithm parameters are hand-optimized as follows: They allow the algorithms that only converge to an inexact solution (i.e., DQN-0 and D-BFGS) to achieve a minimal $\epsilon^k = \frac{1}{N}\sum_{i\in\mathcal{V}} \|x_i^k - x^\star\|^2$ when $k = 250$ and allow the algorithms that converge to the exact optimum (i.e., SoPro, ESOM-0, NRC, and DQM) to reach $\epsilon^k \leq 10^{-6}$ as fast as possible.

Figure 1 plots the error ϵ^k versus the number k of iterations for SoPro, DQN-0, ESOM-0, NRC, DQM, D-BFGS with the above settings. In all the four cases, SoPro converges faster and reaches higher accuracy than other methods. Further, by comparing Figure 1(a) with Figure 1(b), we can observe that the more connected the network is, the faster SoPro, ESOM-0, NRC, DQM converge, while connectivity ratio seems to have little impact on DQN-0 and D-BFGS. By comparing Figure 1(a) with Figure 1(c), we can see that larger convexity parameters may accelerate the convergence of NRC and may smoothen the curve of SoPro. Finally, by comparing Figure 1(a) with 1(d), we demonstrate that SoPro is highly scalable with respect to the network size.

VI. Conclusion

We have developed a distributed second-order proximal algorithm (SoPro) for consensus optimization over undirected networks. It achieves a linear rate of convergence to the optimum under local restricted strong convexity, which is weaker than the global strong convexity required by the existing linearly convergent distributed optimization algorithms. We have also provided an explicit parameter condition to guarantee the linear convergence. In addition, we have demonstrated that SoPro is more efficient than most second-order methods in computation and communication costs. Simulations have validated the competent performance of SoPro.

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*, Princeton, USA, 2005, p. 969974.

[2] A. Beck, A. Nedić, A. Ozdaglar, and M. Teboulle, "An $O(1/k)$ gradient method for network resource allocation problems," *IEEE Transactions on Control of Network Systems*, vol. 1, no. 1, pp. 64–73, 2014.

[3] P. Forero, A. Cano, and G. Giannakis, "Consensus-based distributed support vector machines," *Journal of Machine Learning Research*, vol. 11, no. May, pp. 1663–1707, 2010.

[4] A. Nedić, A. Ozdaglar, and P. A. Parrilo, "Constrained consensus and optimization in multi-agent networks," *IEEE Transactions on Automatic Control*, vol. 55, no. 4, pp. 922–938, 2010.

[5] A. Nedić and A. Olshevsky, "Distributed optimization over time-varying directed graphs," *IEEE Transactions on Automatic Control*, vol. 60, no. 3, pp. 601– 615, 2015.

[6] ——, "Stochastic gradient-push for strongly convex functions on time-varying directed graphs," *IEEE Transactions on Automatic Control*, vol. 61, no. 12, pp. 3936–3947, 2016.

[7] D. Jakovetić, J. Xavier, and J. Moura, "Fast distributed gradient methods," *IEEE Transactions on Automatic Control*, vol. 59, no. 5, pp. 1131–1146, 2014.

[8] W. Shi, Q. Ling, G. Wu, and W. Yin, "EXTRA: an exact first-order algorithm for decentralized consensus optimization," *SIAM Journal on Optimization*, vol. 25, no. 2, pp. 944–966, 2015.

[9] ——, "A proximal gradient algorithm for decentralized composite optimization," *IEEE Transactions on Signal Processing*, vol. 63, no. 22, pp. 6013–6023, 2015.

[10] G. Qu and N. Li, "Harnessing smoothness to accelerate distributed optimization," *IEEE Transactions on Control of Network Systems*, 2017.

[11] ——, "Accelerated distributed Nesterov gradient descent," *arXiv preprint arXiv:1705.07176*, 2017.

[12] D. Jakovetić, "A unification and generalization of exact distributed first order methods," accepted to *IEEE Transactions on Signal and Information Processing over Networks*, 2018.

[13] C. Xi and U. Khan, "DEXTRA: A fast algorithm for optimization over directed graphs," *IEEE Transactions on Automatic Control*, vol. 62, no. 10, pp. 4980–4993, 2017.

[14] C. Xi, R. Xin, , and U. Khan, "ADD-OPT: Accelerated distributed directed optimization," *IEEE Transactions on Automatic Control*, vol. 63, no. 5, pp. 1329–1339, 2018.

[15] C. Xi, V. Mai, R. Xin, E. Abed, , and U. Khan, "Linear convergence in optimization over directed graphs with row-stochastic matrices," *IEEE Transactions on Automatic Control*, vol. 63, no. 10, pp. 3558–3565, 2018.

[16] R. Xin, , and U. Khan, "A linear algorithm for optimization over directed graphs with geometric convergence," *IEEE Control Systems Letter*, vol. 2, no. 3, pp. 315–310, 2018.

[17] J. Duchi, A. Agarwal, and M. Wainwright, "Dual averaging for distributed optimization: Convergence and network scaling," *IEEE Transactions on Automatic Control*, vol. 57, no. 3, pp. 592–606, 2012.

[18] P. Bianchi, W. Hachem, and F. Iutzeler, "A coordinate descent primal-dual algorithm and application to distributed asynchronous optimization," *IEEE Transactions on Automatic Control*, vol. 61, no. 10, pp. 2947–2957, 2016.

[19] A. Makhdoumi and A. Ozdaglar, "Convergence rate of distributed admm over networks," *IEEE Transactions on Automatic Control*, vol. 62, no. 10, pp. 5082–5095, 2017.

[20] C. Shi and G. Yang, "Augmented lagrange algorithms for distributed optimization over multi-agent networks via edge-based method," *Automatica*, vol. 94, pp. 55–62, 2018.

[21] M. Eisen, A. Mokhtari, and A. Ribeiro, "Decentralized quasi-newton methods," *IEEE Transactions on Signal Processing*, vol. 65, no. 10, pp. 2613–2628, 2017.

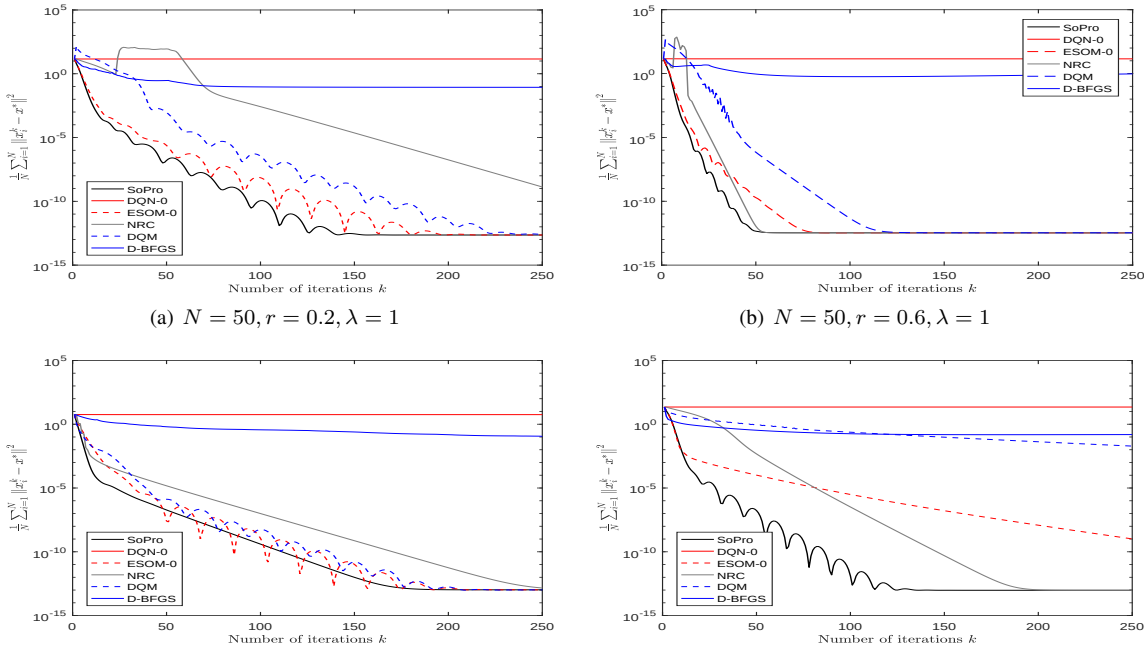
[22] D. Bajović, D. Jakovetić, N. Krejić, and N. K. Jerinikić, "Newton-like method with diagonal correction for distributed optimization," *SIAM Journal on Optimization*, vol. 27, no. 2, pp. 1171–1203, 2017.

[23] A. Mokhtari, Q. Ling, and A. Ribeiro, "Network newton distributed optimization methods," *IEEE Transactions on Signal Processing*, vol. 65, no. 1, pp. 146–161, 2017.

[24] D. Varagnolo, F. Zanella, A. Cenedese, G. Pillonetto, and L. Schenato, "Newton-Raphson consensus for distributed convex optimization," *IEEE Transactions on Automatic Control*, vol. 61, no. 4, pp. 994–1009, 2016.

[25] A. Mokhtari, W. Shi, Q. Ling, and A. Ribeiro, "A decentralized second-order method with exact linear convergence rate for consensus optimization," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 2, no. 4, pp. 507–522, 2016.

[26] ——, "Dqm: Decentralized quadratically approximated alternating direction method of multipliers," *IEEE Transactions on Signal Processing*, vol. 64, no. 19, pp. 5158–5173, 2016.



(a) $N = 50, r = 0.2, \lambda = 1$ (b) $N = 50, r = 0.6, \lambda = 1$

(c) $N = 50, r = 0.2, \lambda = 10$ (d) $N = 200, r = 0.2, \lambda = 1$

Fig. 1. Convergence performance comparison.

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

[28] F. Bach, "Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression," *Journal of Machine Learning Research*, vol. 15, pp. 595–627, 2014.

[29] J.-Y. Chen, G. Pandurangan, and D. Xu, "Robust computation of aggregates in wireless sensor networks: Distributed randomized algorithms and analysis," *IEEE Transactions on Parallel and Distributed Systems*, vol. 17, no. 9, pp. 987–1000, 2006.

[30] B. Mohar, Y. Alavi, G. Chartrand, and O. Oellermann, "The laplacian spectrum of graphs," *Graph theory, combinatorics, and applications*, vol. 2, pp. 871–898, 1991.

APPENDIX

A. Proof of Lemma 1

Since C is compact, so is $\bar{C}$. Due to Assumption 1, $\sum_{i\in\mathcal{V}} f_i(x)$ is restricted strongly convex with respect to $x^\star$ on $\bar{C}$ with convexity parameter $m_{\bar{C}} \in (0,\infty)$. Thus, for any $\mathbf{x} \in C$, since $\frac{1}{N}\sum_{i\in\mathcal{V}} x_i \in \bar{C}$, $\langle \sum_{i\in\mathcal{V}} \nabla f_i(\frac{1}{N}\sum_{i\in\mathcal{V}} x_i) - \nabla f_i(x^\star), \frac{1}{N}\sum_{i\in\mathcal{V}} x_i - x^\star\rangle \geq m_{\bar{C}}\|(\frac{1}{N}\sum_{i\in\mathcal{V}} x_i) - x^\star\|^2$. Then, (12) can be proved in a similar way to [8, Appendix A].

B. Proof of Lemma 2

Define $H^k = \nabla^2 f(\mathbf{x}^k) + D$ for simplicity. We first introduce a lemma to bound the descent of $\|\mathbf{z}^k - \mathbf{z}^\star\|_Q^2$.

Lemma 3. *Suppose Assumption 1 holds. For any $\eta \in (0,1)$, $\beta > 0$, and $k \geq 0$,*

$$\begin{aligned}&\|\mathbf{z}^k - \mathbf{z}^\star\|_Q^2 - \|\mathbf{z}^{k+1} - \mathbf{z}^\star\|_Q^2\\ \geq& 2\rho\eta\langle \mathbf{x}^k - \mathbf{x}^\star, \nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^\star)\rangle - \rho\beta\|\mathbf{x}^k - \mathbf{x}^\star\|^2\\ &-\rho\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2_{\mathcal{A}_{\beta,\eta} - R} + \rho^2\|\mathbf{x}^{k+1}\|_W^2,\end{aligned} \tag{20}$$

where $\mathcal{A}_{\beta,\eta} = \Lambda_M/(2(1-\eta)) + (\Lambda_M - \Lambda_m)^2/(4\beta) + \Lambda_M - \Lambda_m$.

Proof. Using (6) and $W\mathbf{x}^\star = \mathbf{0}_{Nd}$,

$$\begin{aligned}&\langle \mathbf{v}^k - \mathbf{v}^{k+1}, \mathbf{v}^{k+1} - \mathbf{v}^\star\rangle\\ =& -\rho\langle \mathbf{x}^{k+1} - \mathbf{x}^\star, W^{\frac{1}{2}}(\mathbf{v}^{k+1} - \mathbf{v}^\star)\rangle.\end{aligned} \tag{21}$$

From (7) and (6),

$$\begin{aligned}W^{\frac{1}{2}}\mathbf{v}^{k+1} &= W^{\frac{1}{2}}(\mathbf{v}^{k+1} - \mathbf{v}^k) + W^{\frac{1}{2}}\mathbf{v}^k\\ &= (\rho W - H^k)(\mathbf{x}^{k+1} - \mathbf{x}^k) - \nabla f(\mathbf{x}^k).\end{aligned} \tag{22}$$

Combining (21), (22), and (13) results in

$$\begin{aligned}&\langle \mathbf{v}^k - \mathbf{v}^{k+1}, \mathbf{v}^{k+1} - \mathbf{v}^\star\rangle\\ =&\rho\langle \mathbf{x}^{k+1} - \mathbf{x}^\star, \nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^\star)\rangle\\ &+ \rho\langle \mathbf{x}^{k+1} - \mathbf{x}^\star, (H^k - \rho W)(\mathbf{x}^{k+1} - \mathbf{x}^k)\rangle.\end{aligned} \tag{23}$$

In addition,

$$\begin{aligned}&\|\mathbf{z}^k - \mathbf{z}^\star\|_Q^2 - \|\mathbf{z}^{k+1} - \mathbf{z}^\star\|_Q^2 - \|\mathbf{z}^k - \mathbf{z}^{k+1}\|_Q^2\\ =&2\langle \mathbf{v}^k - \mathbf{v}^{k+1}, \mathbf{v}^{k+1} - \mathbf{v}^\star\rangle\\ &+ 2\rho\langle \mathbf{x}^{k+1} - \mathbf{x}^\star, R(\mathbf{x}^k - \mathbf{x}^{k+1})\rangle.\end{aligned} \tag{24}$$

Note that (23) and (24) together lead to

$$\begin{aligned}&\|\mathbf{z}^k - \mathbf{z}^\star\|_Q^2 - \|\mathbf{z}^{k+1} - \mathbf{z}^\star\|_Q^2 - \|\mathbf{z}^k - \mathbf{z}^{k+1}\|_Q^2\\ =&2\rho\langle \mathbf{x}^{k+1} - \mathbf{x}^\star, \nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^\star)\rangle\\ &+ 2\rho\langle \mathbf{x}^{k+1} - \mathbf{x}^\star, (H^k - \rho W - R)(\mathbf{x}^{k+1} - \mathbf{x}^k)\rangle.\end{aligned} \tag{25}$$

Moreover, due to the Lipschitz continuity of each ∇f_i,

$$\langle \mathbf{x}^k - \mathbf{x}^\star, \nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^\star)\rangle \geq \|\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^\star)\|^2_{\Lambda_M^{-1}}.$$

Also, due to the AM-GM inequality,

$$\begin{aligned}&\langle \mathbf{x}^{k+1} - \mathbf{x}^k, \nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^\star)\rangle\\ \geq& -(1-\eta)\|\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^\star)\|^2_{\Lambda_M^{-1}} - \frac{\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2_{\Lambda_M}}{4(1-\eta)}.\end{aligned}$$

It follows that

$$\begin{aligned}&\langle \mathbf{x}^{k+1}-\mathbf{x}^\star, \nabla f(\mathbf{x}^k)-\nabla f(\mathbf{x}^\star)\rangle\\ =&\langle \mathbf{x}^{k}-\mathbf{x}^\star, \nabla f(\mathbf{x}^k)-\nabla f(\mathbf{x}^\star)\rangle\\ &+\langle \mathbf{x}^{k+1}-\mathbf{x}^k, \nabla f(\mathbf{x}^k)-\nabla f(\mathbf{x}^\star)\rangle\\ \geq&\eta\langle \mathbf{x}^{k}-\mathbf{x}^\star, \nabla f(\mathbf{x}^k)-\nabla f(\mathbf{x}^\star)\rangle-\frac{\|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2_{\Lambda_M}}{4(1-\eta)}. \quad (26)\end{aligned}$$

In addition, we have $H^k-\rho W-R=\nabla^2 f(\mathbf{x}^k)-\frac{\Lambda_m+\Lambda_M}{2}=\mathrm{diag}(\nabla^2 f_1(x_1^k)-\frac{(M_1+m_1)I_d}{2},\ldots,\nabla^2 f_N(x_N^k)-\frac{(M_N+m_N)I_d}{2})$ and $\|\nabla^2 f_i(x_i^k)-\frac{(M_i+m_i)I_d}{2}\|\leq\frac{1}{2}(M_i-m_i)\ \forall i\in\mathcal{V}$. Hence,

$$\frac{\Lambda_m-\Lambda_M}{2}\preceq H^k-\rho W-R\preceq\frac{\Lambda_M-\Lambda_m}{2}.$$

Besides, for any $\beta>0$,

$$\begin{aligned}&\langle \mathbf{x}^{k+1}-\mathbf{x}^\star, (H^k-\rho W-R)(\mathbf{x}^{k+1}-\mathbf{x}^k)\rangle\\ =&\langle \mathbf{x}^{k}-\mathbf{x}^\star, (H^k-\rho W-R)(\mathbf{x}^{k+1}-\mathbf{x}^k)\rangle\\ &+\langle \mathbf{x}^{k+1}-\mathbf{x}^k, (H^k-\rho W-R)(\mathbf{x}^{k+1}-\mathbf{x}^k)\rangle\\ \geq&-\frac{\beta}{2}\|\mathbf{x}^k-\mathbf{x}^\star\|^2-\frac{1}{8\beta}\|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2_{(\Lambda_M-\Lambda_m)^2}\\ &-\frac{1}{2}\|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2_{\Lambda_M-\Lambda_m}. \quad (27)\end{aligned}$$

Combining (25), (26) and (27) yields

$$\begin{aligned}&\|\mathbf{z}^k-\mathbf{z}^\star\|^2_Q-\|\mathbf{z}^{k+1}-\mathbf{z}^\star\|^2_Q-\|\mathbf{z}^k-\mathbf{z}^{k+1}\|^2_Q\\ \geq&2\rho\eta\langle \mathbf{x}^k-\mathbf{x}^\star, \nabla f(\mathbf{x}^k)-\nabla f(\mathbf{x}^\star)\rangle-\rho\beta\|\mathbf{x}^k-\mathbf{x}^\star\|^2\\ &-\rho\|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2_{\mathcal{A}_{\beta,\eta}},\end{aligned}$$

which is equivalent to (20). □

Now we show $\mathbf{x}^k\in S_0$, $\forall k\geq 0$ by induction. Clearly, $\mathbf{x}^0\in S_0$. Then, suppose $\mathbf{x}^k\in S_0$ and we want to show that $\mathbf{x}^{k+1}\in S_0$. Adding $\rho^2\|\mathbf{x}^k\|^2_W$ on both sides of (20) gives

$$\begin{aligned}&(\|\mathbf{z}^k-\mathbf{z}^\star\|^2_Q+\rho^2\|\mathbf{x}^k\|^2_W)-(\|\mathbf{z}^{k+1}-\mathbf{z}^\star\|^2_Q+\rho^2\|\mathbf{x}^{k+1}\|^2_W)\\ &\geq 2\rho\eta\langle \mathbf{x}^k-\mathbf{x}^\star, \nabla f(\mathbf{x}^k)-\nabla f(\mathbf{x}^\star)\rangle-\rho\beta\|\mathbf{x}^k-\mathbf{x}^\star\|^2\\ &-\rho\|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2_{\mathcal{A}_{\beta,\eta}-R}+\rho^2\|\mathbf{x}^k\|^2_W.\end{aligned}$$

Moreover, since $\mathbf{x}^k\in S_0$, by the restricted strong convexity of $f(\mathbf{x})+\frac{\rho}{4}\|\mathbf{x}\|^2_W$ on S_0 with respect to $\mathbf{x}^\star$ and $W\mathbf{x}^\star=\mathbf{0}_{Nd}$,

$$\langle \nabla f(\mathbf{x}^k)-\nabla f(\mathbf{x}^\star), \mathbf{x}^k-\mathbf{x}^\star\rangle\geq m_\rho\|\mathbf{x}^k-\mathbf{x}^\star\|^2-\frac{\rho}{2}\|\mathbf{x}^k\|^2_W.$$

Thus,

$$\begin{aligned}&(\|\mathbf{z}^k-\mathbf{z}^\star\|^2_Q+\rho^2\|\mathbf{x}^k\|^2_W)-(\|\mathbf{z}^{k+1}-\mathbf{z}^\star\|^2_Q+\rho^2\|\mathbf{x}^{k+1}\|^2_W)\\ &\geq\rho(2\eta m_\rho-\beta)\|\mathbf{x}^k-\mathbf{x}^\star\|^2+\rho^2(1-\eta)\|\mathbf{x}^k\|^2_W\\ &-\rho\|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2_{\mathcal{A}_{\beta,\eta}-R}. \quad (28)\end{aligned}$$

Notice that when (15) hold, there exists $\beta\leq 2\eta m_\rho$ such that $\mathcal{A}_{\beta,\eta}\preceq R$. This, along with (28), leads to

$$\begin{aligned}&(\|\mathbf{z}^k-\mathbf{z}^\star\|^2_Q+\rho^2\|\mathbf{x}^k\|^2_W)-(\|\mathbf{z}^{k+1}-\mathbf{z}^\star\|^2_Q+\rho^2\|\mathbf{x}^{k+1}\|^2_W)\\ &\geq\rho(2\eta m_\rho-\beta)\|\mathbf{x}^k-\mathbf{x}^\star\|^2+\rho^2(1-\eta)\|\mathbf{x}^k\|^2_W\geq 0,\end{aligned}$$

which implies $\mathbf{x}^{k+1}\in S_0$.

C. Proof of Theorem 1

From the proof of Lemma 2, (28) holds for every $k\geq 0$. Based on this, we derive the following lemma to bound $\|\mathbf{z}^k-\mathbf{z}^\star\|^2_Q$:

Lemma 4. *Suppose Assumptions 1 holds. If there exists $\eta\in(0,1)$ satisfying (15), then for any $c_1,c_2>0$ and $k\geq 0$,*

$$\begin{aligned}&\|\mathbf{z}^k-\mathbf{z}^\star\|^2_Q\\ \leq&\frac{1+c_1}{\lambda_W}\|\mathbf{x}^{k+1}-\mathbf{x}^k\|_{(\Lambda_M+D)^2}+\rho^2(1+\frac{1}{c_1})(1+c_2)\|\mathbf{x}^k\|^2_W\\ &+\|\mathbf{x}^k-\mathbf{x}^\star\|^2_{\frac{(1+1/c_1)(1+1/c_2)\Lambda_M^2}{\lambda_W}+\rho R}, \quad (29)\end{aligned}$$

where λ_W is defined in Lemma 1.

Proof. From (7), we have

$$W^{\frac{1}{2}}\mathbf{v}^k=H^k(\mathbf{x}^k-\mathbf{x}^{k+1})-\rho W\mathbf{x}^k-\nabla f(\mathbf{x}^k),$$

which, together with (3), (14) and (13), implies

$$\begin{aligned}&\|\mathbf{v}^k-\mathbf{v}^\star\|^2=\|(W^\dagger)^{\frac{1}{2}}W^{\frac{1}{2}}(\mathbf{v}^k-\mathbf{v}^\star)\|^2\\ =&\|(W^\dagger)^{\frac{1}{2}}(H^k(\mathbf{x}^k-\mathbf{x}^{k+1})-\rho W\mathbf{x}^k-\nabla f(\mathbf{x}^k)+\nabla f(\mathbf{x}^\star))\|^2\\ \leq&(1+c_1)\|(W^\dagger)^{\frac{1}{2}}H^k(\mathbf{x}^k-\mathbf{x}^{k+1})\|^2+\rho^2(1+\tfrac{1}{c_1})(1+c_2)\|\mathbf{x}^k\|^2_W\\ &+(1+\tfrac{1}{c_1})(1+\tfrac{1}{c_2})\|(W^\dagger)^{\frac{1}{2}}(\nabla f(\mathbf{x}^k)-\nabla f(\mathbf{x}^\star))\|^2\\ \leq&\frac{1+c_1}{\lambda_W}\|\mathbf{x}^{k+1}-\mathbf{x}^k\|_{(H^k)^2}+\rho^2(1+\tfrac{1}{c_1})(1+c_2)\|\mathbf{x}^k\|^2_W\\ &+\frac{(1+1/c_1)(1+1/c_2)}{\lambda_W}\|\mathbf{x}^k-\mathbf{x}^\star\|^2_{\Lambda_M^2}.\end{aligned}$$

In addition, since $(H^k)^2\preceq(\Lambda_M+D)^2$, we obtain (29). □

Note from (28) and Lemma 4 that for any $\delta\in(0,1)$,

$$\begin{aligned}&(1-\delta)(\|\mathbf{z}^k-\mathbf{z}^\star\|^2_Q+\rho^2\|\mathbf{x}^k\|^2_W)-\|\mathbf{z}^{k+1}-\mathbf{z}^\star\|^2_Q-\rho^2\|\mathbf{x}^{k+1}\|^2_W\\ \geq&\lambda_{\min}\left(\rho(2\eta m_\rho-\beta)I_{Nd}-\frac{\delta(1+1/c_1)(1+1/c_2)\Lambda_M^2}{\lambda_W}-\delta\rho R\right)\\ &\cdot\|\mathbf{x}^k-\mathbf{x}^\star\|^2+(\rho^2(1-\eta-\delta)-\delta\rho^2(1+c_2)(1+1/c_1))\|\mathbf{x}^k\|^2_W\\ &-\|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2_{\frac{\delta(1+c_1)}{\lambda_W}(\Lambda_M+D)^2+\rho(\mathcal{A}_{\beta,\eta}-R)}. \quad (30)\end{aligned}$$

To satisfy (16), we require

$$\rho(2\eta m_\rho-\beta)I_{Nd}-\frac{\delta(1+1/c_1)(1+1/c_2)\Lambda_M^2}{\lambda_W}-\delta\rho R\succeq\mathbf{O}_{Nd}, \quad (31)$$

$$\rho^2(1-\eta-\delta)-\delta\rho^2(1+c_2)(1+1/c_1)\geq 0, \quad (32)$$

$$\frac{\delta(1+c_1)}{\lambda_W}(\Lambda_M+D)^2+\rho(\mathcal{A}_{\beta,\eta}-R)\preceq\mathbf{O}_{Nd}. \quad (33)$$

Due to (15), there always exists some $\beta\in(0,2\eta m_\rho)$ such that $\kappa_{\beta,\eta}>0$. Also, since $2\eta m_\rho-\beta>0$, $\eta<1$, and $\mathcal{A}_{\beta,\eta}\prec R$, the largest $\delta\in(0,1)$ satisfying (31), (32) and (33) is given by (17). This completes the proof.

D. Proof of Proposition 1

For each $j\in\mathcal{V}$, let $g_j(x)=f_j(x)+|f_j^\star|$, whose value is nonnegative for all $x\in\mathbb{R}^d$. Then, $\|x_j^0-x^\star\|\leq\mathrm{diam}\{x\in\mathbb{R}^d:\sum_{i\in\mathcal{V}}f_i(x)\leq\sum_{i\in\mathcal{V}}f_i(x_j^0)\}=\mathrm{diam}\{x\in\mathbb{R}^d:\sum_{i\in\mathcal{V}}g_i(x)\leq c_j\}\leq\mathrm{diam}\{x\in\mathbb{R}^d:g_i(x)\leq c_j\}\ \forall i\in\mathcal{V}$.