A Newton Tracking Algorithm with Exact Linear Convergence Rate for Decentralized Consensus Optimization

Jiaojiao Zhang*, Qing Ling†, and Anthony Man-Cho So*

***Abstract*— This paper considers the decentralized consensus optimization problem defined over a network where each node holds a twice continuously differentiable local objective function. Our goal is to minimize the summation of local objective functions and find the exact optimal solution using only local computation and neighboring communications. We propose a novel Newton tracking algorithm, in which each node updates its local variable along a local Newton direction modified with neighboring and historical information. We investigate the connections between the proposed Newton tracking algorithm and several existing methods, including gradient tracking and second-order algorithms. Under the strong convexity assumption, we prove that our proposed algorithm converges to the exact optimal solution at a linear rate. We also present numerical results to demonstrate the efficacy of Newton tracking and validate the theoretical findings.**

I. Introduction

In this paper, we focus on the decentralized consensus optimization problem defined over an undirected and connected network with n nodes, in the form of

$$x^* = \arg\min_{x\in\mathbb{R}^p} \sum_{i=1}^{n} f_i(x). \tag{1}$$

Here, $f_i : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ is a convex and twice continuously differentiable function privately owned by node i. Every agent aims to obtain an optimal solution x^* to (1) via local computation and communication with its neighbors. Decentralized consensus optimization problem in the form of (1) arises in various applications, such as resource allocation [1], [2], smart grid control [3], [4], federate learning [5], [6], decentralized machine learning [7], [8], and so on.

Decentralized consensus optimization methods have been extensively studied in the literature. Among the first-order methods, a popular algorithm is decentralized gradient descent (DGD) [9]–[11]. However, DGD has to employ a diminishing step size to obtain an exact optimal solution. With a fixed step size, DGD converges fast but only to a neighborhood of an optimal solution [11]. There are other first-order methods that use fixed step sizes but still converge to an exact optimal solution, including EXTRA [12], exact diffusion [13], NIDS [14], and gradient tracking [15]–[17]. Take gradient tracking as an example, each node maintains a local estimate of the global gradient descent direction based on neighboring and historical information and uses it to correct the convergence error in DGD.

Although the first-order algorithms enjoy the advantage of low iteration-wise computational complexity, second-order methods are still attractive due to their faster convergence speeds and hence lower communication costs. Some works such as [18]–[20] penalize the implicit consensus constraints in the objective function, thereby allowing the use of unconstrained optimization techniques. However, these penalized second-order algorithms can only converge to a neighborhood of an optimal solution, as the penalty parameter trade-offs the convergence error and convergence speed. Primal-dual methods are effective in handling this accuracy-speed trade-off. This leads to the second-order methods operating in the primal-dual domain [21]–[23], which achieve exact convergence with linear rates. There are other second-order methods with superlinear convergence rates under stricter conditions [24], [25]. For example, the work [24] proposes the distributed averaged quasi-Newton method for a master-slave network, but the initialization is required to be close to an optimal solution so as to guarantee local superlinear convergence. The work [25] runs a finite-time set-consensus inner loop in each iteration of the Polyak's adaptive Newton method and achieves global superlinear convergence.

In this paper, we propose a novel second-order Newton tracking algorithm, in which each node updates its local variable along a local Newton direction modified with neighboring and historical information. As its name suggests, Newton tracking inherits the idea of gradient tracking but improves its convergence speed through the use of second-order information. We investigate the connections between the proposed Newton tracking algorithm and several existing methods, including gradient tracking and second-order algorithms. Under the strong convexity assumption, we prove that our proposed algorithm converges to the exact optimal solution at a linear rate. We also present numerical results to demonstrate the efficacy of Newton tracking and validate the theoretical findings.

Notations. $\mathbf{I} \in \mathbb{R}^{np\times np}$ and $I_n \in \mathbb{R}^{n\times n}$ denote identity matrices of different sizes. $\mathbf{0} \in \mathbb{R}^{np}$ and $0_p \in \mathbb{R}^p$ denote all-zero vectors of different sizes. $1_n \in \mathbb{R}^n$ is the all-one vector. $\lambda_{\max}(\cdot)$, $\lambda_{\min}(\cdot)$ and $\hat{\lambda}_{\min}(\cdot)$ denote the largest, smallest, and smallest nonzero eigenvalues of a matrix, respectively.

II. Problem Formulation and Algorithm Development

In this section, we rewrite the decentralized consensus optimization problem (1) into an equivalent constrained form

Qing Ling is supported in part by NSF China Grants 61573331 and 61973324, and Fundamental Research Funds for the Central Universities.

Jiaojiao Zhang and Anthony Man-Cho So are with Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong.

Qing Ling is with School of Data and Computer Science and Guangdong Province Key Laboratory of Computational Science, Sun Yat-Sen University.

and propose the Newton tracking algorithm to solve it.

A. Problem Formulation

Consider a bidirectionally connected network of n nodes. Two nodes are neighbors if they are connected with an edge. Define $\mathcal{N}_i$ as the set of neighbors of node i and let $x_i \in \mathbb{R}^p$ be the local copy of the decision variable x kept at node i. Since the network is bidirectionally connected, problem (1) is equivalent to

$$\begin{aligned}\{x_i^*\}_{i=1}^n =& \underset{\{x_i\}_{i=1}^n}{\arg\min} \sum_{i=1}^n f_i\left(x_i\right) \\ & \text{s.t. } x_i = x_j,\ \forall j \in \mathcal{N}_i,\ \forall i.\end{aligned} \tag{2}$$

Indeed, the constraints in (2) enforce the consensus condition $x_1 = \cdots = x_n$ for any feasible solution of (2). When the consensus condition is satisfied, the objective functions in (1) and (2) are equivalent, so that the optimal local variables x_i^* of (2) are all equal to the optimal solution x^* of (1), i.e., $x_1^* = \cdots = x_n^* = x^*$.

B. Algorithm Development

Let us introduce a nonnegative mixing matrix $W \in \mathbb{R}^{n\times n}$ whose (i,j)-th element $w_{ij} \geq 0$ represents the weight that node i assigns to node j. The weight $w_{ij} = 0$ if and only if $j \notin \mathcal{N}_i \cup \{i\}$. The mixing matrix W is further required to satisfy the following assumption.

Assumption 1: The mixing matrix W is symmetric and doubly stochastic, i.e., $W = W^T$ and $W1_n = 1_n$. The null space of $I_n - W$ is $\text{span}\,(1_n)$.

When the underlying network is bidirectionally connected, the mixing matrix W satisfying Assumption 1 can be generated using various techniques, such as those introduced in [26]. According to the Perron-Frobenius theorem [27], Assumption 1 means that the eigenvalues of W lie in $(-1, 1]$ and W has a single eigenvalue at 1.

At time t of our proposed Newton tracking algorithm, each node i keeps a local copy $x_i^t \in \mathbb{R}^p$ and a vector $u_i^t \in \mathbb{R}^p$ that estimates the negative Newton direction u^t as

$$u_i^t \approx u^t \triangleq \left(\frac{1}{n}\sum_{i=1}^n \nabla^2 f_i(\bar{x}^t)\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \nabla f_i(\bar{x}^t)\right),$$

where $\bar{x}^t \triangleq \frac{1}{n}\sum_{i=1}^n x_i^t$ is the average of local copies. Each node i updates x_i^{t+1} from x_i^t by descending along the direction $-u_i^t$ with unit step size. Since it is unaffordable to compute the exact Newton direction in a decentralized manner, we propose to estimate the Newton direction by a novel Newton tracking technique.

To be specific, the proposed Newton tracking algorithm starts with $x_i^0 = 0_p$ and $u_i^0 = \left(\nabla^2 f_i(0_p) + \epsilon I_p\right)^{-1}\nabla f_i(0_p)$, then proceeds with

$$x_i^{t+1} = x_i^t - u_i^t, \tag{3}$$

$$\begin{aligned}u_i^{t+1} =& (\nabla^2 f_i(x_i^{t+1}) + \epsilon I_p)^{-1} \\ & \big[(\nabla^2 f_i(x_i^t) + \epsilon I_p)u_i^t + \nabla f_i(x_i^{t+1}) - \nabla f_i(x_i^t) \\ & + 2\alpha(x_i^{t+1} - \sum_{j\in\mathcal{N}_i} w_{ij}x_j^{t+1}) - \alpha(x_i^t - \sum_{j\in\mathcal{N}_i} w_{ij}x_j^t)\big],\end{aligned} \tag{4}$$

where $\epsilon > 0$ and $\alpha > 0$ are parameters. Comparing $-u_i^{t+1}$ with the true Newton direction, we have two observations: (i) The exact global Hessian $\frac{1}{n}\sum_{i=1}^n \nabla^2 f_i(\bar{x}^{t+1})$ is replaced by the regularized local Hessian $\nabla^2 f_i(x_i^{t+1}) + \epsilon I_p$. The regularization parameter ϵ is necessary because the local Hessian $\nabla^2 f_i(x_i^{t+1})$ may be unreliable, especially in the beginning stage of the algorithm. (ii) The exact gradient $\frac{1}{n}\sum_{i=1}^n \nabla f_i(\bar{x}^t)$ is replaced by three terms that are locally computable. The first term $(\nabla^2 f_i(x_i^t) + \epsilon I_p)u_i^t$ involves the previous local Hessian and estimated Newton direction. The second term $\nabla f_i(x_i^{t+1}) - \nabla f_i(x_i^t)$ is the difference between the current and previous gradient directions. The third term $2\alpha(x_i^{t+1} - \sum_{j\in\mathcal{N}_i} w_{ij}x_j^{t+1}) - \alpha(x_i^t - \sum_{j\in\mathcal{N}_i} w_{ij}x_j^t)$ extrapolates the current and previous consensus errors.

Now, we manipulate (4) to better illustrate the idea of Newton tracking. From (4) we have

$$\begin{aligned}& (\nabla^2 f_i(x_i^{t+1}) + \epsilon I_p)u_i^{t+1} \\ =& (\nabla^2 f_i(x_i^t) + \epsilon I_p)u_i^t + \nabla f_i(x_i^{t+1}) - \nabla f_i(x_i^t) \\ & + 2\alpha(x_i^{t+1} - \sum_{j\in\mathcal{N}_i} w_{ij}x_j^{t+1}) - \alpha(x_i^t - \sum_{j\in\mathcal{N}_i} w_{ij}x_j^t).\end{aligned} \tag{5}$$

Summing up (5) over $i = 1, \ldots, n$ and invoking the double stochasticity of W, we have

$$\begin{aligned}& \sum_{i=1}^n \left(\nabla^2 f_i(x_i^{t+1}) + \epsilon I_p\right) u_i^{t+1} \\ =& \sum_{i=1}^n \left(\nabla^2 f_i(x_i^t) + \epsilon I_p\right) u_i^t + \sum_{i=1}^n \left(\nabla f_i(x_i^{t+1}) - \nabla f_i(x_i^t)\right).\end{aligned} \tag{6}$$

When the algorithm is initialized such that $\sum_{i=1}^n \nabla f_i(x_i^0) = \sum_{i=1}^n \left(\nabla^2 f_i(x_i^0) + \epsilon I_p\right) u_i^0$, summing up (6) from time 0 to time t yields

$$\sum_{i=1}^n \left(\nabla^2 f_i(x_i^t) + \epsilon I_p\right) u_i^t = \sum_{i=1}^n \nabla f_i(x_i^t).$$

In comparison, the global Newton direction $-u^t$ satisfies

$$\sum_{i=1}^n \nabla^2 f_i(\bar{x}^t)u^t = \sum_{i=1}^n \nabla f_i(\bar{x}^t).$$

When the local variable pairs (x_i^t, u_i^t) are similar across the nodes, we observe that x_i^t is close to $\bar{x}^t$ and u_i^t tracks a regularized Newton direction.

The recursion (3)–(4) can be written in a compact form. Define $\mathbf{x} \triangleq [x_1; \ldots; x_n] \in \mathbb{R}^{np}$ and $\mathbf{u} \triangleq [u_1; \ldots; u_n] \in \mathbb{R}^{np}$ as the stacks of local variables. Define the aggregate function $f : \mathbb{R}^{np} \to \mathbb{R}$ as $f(\mathbf{x}) = f(x_1, \cdots, x_n) = \sum_{i=1}^n f_i(x_i)$ that sums up all the local functions $f_i\left(x_i\right)$. The gradient of f at $\mathbf{x}$ is $\nabla f(\mathbf{x}) = [\nabla f_1(x_1); \ldots; \nabla f_n(x_n)] \in \mathbb{R}^{np}$. The Hessian of f at $\mathbf{x}$, denoted by $\nabla^2 f(\mathbf{x}) \in \mathbb{R}^{np\times np}$, is a block diagonal matrix whose i-th diagonal block is $\nabla^2 f_i(x)$. Define $\mathbf{H} \triangleq \nabla^2 f(\mathbf{x}) + \epsilon\mathbf{I} \in \mathbb{R}^{np\times np}$ and $\mathbf{W} \triangleq W \otimes I_p \in \mathbb{R}^{np\times np}$ as the Kronecker product of the weight matrix W and the identity

matrix I_p. Then, the recursion (3)–(4) can be written as

$$\mathbf{x}^{t+1} = \mathbf{x}^t - \mathbf{u}^t, \tag{7}$$

$$\mathbf{u}^{t+1} = (\mathbf{H}^{t+1})^{-1}\big[\mathbf{H}^t\mathbf{u}^t + \nabla f\left(\mathbf{x}^{t+1}\right) - \nabla f\left(\mathbf{x}^t\right) + \alpha(\mathbf{I}-\mathbf{W})(2\mathbf{x}^{t+1}-\mathbf{x}^t)\big]. \tag{8}$$

The algorithm is initialized by $\mathbf{x}^0 = \mathbf{0}$ and $\mathbf{u}^0 = (\nabla^2 f(\mathbf{0}) + \epsilon\mathbf{I})^{-1}\nabla f(\mathbf{0})$.

III. Connections with Existing Approaches

This section investigates the connections of the proposed Newton tracking algorithm with several existing approaches, such as gradient tracking and primal-dual methods.

A. Connection with Gradient Tracking

The recursion of gradient tracking [16] is given by

$$\mathbf{x}^{t+1} = \mathbf{W}\mathbf{x}^t - \alpha\mathbf{y}^t, \tag{9}$$

$$\mathbf{y}^{t+1} = \mathbf{W}\mathbf{y}^t + \nabla f(\mathbf{x}^{t+1}) - \nabla f(\mathbf{x}^t), \tag{10}$$

where $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{np}$. To see the connection between gradient tracking and Newton tracking, let us rewrite (9)–(10). First, write $\mathbf{x}^{t+1} = \mathbf{W}\mathbf{x}^t - \alpha\mathbf{y}^t$ as $\mathbf{x}^{t+1} = \mathbf{x}^t - [(\mathbf{I}-\mathbf{W})\mathbf{x}^t + \alpha\mathbf{y}^t]$. Then, define $\mathbf{r}^t = (\mathbf{I}-\mathbf{W})\mathbf{x}^t + \alpha\mathbf{y}^t \in \mathbb{R}^{np}$. Replacing $\mathbf{y}$ with $\mathbf{r}$ shows that (9)–(10) are equivalent to

$$\mathbf{x}^{t+1} = \mathbf{x}^t - \mathbf{r}^t, \tag{11}$$

$$\mathbf{r}^{t+1} = \mathbf{W}\mathbf{r}^t + \alpha[\nabla f(\mathbf{x}^{t+1}) - \nabla f(\mathbf{x}^t)] + (\mathbf{I}-\mathbf{W})(\mathbf{x}^{t+1} - \mathbf{W}\mathbf{x}^t). \tag{12}$$

Similar to the update of $\mathbf{u}^{t+1}$ in (8), the update of $\mathbf{r}^{t+1}$ in (12) also involves three parts: the previous direction $\mathbf{r}^t$, the difference between current and previous gradient directions $\alpha[\nabla f(\mathbf{x}^{t+1}) - \nabla f(\mathbf{x}^t)]$, and the combination of current and previous consensus errors $(\mathbf{I}-\mathbf{W})(\mathbf{x}^{t+1} - \mathbf{W}\mathbf{x}^t)$. The major difference between $\mathbf{u}^{t+1}$ and $\mathbf{r}^{t+1}$ is that the former utilizes the current and previous Hessians, which can improve the convergence speed, especially when the local objective functions are ill-conditioned.

B. Connection with Primal-dual Algorithms

The proposed Newton tracking algorithm has a primal-dual interpretation. Note that the null space of $I_n - W$ is $\operatorname{span}(1_n)$. Thus, the same is true for its square root $(I_n - W)^{\frac{1}{2}}$. Since $(\mathbf{I}-\mathbf{W})^{\frac{1}{2}} = (I_n - W)^{\frac{1}{2}} \otimes I_p$, we have $(\mathbf{I}-\mathbf{W})^{\frac{1}{2}}\mathbf{x} = \mathbf{0}$ if and only if $x_1 = \cdots = x_n$. Thus, problem (2) is equivalent to

$$\mathbf{x}^* \triangleq \arg\min_{\mathbf{x}} \ f(\mathbf{x}) \quad \text{s.t. } (\mathbf{I}-\mathbf{W})^{\frac{1}{2}}\mathbf{x} = \mathbf{0}. \tag{13}$$

The augmented Lagrangian of (13) is

$$L(\mathbf{x}, \mathbf{v}) = f(\mathbf{x}) + \langle \mathbf{v}, (\mathbf{I}-\mathbf{W})^{\frac{1}{2}}\mathbf{x}\rangle + \frac{\alpha}{2}\mathbf{x}^T(\mathbf{I}-\mathbf{W})\mathbf{x}, \tag{14}$$

where $\mathbf{v} \in \mathbb{R}^{np}$ is the dual variable. Therefore, the augmented Lagrangian method to solve (13) is given by

$$\mathbf{x}^{t+1} = \arg\min_{\mathbf{x}} \ L(\mathbf{x}, \mathbf{v}^k), \tag{15}$$

$$\mathbf{v}^{t+1} = \mathbf{v}^t + \alpha(\mathbf{I}-\mathbf{W})^{\frac{1}{2}}\mathbf{x}^{t+1}. \tag{16}$$

Note that solving (15) is nontrivial. First, the update (15) may not admit a closed-form solution when f is a general objective function. Second, even if f is quadratic, the topology-dependent quadratic term $\frac{\alpha}{2}\mathbf{x}^T(\mathbf{I}-\mathbf{W})\mathbf{x}$ makes the closed-form solution not implementable in a decentralized manner. Motivated by these observations, we use a quadratic approximation of f and a linear approximation of $\mathbf{x} \mapsto \frac{\alpha}{2}\mathbf{x}^T(\mathbf{I}-\mathbf{W})\mathbf{x}$ around $\mathbf{x}^t$, and then add a proximal term $\frac{\epsilon}{2}\|\mathbf{x}-\mathbf{x}^t\|^2$ to the objective function of (15). This way, the update of $\mathbf{x}^{t+1}$ is given by the solution of

$$\min_{\mathbf{x}} \ \left\langle \nabla f(\mathbf{x}^t) + (\mathbf{I}-\mathbf{W})^{\frac{1}{2}}\mathbf{v}^t + \alpha(\mathbf{I}-\mathbf{W})\mathbf{x}^t, \mathbf{x}-\mathbf{x}^t \right\rangle + \frac{1}{2}(\mathbf{x}-\mathbf{x}^t)^T\nabla^2 f(\mathbf{x}^t)(\mathbf{x}-\mathbf{x}^t) + \frac{\epsilon}{2}\|\mathbf{x}-\mathbf{x}^t\|^2,$$

which is

$$\mathbf{x}^{t+1} = \mathbf{x}^t - \left(\mathbf{H}^t\right)^{-1}\left[\nabla f\left(\mathbf{x}^t\right) + (\mathbf{I}-\mathbf{W})^{\frac{1}{2}}\mathbf{v}^t + \alpha(\mathbf{I}-\mathbf{W})\mathbf{x}^t\right]. \tag{17}$$

Next, we show that (17)–(16) initialized by $\mathbf{x}^0 = \mathbf{0}$ and $\mathbf{v}^0 = \mathbf{0}$ are equivalent to (7)–(8) initialized by $\mathbf{x}^0 = \mathbf{0}$ and $\mathbf{u}^0 = (\nabla^2 f(\mathbf{0}) + \epsilon\mathbf{I})^{-1}\nabla f(\mathbf{0})$. By (17), the two recursions have the same $\mathbf{x}^1 = -(\nabla^2 f(\mathbf{0}) + \epsilon\mathbf{I})^{-1}\nabla f(\mathbf{0})$. Also by (17), we have

$$\mathbf{H}^t\mathbf{x}^{t+1} = \mathbf{H}^t\mathbf{x}^t - \left[\nabla f\left(\mathbf{x}^t\right) + (\mathbf{I}-\mathbf{W})^{\frac{1}{2}}\mathbf{v}^t + \alpha(\mathbf{I}-\mathbf{W})\mathbf{x}^t\right], \tag{18}$$

$$\mathbf{H}^{t+1}\mathbf{x}^{t+2} = \mathbf{H}^{t+1}\mathbf{x}^{t+1} - \left[\nabla f\left(\mathbf{x}^{t+1}\right) + (\mathbf{I}-\mathbf{W})^{\frac{1}{2}}\mathbf{v}^{t+1} + \alpha(\mathbf{I}-\mathbf{W})\mathbf{x}^{t+1}\right]. \tag{19}$$

Subtracting (18) from (19) and substituting (16) to eliminate the terms $(\mathbf{I}-\mathbf{W})^{\frac{1}{2}}\mathbf{v}^t$ and $(\mathbf{I}-\mathbf{W})^{\frac{1}{2}}\mathbf{v}^{t+1}$, we have

$$\begin{aligned}\mathbf{H}^{t+1}\mathbf{x}^{t+2} - \left[\mathbf{H}^{t+1} - \alpha(\mathbf{I}-\mathbf{W})\right]\mathbf{x}^{t+1} &= \mathbf{H}^t\mathbf{x}^{t+1} - \left[\mathbf{H}^t - \alpha(\mathbf{I}-\mathbf{W})\right]\mathbf{x}^t - \alpha(\mathbf{I}-\mathbf{W})\mathbf{x}^{t+1} \\ &\quad - \left[\nabla f\left(\mathbf{x}^{t+1}\right) - \nabla f\left(\mathbf{x}^t\right)\right].\end{aligned} \tag{20}$$

Define $\mathbf{s}^t \triangleq \mathbf{H}^t\mathbf{x}^{t+1} - [\mathbf{H}^t - \alpha(\mathbf{I}-\mathbf{W})]\mathbf{x}^t$ and rewrite (20) as

$$\mathbf{s}^{t+1} = \mathbf{s}^t - \alpha(\mathbf{I}-\mathbf{W})\mathbf{x}^{t+1} - \left[\nabla f\left(\mathbf{x}^{t+1}\right) - \nabla f\left(\mathbf{x}^t\right)\right]. \tag{21}$$

From the definition of $\mathbf{s}^t$, we have

$$\mathbf{x}^{t+1} = \mathbf{x}^t - (\mathbf{H}^t)^{-1}\left[\alpha(\mathbf{I}-\mathbf{W})\mathbf{x}^t - \mathbf{s}^t\right]. \tag{22}$$

Upon defining $\mathbf{q}^t \triangleq \alpha(\mathbf{I}-\mathbf{W})\mathbf{x}^t - \mathbf{s}^t = \mathbf{H}^t(\mathbf{x}^t - \mathbf{x}^{t+1})$, we rewrite (22) and (21) as

$$\mathbf{x}^{t+1} = \mathbf{x}^t - (\mathbf{H}^t)^{-1}\mathbf{q}^t, \tag{23}$$

$$\mathbf{q}^{t+1} = \mathbf{q}^t + \nabla f\left(\mathbf{x}^{t+1}\right) - \nabla f\left(\mathbf{x}^t\right) + \alpha(\mathbf{I}-\mathbf{W})(2\mathbf{x}^{t+1} - \mathbf{x}^t). \tag{24}$$

Observe that (23)–(24) are equivalent to (7)–(8) in the sense that $\mathbf{u}^t = (\mathbf{H}^t)^{-1}\mathbf{q}^t$.

Remark 1: The primal-dual second-order algorithm ESOM in [22] also uses a quadratic approximation of the augmented Lagrangian when solving (15). However, unlike the proposed Newton tracking algorithm, ESOM

does not linearize the topology-dependent quadratic term $\frac{\alpha}{2}\mathbf{x}^T(\mathbf{I}-\mathbf{W})\mathbf{x}$, which, as we have indicated earlier, makes the closed-form solution not implementable in a decentralized manner. In fact, the primal update of ESOM is given by

$$\begin{aligned}\mathbf{x}^{t+1} =&\mathbf{x}^t - \left(\triangledown^2 f(\mathbf{x}) + \alpha(\mathbf{I}-\mathbf{W}) + \epsilon\mathbf{I}\right)^{-1} \\ &\left[\nabla f\left(\mathbf{x}^t\right) + (\mathbf{I}-\mathbf{W})^{\frac{1}{2}}\mathbf{v}^t + \alpha(\mathbf{I}-\mathbf{W})\mathbf{x}^t\right].\end{aligned} \tag{25}$$

In (25), computing the inverse of $\triangledown^2 f(\mathbf{x}) + \alpha(\mathbf{I}-\mathbf{W}) + \epsilon\mathbf{I}$ requires multiple rounds of communication and computation. Therefore, ESOM introduces an inner loop to approximately solve (25), which leads to extra communication and computation costs [22].

IV. Convergence Analysis

Since the Newton tracking recursion (7)–(8) is equivalent to the primal-dual iteration in (17)–(16), once we show that the primal-dual iteration in (17)–(16) exhibits a linear convergence rate, then so does the Newton tracking recursion (7)–(8). The proofs of the results in this section can be found in the full version of the paper [28]. We begin by stating the following assumption, which will be used in our subsequent development.

Assumption 2: The local objective functions $\{f_i\}_{i=1}^n$ are twice continuously differentiable. The eigenvalues of the Hessians $\{\nabla^2 f_i\}_{i=1}^n$ are bounded by positive constants $\mu_f, L_f \in (0,\infty)$, i.e.

$$\mu_f I_p \preceq \nabla^2 f_i\left(x_i\right) \preceq L_f I_p \tag{26}$$

for all $x_i \in \mathbb{R}^p$ and $i = 1,\dots,n$.

The lower bound in (26) implies that the local objective functions $\{f_i\}_{i=1}^n$ are strongly convex with constant $\mu_f > 0$. The upper bound implies that the local gradients $\{\nabla f_i\}_{i=1}^n$ are Lipschitz continuous with constant L_f. Note that the aggregate Hessian $\nabla^2 f(\mathbf{x})$ is a block diagonal matrix whose i-th diagonal block is $\nabla^2 f_i(x)$. Therefore, the bound on the eigenvalues of the Hessians $\{\nabla^2 f_i\}_{i=1}^n$ in (26) also holds for the aggregate Hessian, i.e.

$$\mu_f\mathbf{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L_f\mathbf{I}$$

for all $\mathbf{x} \in \mathbb{R}^{np}$. Thus, the aggregate objective function f is also strongly convex with constant μ_f and its gradient ∇f is Lipschitz continuous with constant L_f.

Our analysis involves the optimal primal-dual pair $(\mathbf{x}^*, \mathbf{v}^*)$ of (13). According to the KKT condition of (13), we have

$$\triangledown f(\mathbf{x}^*) + (\mathbf{I}-\mathbf{W})^{\frac{1}{2}}\mathbf{v}^* = \mathbf{0}, \tag{27}$$

$$(\mathbf{I}-\mathbf{W})^{\frac{1}{2}}\mathbf{x}^* = 0 \quad \text{or} \quad (\mathbf{I}-\mathbf{W})\mathbf{x}^* = \mathbf{0}. \tag{28}$$

Lemma 1: Consider the equivalent Newton tracking iteration in (17)–(16). The primal-dual iterate satisfies

$$\begin{aligned}&\nabla f\left(\mathbf{x}^{t+1}\right) - \nabla f\left(\mathbf{x}^*\right) + (\mathbf{I}-\mathbf{W})^{\frac{1}{2}}\left(\mathbf{v}^{t+1}-\mathbf{v}^*\right) \\ &\quad + \epsilon\left(\mathbf{x}^{t+1}-\mathbf{x}^t\right) + \mathbf{e}^t = \mathbf{0},\end{aligned} \tag{29}$$

where $\mathbf{e}^t$ is defined as

$$\begin{aligned}\mathbf{e}^t \triangleq&\nabla f\left(\mathbf{x}^t\right) - \nabla f\left(\mathbf{x}^{t+1}\right) + \nabla^2 f\left(\mathbf{x}^t\right)\left(\mathbf{x}^{t+1}-\mathbf{x}^t\right) \\ &- \alpha(\mathbf{I}-\mathbf{W})(\mathbf{x}^{t+1}-\mathbf{x}^t).\end{aligned} \tag{30}$$

The result in Lemma 1 shows the relationship of the primal-dual pairs $(\mathbf{x}^t, \mathbf{v}^t)$ and $(\mathbf{x}^{t+1}, \mathbf{v}^{t+1})$ with the optimal primal-dual pair $(\mathbf{x}^*, \mathbf{v}^*)$. The arguments used in the proof of Lemma 1 are similar to those in [22].

Proof: By the definition of $\mathbf{e}^t$, (17) can be rewritten as

$$\begin{aligned}&\nabla f\left(\mathbf{x}^{t+1}\right) + (\mathbf{I}-\mathbf{W})^{\frac{1}{2}}\mathbf{v}^t + \alpha(\mathbf{I}-\mathbf{W})\mathbf{x}^{t+1} \\ &\quad + \epsilon\left(\mathbf{x}^{t+1}-\mathbf{x}^t\right) + \mathbf{e}^t = \mathbf{0}.\end{aligned} \tag{31}$$

Combining (27) and (28) with (31), we have

$$\begin{aligned}&\nabla f\left(\mathbf{x}^{t+1}\right) - \nabla f\left(\mathbf{x}^*\right) + (\mathbf{I}-\mathbf{W})^{\frac{1}{2}}\left(\mathbf{v}^t-\mathbf{v}^*\right) \\ &\quad + \alpha(\mathbf{I}-\mathbf{W})\left(\mathbf{x}^{t+1}-\mathbf{x}^*\right) + \epsilon\left(\mathbf{x}^{t+1}-\mathbf{x}^t\right) + \mathbf{e}^t = \mathbf{0}.\end{aligned} \tag{32}$$

Observe that $\mathbf{v}^t$ in (32) can be further replaced by $\mathbf{v}^{t+1}$. To be specific, substituting (28) into (16) and then regrouping terms, we know that $\mathbf{v}^t$ can be represented as

$$\mathbf{v}^t = \mathbf{v}^{t+1} - \alpha(\mathbf{I}-\mathbf{W})^{\frac{1}{2}}\left(\mathbf{x}^{t+1}-\mathbf{x}^*\right). \tag{33}$$

Substituting (33) into (32) yields (29). ∎

Observe that the term $\mathbf{e}^t$ can be interpreted as the approximation error at time t. This motivates us to find an upper bound for $\|\mathbf{e}^t\|$. In the following lemma, we provide an upper bound on $\|\mathbf{e}^t\|$ in terms of $\left\|\mathbf{x}^{t+1}-\mathbf{x}^t\right\|$.

Lemma 2: Consider the equivalent Newton tracking iteration in (17)–(16) and recall the definition of the error vector $\mathbf{e}^t$ in (30). If Assumption 2 holds, then $\|\mathbf{e}^t\|$ is bounded by

$$\left\|\mathbf{e}^t\right\| \leq \kappa\left\|\mathbf{x}^{t+1}-\mathbf{x}^t\right\|, \tag{34}$$

where $\kappa \triangleq 2L_f + \alpha\lambda_{\max}(\mathbf{I}-\mathbf{W})$.

Proof: By the triangle inequality, $\|\mathbf{e}^t\|$ is bounded by

$$\begin{aligned}\|\mathbf{e}^t\| \leq& \|\nabla f\left(\mathbf{x}^t\right) - \nabla f\left(\mathbf{x}^{t+1}\right)\| \\ &+ \|\nabla^2 f\left(\mathbf{x}^t\right)\left(\mathbf{x}^{t+1}-\mathbf{x}^t\right)\| + \|\alpha(\mathbf{I}-\mathbf{W})(\mathbf{x}^{t+1}-\mathbf{x}^t)\|.\end{aligned} \tag{35}$$

By Assumption 2, $\|\nabla f\left(\mathbf{x}^t\right) - \nabla f\left(\mathbf{x}^{t+1}\right)\| \leq L_f\|\mathbf{x}^{t+1}-\mathbf{x}^t\|$. As the largest eigenvalue of $\nabla^2 f\left(\mathbf{x}^t\right)$ and $\mathbf{I}-\mathbf{W}$ are L_f and $\lambda_{\max}(\mathbf{I}-\mathbf{W})$, respectively, we know $\|\nabla^2 f\left(\mathbf{x}^t\right)\left(\mathbf{x}^{t+1}-\mathbf{x}^t\right)\| \leq L_f\|\mathbf{x}^{t+1}-\mathbf{x}^t\|$ and $\|\alpha(\mathbf{I}-\mathbf{W})(\mathbf{x}^{t+1}-\mathbf{x}^t)\| \leq \lambda_{\max}(\mathbf{I}-\mathbf{W})\|\mathbf{x}^{t+1}-\mathbf{x}^t\|$. Substituting these inequalities into (35) completes the proof. ∎

The result in (34) implies that the errors $\{\mathbf{e}^t\}_t$ introduced by approximation tend to zero as the sequence of iterates $\{\mathbf{x}^t\}_t$ approaches the optimal solution $\mathbf{x}^*$. This will be shown in Theorem 1.

Given the preliminary results in Lemmas 1 and 2, we are ready to establish the linear convergence of the proposed Newton tracking method. Specifically, we define the vectors $\zeta, \zeta^* \in \mathbb{R}^{2np}$ and matrix $\mathbf{G} \in \mathbb{R}^{np\times np}$ as

$$\zeta^t = \begin{bmatrix}\mathbf{x}^t \\ \mathbf{v}^t\end{bmatrix},\ \zeta^* = \begin{bmatrix}\mathbf{x}^* \\ \mathbf{v}^*\end{bmatrix},\ \mathbf{G} = \begin{bmatrix}\mathbf{Q} & \mathbf{0} \\ \mathbf{0} & \frac{1}{\alpha}\mathbf{I}\end{bmatrix},$$

where $\mathbf{Q} \triangleq \epsilon\mathbf{I} - \alpha(\mathbf{I}-\mathbf{W})$. Note that $\mathbf{Q}$ and hence $\mathbf{G}$ are positive definite when $\epsilon - \alpha\lambda_{\max}(\mathbf{I}-\mathbf{W}) > 0$. Then, we can

show that the sequence $\{\|\zeta^t-\zeta^*\|_{\mathbf{G}}^2\}_t$ converges to zero at a linear rate.

Theorem 1: Consider the equivalent Newton tracking iteration in (17)–(16). Suppose that the parameters ϵ and α satisfy $\lambda_{\min}(\mathbf{Q})=\epsilon-\alpha\lambda_{\max}(\mathbf{I}-\mathbf{W})>\frac{4L_f^2}{\mu_f}$. Then, the sequence $\{\|\zeta^t-\zeta^*\|_{\mathbf{G}}^2\}_t$ satisfies

$$\left\|\zeta^{t+1}-\zeta^*\right\|_{\mathbf{G}}^2\leq\frac{1}{1+\delta'}\left\|\zeta^t-\zeta^*\right\|_{\mathbf{G}}^2, \tag{36}$$

where

$$\delta'=\min\Big\{\frac{\mu_f\delta}{(1+\delta)\left[\epsilon+\frac{\beta\phi L_f^2}{\alpha\hat{\lambda}_{\min}(\mathbf{I}-\mathbf{W})}\right]},\\ \frac{\alpha\delta^2(\epsilon-\alpha\lambda_{\max}(\mathbf{I}-\mathbf{W}))\hat{\lambda}_{\min}(\mathbf{I}-\mathbf{W})}{\frac{\beta\epsilon^2}{(\beta-1)}+\frac{\beta\phi(2L_f+\alpha\lambda_{\max}(\mathbf{I}-\mathbf{W}))^2}{(\phi-1)}}\Big\}, \tag{37}$$

$\beta>1$ and $\phi>1$ are arbitrary constants, and

$$\delta\triangleq 1-\frac{4L_f^2}{\mu_f\lambda_{\min}(\mathbf{Q})}=1-\frac{4L_f^2}{\mu_f(\epsilon-\alpha\lambda_{\max}(\mathbf{I}-\mathbf{W}))}>0.$$

Proof: **Step 1.** By reorganizing (17), we get

$$\epsilon(\mathbf{x}^t-\mathbf{x}^{t+1})+\triangledown^2 f(\mathbf{x}^t)\left(\mathbf{x}^t-\mathbf{x}^{t+1}\right)\\ -\left[\nabla f\left(\mathbf{x}^t\right)+(\mathbf{I}-\mathbf{W})^{\frac{1}{2}}\mathbf{v}^t+\alpha(\mathbf{I}-\mathbf{W})\mathbf{x}^t\right]=\mathbf{0}.$$

Thus, it holds

$$\left\langle\mathbf{x}^*-\mathbf{x}^{t+1},\epsilon(\mathbf{x}^t-\mathbf{x}^{t+1})+\triangledown^2 f(\mathbf{x}^t)\left(\mathbf{x}^t-\mathbf{x}^{t+1}\right)\right.\\ \left.-\left[\nabla f\left(\mathbf{x}^t\right)+(\mathbf{I}-\mathbf{W})^{\frac{1}{2}}\mathbf{v}^t+\alpha(\mathbf{I}-\mathbf{W})\mathbf{x}^t\right]\right\rangle=0. \tag{38}$$

Substituting the dual update $\mathbf{v}^t=\mathbf{v}^{t+1}-\alpha(\mathbf{I}-\mathbf{W})^{\frac{1}{2}}\mathbf{x}^{t+1}$ and regrouping the terms, we can rewrite (38) to

$$\left\langle\mathbf{x}^*-\mathbf{x}^{t+1},\underbrace{(\epsilon\mathbf{I}-\alpha(\mathbf{I}-\mathbf{W}))}_{\triangleq\mathbf{Q}}(\mathbf{x}^t-\mathbf{x}^{t+1})\right\rangle \tag{39}$$
$$-\left\langle\mathbf{x}^*-\mathbf{x}^{t+1},\nabla f\left(\mathbf{x}^t\right)\right\rangle-\left\langle\mathbf{x}^*-\mathbf{x}^{t+1},(\mathbf{I}-\mathbf{W})^{\frac{1}{2}}\mathbf{v}^{t+1}\right\rangle\\ +\left\langle\mathbf{x}^*-\mathbf{x}^{t+1},\triangledown^2 f(\mathbf{x}^t)(\mathbf{x}^t-\mathbf{x}^{t+1})\right\rangle=0.$$

For the first term at the left-hand side of (39), we have

$$\left\langle\mathbf{x}^*-\mathbf{x}^{t+1},\mathbf{Q}(\mathbf{x}^t-\mathbf{x}^{t+1})\right\rangle \tag{40}$$
$$=\frac{1}{2}\big(\|\mathbf{x}^*-\mathbf{x}^{t+1}\|_{\mathbf{Q}}^2+\|\mathbf{x}^t-\mathbf{x}^{t+1}\|_{\mathbf{Q}}^2-\|\mathbf{x}^*-\mathbf{x}^t\|_{\mathbf{Q}}^2\big).$$

For the second term at the left-hand side of (39), according to the μ_f-strong convexity of f, we have

$$\left\langle\mathbf{x}^*-\mathbf{x}^{t+1},\nabla f\left(\mathbf{x}^t\right)\right\rangle \tag{41}$$
$$=\left\langle\mathbf{x}^*-\mathbf{x}^{t+1},\nabla f\left(\mathbf{x}^{t+1}\right)\right\rangle\\ +\left\langle\mathbf{x}^*-\mathbf{x}^{t+1},\nabla f\left(\mathbf{x}^t\right)-\nabla f\left(\mathbf{x}^{t+1}\right)\right\rangle\\ \leq f(\mathbf{x}^*)-f\left(\mathbf{x}^{t+1}\right)-\frac{\mu_f}{2}\left\|\mathbf{x}^*-\mathbf{x}^{t+1}\right\|^2\\ +\left\langle\mathbf{x}^*-\mathbf{x}^{t+1},\nabla f\left(\mathbf{x}^t\right)-\nabla f\left(\mathbf{x}^{t+1}\right)\right\rangle.$$

Substituting (41) and (40) into (39), we get

$$\frac{1}{2}\big(\|\mathbf{x}^*-\mathbf{x}^{t+1}\|_{\mathbf{Q}}^2+\|\mathbf{x}^t-\mathbf{x}^{t+1}\|_{\mathbf{Q}}^2-\|\mathbf{x}^*-\mathbf{x}^t\|_{\mathbf{Q}}^2\big) \tag{42}$$
$$-f(\mathbf{x}^*)+f\left(\mathbf{x}^{t+1}\right)+\frac{\mu_f}{2}\left\|\mathbf{x}^*-\mathbf{x}^{t+1}\right\|^2\\ +\left\langle\mathbf{x}^*-\mathbf{x}^{t+1},\nabla f\left(\mathbf{x}^{t+1}\right)-\nabla f\left(\mathbf{x}^t\right)+\triangledown^2 f(\mathbf{x}^t)(\mathbf{x}^t-\mathbf{x}^{t+1})\right\rangle\\ -\left\langle\mathbf{x}^*-\mathbf{x}^{t+1},(\mathbf{I}-\mathbf{W})^{\frac{1}{2}}\mathbf{v}^{t+1}\right\rangle\leq 0.$$

After being regrouped, (42) becomes

$$\underbrace{f(\mathbf{x}^*)-f\left(\mathbf{x}^{t+1}\right)}_{(i)}+\underbrace{\left\langle\mathbf{x}^*-\mathbf{x}^{t+1},(\mathbf{I}-\mathbf{W})^{\frac{1}{2}}\mathbf{v}^{t+1}\right\rangle}_{(ii)} \tag{43}$$
$$-\frac{1}{2}\big(\|\mathbf{x}^*-\mathbf{x}^{t+1}\|_{\mathbf{Q}}^2-\|\mathbf{x}^*-\mathbf{x}^t\|_{\mathbf{Q}}^2\big)\\ \geq\frac{1}{2}\|\mathbf{x}^t-\mathbf{x}^{t+1}\|_{\mathbf{Q}}^2+\frac{\mu_f}{2}\left\|\mathbf{x}^*-\mathbf{x}^{t+1}\right\|^2\\ +\left\langle\mathbf{x}^*-\mathbf{x}^{t+1},\nabla f\left(\mathbf{x}^{t+1}\right)-\nabla f\left(\mathbf{x}^t\right)+\triangledown^2 f(\mathbf{x}^t)(\mathbf{x}^t-\mathbf{x}^{t+1})\right\rangle.$$

Step 2. We proceed to simplify (43). According to the dual update (16), $\mathbf{v}^{t+1}=\mathbf{v}^t+\alpha(\mathbf{I}-\mathbf{W})^{\frac{1}{2}}\mathbf{x}^{t+1}$ and consequently

$$\left\langle\mathbf{v}^*-\mathbf{v}^{t+1},-(\mathbf{I}-\mathbf{W})^{\frac{1}{2}}\mathbf{x}^{t+1}\right\rangle\\ =\left\langle\mathbf{v}^*-\mathbf{v}^{t+1},\frac{\mathbf{v}^t-\mathbf{v}^{t+1}}{\alpha}\right\rangle\\ =\frac{1}{2\alpha}\left(\|\mathbf{v}^{t+1}-\mathbf{v}^t\|^2-\|\mathbf{v}^*-\mathbf{v}^t\|^2+\|\mathbf{v}^*-\mathbf{v}^{t+1}\|^2\right).$$

Reorganizing the terms, we have

$$\underbrace{\left\langle\mathbf{v}^*,-(\mathbf{I}-\mathbf{W})^{\frac{1}{2}}\mathbf{x}^{t+1}\right\rangle}_{(i')}+\underbrace{\left\langle\mathbf{v}^{t+1},(\mathbf{I}-\mathbf{W})^{\frac{1}{2}}\mathbf{x}^{t+1}\right\rangle}_{(ii')} \tag{44}$$
$$+\frac{1}{2\alpha}\left(\|\mathbf{v}^*-\mathbf{v}^t\|^2-\|\mathbf{v}^*-\mathbf{v}^{t+1}\|^2\right)\\ =\frac{1}{2\alpha}\|\mathbf{v}^{t+1}-\mathbf{v}^t\|^2.$$

Next, we sum up (43) and (44). The summation of (i) and (i') can be simplified as

$$f(\mathbf{x}^*)-f(\mathbf{x}^{t+1})+\left\langle\mathbf{v}^*,-(\mathbf{I}-\mathbf{W})^{\frac{1}{2}}\mathbf{x}^{t+1}\right\rangle \tag{45}$$
$$=\hat{L}(\mathbf{x}^*,\mathbf{v}^*)-\hat{L}(\mathbf{x}^{t+1},\mathbf{v}^*)\leq 0,$$

where $\hat{L}(\mathbf{x},\mathbf{v})=f(\mathbf{x})+\langle\mathbf{v},(\mathbf{I}-\mathbf{W})^{\frac{1}{2}}\mathbf{x}\rangle$ is the Lagrangian of (13). The inequality holds because $(\mathbf{x}^*,\mathbf{v}^*)$ is the saddle point of $\hat{L}(\mathbf{x},\mathbf{v})$. The summation of (ii) and (ii') is

$$\left\langle\mathbf{x}^*-\mathbf{x}^{t+1},(\mathbf{I}-\mathbf{W})^{\frac{1}{2}}\mathbf{v}^{t+1}\right\rangle+\left\langle\mathbf{v}^{t+1},(\mathbf{I}-\mathbf{W})^{\frac{1}{2}}\mathbf{x}^{t+1}\right\rangle\\ =\left\langle\mathbf{x}^*,(\mathbf{I}-\mathbf{W})^{\frac{1}{2}}\mathbf{v}^{t+1}\right\rangle=0. \tag{46}$$

Note that in deriving both (45) and (46), we utilize the consensus condition $(\mathbf{I}-\mathbf{W})^{\frac{1}{2}}\mathbf{x}^*=\mathbf{0}$. With (45) and (46), the summation of (43) and (44) is

$$\frac{1}{2}\left(\|\mathbf{x}^*-\mathbf{x}^t\|_{\mathbf{Q}}^2-\|\mathbf{x}^*-\mathbf{x}^{t+1}\|_{\mathbf{Q}}^2\right) \tag{47}$$
$$+\frac{1}{2\alpha}\left(\|\mathbf{v}^*-\mathbf{v}^t\|^2-\|\mathbf{v}^*-\mathbf{v}^{t+1}\|^2\right)\\ \geq\frac{1}{2}\|\mathbf{x}^t-\mathbf{x}^{t+1}\|_{\mathbf{Q}}^2+\frac{1}{2\alpha}\|\mathbf{v}^{t+1}-\mathbf{v}^t\|^2+\frac{\mu_f}{2}\left\|\mathbf{x}^*-\mathbf{x}^{t+1}\right\|^2\\ +\left\langle\mathbf{x}^*-\mathbf{x}^{t+1},\nabla f\left(\mathbf{x}^{t+1}\right)-\nabla f\left(\mathbf{x}^t\right)+\triangledown^2 f(\mathbf{x}^t)(\mathbf{x}^t-\mathbf{x}^{t+1})\right\rangle.$$

It is the μ_f-strong convexity of f that brings the quadratic term $\frac{\mu_f}{2}\|\mathbf{x}^*-\mathbf{x}^{t+1}\|^2$ in (47), which enables us to establish the linear convergence. Indeed, by Cauchy-Schwarz inequality, for any $\theta>0$ we have

$$\begin{aligned}&\left\langle \mathbf{x}^*-\mathbf{x}^{t+1}, \nabla f\left(\mathbf{x}^{t+1}\right)-\nabla f\left(\mathbf{x}^t\right)+\triangledown^2 f(\mathbf{x}^t)(\mathbf{x}^t-\mathbf{x}^{t+1})\right\rangle\\ &\geq -\frac{1}{\theta}\|\nabla f\left(\mathbf{x}^{t+1}\right)-\nabla f\left(\mathbf{x}^t\right)+\triangledown^2 f(\mathbf{x}^t)(\mathbf{x}^t-\mathbf{x}^{t+1})\|^2\\ &\quad -\theta\|\mathbf{x}^*-\mathbf{x}^{t+1}\|^2. \end{aligned}\tag{48}$$

By Lipschitz continuity of ∇f, it holds

$$\begin{aligned}&-\frac{1}{\theta}\|\nabla f\left(\mathbf{x}^{t+1}\right)-\nabla f\left(\mathbf{x}^t\right)+\triangledown^2 f(\mathbf{x}^t)(\mathbf{x}^t-\mathbf{x}^{t+1})\|^2\\ &\geq -\frac{2}{\theta}\|\nabla f\left(\mathbf{x}^{t+1}\right)-\nabla f\left(\mathbf{x}^t\right)\|^2-\frac{2}{\theta}\|\triangledown^2 f(\mathbf{x}^t)(\mathbf{x}^t-\mathbf{x}^{t+1})\|^2\\ &\geq -\frac{4L_f^2}{\theta}\|\mathbf{x}^t-\mathbf{x}^{t+1}\|^2. \end{aligned}\tag{49}$$

Thus, combining (48) and (49) yields

$$\begin{aligned}&\left\langle \mathbf{x}^*-\mathbf{x}^{t+1}, \nabla f\left(\mathbf{x}^{t+1}\right)-\nabla f\left(\mathbf{x}^t\right)+\triangledown^2 f(\mathbf{x}^t)(\mathbf{x}^t-\mathbf{x}^{t+1})\right\rangle\\ &\geq -\theta\|\mathbf{x}^*-\mathbf{x}^{t+1}\|^2-\frac{4L_f^2}{\theta}\|\mathbf{x}^t-\mathbf{x}^{t+1}\|^2. \end{aligned}\tag{50}$$

substituting (50) into (47), we obtain

$$\begin{aligned}&\|\mathbf{x}^*-\mathbf{x}^t\|_{\mathbf{Q}}^2-\|\mathbf{x}^*-\mathbf{x}^{t+1}\|_{\mathbf{Q}}^2\\ &+\frac{1}{\alpha}\left(\|\mathbf{v}^*-\mathbf{v}^t\|^2-\|\mathbf{v}^*-\mathbf{v}^{t+1}\|^2\right)\\ \geq&\|\mathbf{x}^t-\mathbf{x}^{t+1}\|_{\mathbf{Q}}^2+\frac{1}{\alpha}\|\mathbf{v}^{t+1}-\mathbf{v}^t\|^2+\mu_f\left\|\mathbf{x}^*-\mathbf{x}^{t+1}\right\|^2\\ &-\theta\|\mathbf{x}^*-\mathbf{x}^{t+1}\|^2-\frac{4L_f^2}{\theta}\|\mathbf{x}^t-\mathbf{x}^{t+1}\|^2\\ =&\|\mathbf{x}^t-\mathbf{x}^{t+1}\|^2_{(\mathbf{Q}-\frac{4L_f^2}{\theta}\mathbf{I})}+\frac{1}{\alpha}\|\mathbf{v}^{t+1}-\mathbf{v}^t\|^2\\ &+(\mu_f-\theta)\left\|\mathbf{x}^*-\mathbf{x}^{t+1}\right\|^2. \end{aligned}\tag{51}$$

Step 3. To prove the linear convergence, the parameters in (51) are required to satisfy

$$\begin{cases}\lambda_{\min}(\mathbf{Q})-\dfrac{4L_f^2}{\theta}>0,\\ \mu_f-\theta>0.\end{cases}\tag{52}$$

Hence, (52) is attainable when

$$\delta\triangleq 1-\frac{4L_f^2}{\mu_f\lambda_{\min}(\mathbf{Q})}>0,\tag{53}$$

which holds since $\lambda_{\min}(\mathbf{Q})=\epsilon-\alpha\lambda_{\max}(\mathbf{I}-\mathbf{W})>\frac{4L_f^2}{\mu_f}$ by hypothesis.

When $\delta>0$, then (52) holds true if we choose $\theta=\frac{\mu_f}{1+\delta}$. Substituting this specific θ and the definition of δ, we can rewrite (51) to

$$\begin{aligned}&\|\mathbf{x}^*-\mathbf{x}^t\|_{\mathbf{Q}}^2-\|\mathbf{x}^*-\mathbf{x}^{t+1}\|_{\mathbf{Q}}^2\\ &+\frac{1}{\alpha}\left(\|\mathbf{v}^*-\mathbf{v}^t\|^2-\|\mathbf{v}^*-\mathbf{v}^{t+1}\|^2\right)\\ \geq&\delta^2\lambda_{\min}(\mathbf{Q})\|\mathbf{x}^t-\mathbf{x}^{t+1}\|^2+\frac{1}{\alpha}\|\mathbf{v}^{t+1}-\mathbf{v}^t\|^2\\ &+\frac{\mu_f\delta}{1+\delta}\left\|\mathbf{x}^*-\mathbf{x}^{t+1}\right\|^2. \end{aligned}\tag{54}$$

To establish the linear convergence in (36), we need to show that $\|\zeta^t-\zeta^*\|_{\mathbf{G}}^2-\left\|\zeta^{t+1}-\zeta^*\right\|_{\mathbf{G}}^2\geq\delta'\left\|\zeta^{t+1}-\zeta^*\right\|_{\mathbf{G}}^2$. Given (54), it is enough to show that

$$\begin{aligned}&\frac{\delta'}{\alpha}\left\|\mathbf{v}^{t+1}-\mathbf{v}^*\right\|^2+\delta'\left\|\mathbf{x}^{t+1}-\mathbf{x}^*\right\|_{\mathbf{Q}}^2\\ \leq&\delta^2\lambda_{\min}(\mathbf{Q})\|\mathbf{x}^t-\mathbf{x}^{t+1}\|^2+\frac{1}{\alpha}\|\mathbf{v}^{t+1}-\mathbf{v}^t\|^2\\ &+\frac{\mu_f\delta}{1+\delta}\left\|\mathbf{x}^*-\mathbf{x}^{t+1}\right\|^2. \end{aligned}\tag{55}$$

We proceed to find an upper bound for $\left\|\mathbf{v}^{t+1}-\mathbf{v}^*\right\|^2$ in terms of the summands at the right-hand side of (55). For $\nabla f\left(\mathbf{x}^{t+1}\right)-\nabla f\left(\mathbf{x}^*\right)+(\mathbf{I}-\mathbf{W})^{\frac{1}{2}}\left(\mathbf{v}^{t+1}-\mathbf{v}^*\right)+\epsilon\left(\mathbf{x}^{t+1}-\mathbf{x}^t\right)+\mathbf{e}^t=\mathbf{0}$ in (29), we utilize Cauchy-Schwarz inequality twice to obtain

$$\begin{aligned}\left\|\mathbf{v}^{t+1}-\mathbf{v}^*\right\|_{\mathbf{I}-\mathbf{W}}^2\leq&\frac{\beta\epsilon^2}{\beta-1}\|\mathbf{x}^{t+1}-\mathbf{x}^t\|^2\\ &+\beta\phi\|\nabla f\left(\mathbf{x}^{t+1}\right)-\nabla f\left(\mathbf{x}^*\right)\|^2+\frac{\beta\phi}{\phi-1}\|\mathbf{e}^t\|^2, \end{aligned}\tag{56}$$

where $\beta>1$ and $\phi>1$ are parameters introduced in using Cauchy-Schwarz inequality. By Lipschitz continuity of ∇f, it holds that $\|\nabla f\left(\mathbf{x}^{t+1}\right)-\nabla f\left(\mathbf{x}^*\right)\|^2\leq L_f^2\|\mathbf{x}^{t+1}-\mathbf{x}^*\|^2$. By (35), we have $\|\mathbf{e}^t\|^2\leq\kappa^2\left\|\mathbf{x}^{t+1}-\mathbf{x}^t\right\|^2$. Therefore, (56) implies that

$$\begin{aligned}&\left\|\mathbf{v}^{t+1}-\mathbf{v}^*\right\|_{\mathbf{I}-\mathbf{W}}^2\\ \leq&\left(\frac{\beta\epsilon^2}{\beta-1}+\frac{\beta\phi\kappa^2}{\phi-1}\right)\|\mathbf{x}^{t+1}-\mathbf{x}^t\|^2+\beta\phi L_f^2\|\mathbf{x}^{t+1}-\mathbf{x}^*\|^2. \end{aligned}$$

Further, considering that $\mathbf{v}^{t+1}$ and $\mathbf{v}^*$ both lie in the column space of $(\mathbf{I}-\mathbf{W})^{\frac{1}{2}}$, we have

$$\begin{aligned}&\left\|\mathbf{v}^{t+1}-\mathbf{v}^*\right\|^2\leq\frac{1}{\hat{\lambda}_{\min}(\mathbf{I}-\mathbf{W})}\\ &\left\{\left(\frac{\beta\epsilon^2}{\beta-1}+\frac{\beta\phi\kappa^2}{\phi-1}\right)\|\mathbf{x}^{t+1}-\mathbf{x}^t\|^2+\beta\phi L_f^2\|\mathbf{x}^{t+1}-\mathbf{x}^*\|^2\right\}. \end{aligned}\tag{57}$$

Note that $\hat{\lambda}_{\min}(\mathbf{I}-\mathbf{W})>0$ because $\mathbf{I}-\mathbf{W}\succeq 0$. We also find an upper bound for $\left\|\mathbf{x}^{t+1}-\mathbf{x}^*\right\|_{\mathbf{Q}}^2$ as

$$\left\|\mathbf{x}^{t+1}-\mathbf{x}^*\right\|_{\mathbf{Q}}^2\leq\lambda_{\max}(\mathbf{Q})\left\|\mathbf{x}^{t+1}-\mathbf{x}^*\right\|^2.\tag{58}$$

By substituting the upper bounds in (57) and (58) into (55), we obtain a sufficient condition for (36), given by

$$\begin{aligned}&\lambda_{\max}(\mathbf{Q})\delta'\left\|\mathbf{x}^{t+1}-\mathbf{x}^*\right\|^2+\frac{\delta'}{\alpha\hat{\lambda}_{\min}(\mathbf{I}-\mathbf{W})}\\ &\left\{\left(\frac{\beta\epsilon^2}{\beta-1}+\frac{\beta\phi\kappa^2}{\phi-1}\right)\|\mathbf{x}^{t+1}-\mathbf{x}^t\|^2+\beta\phi L_f^2\|\mathbf{x}^{t+1}-\mathbf{x}^*\|^2\right\}\\ \leq&\delta^2\lambda_{\min}(\mathbf{Q})\|\mathbf{x}^t-\mathbf{x}^{t+1}\|^2+\frac{1}{\alpha}\|\mathbf{v}^{t+1}-\mathbf{v}^t\|^2\\ &+\frac{\mu_f\delta}{1+\delta}\left\|\mathbf{x}^*-\mathbf{x}^{t+1}\right\|^2. \end{aligned}\tag{59}$$

Regrouping the terms, we know that (59) is equivalent to

$$\left(\frac{\mu_f\delta}{1+\delta}-\delta'\lambda_{\max}(\mathbf{Q})-\frac{\delta'\beta\phi L_f^2}{\alpha\hat{\lambda}_{\min}(\mathbf{I}-\mathbf{W})}\right)\left\|\mathbf{x}^{t+1}-\mathbf{x}^*\right\|^2$$
$$+\left(\delta^2\lambda_{\min}(\mathbf{Q})-\frac{\delta'\beta\epsilon^2/(\beta-1)}{\alpha\hat{\lambda}_{\min}(\mathbf{I}-\mathbf{W})}-\frac{\delta'\beta\phi\kappa^2/(\phi-1)}{\alpha\hat{\lambda}_{\min}(\mathbf{I}-\mathbf{W})}\right)$$
$$\left\|\mathbf{x}^{t+1}-\mathbf{x}^t\right\|^2+\frac{1}{\alpha}\left\|\mathbf{v}^{t+1}-\mathbf{v}^t\right\|^2\geq 0. \tag{60}$$

Recall that if (60) is satisfied, then (59) holds, and hence (55) and (36) are also true. To get (60), we need to make sure that the coefficients in (60) are non-negative. Thus, (60) holds if δ' satisfies

$$\delta'\leq\min\left\{\frac{\mu_f\delta}{(1+\delta)\left[\lambda_{\max}(\mathbf{Q})+\frac{\beta\phi L_f^2}{\alpha\hat{\lambda}_{\min}(\mathbf{I}-\mathbf{W})}\right]},\right.$$
$$\left.\frac{\alpha\delta^2\lambda_{\min}(\mathbf{Q})\hat{\lambda}_{\min}(\mathbf{I}-\mathbf{W})}{\frac{\beta\epsilon^2}{(\beta-1)}+\frac{\beta\phi\kappa^2}{(\phi-1)}}\right\}. \tag{61}$$

By the definition of $\mathbf{Q}=\epsilon\mathbf{I}-\alpha(\mathbf{I}-\mathbf{W})$, we have

$$\lambda_{\min}(\mathbf{Q})=\epsilon-\alpha\lambda_{\max}(\mathbf{I}-\mathbf{W})>\frac{4L_f^2}{\mu_f}>0,$$
$$\lambda_{\max}(\mathbf{Q})=\epsilon-\alpha\lambda_{\min}(\mathbf{I}-\mathbf{W})=\epsilon>0.$$

Substituting these connections and the definition of $\kappa=2L_f$ $+\alpha\lambda_{\max}(\mathbf{I}-\mathbf{W})$ to (61), we eventually find the largest δ' to satisfy (61), as in (37). ■

Theorem 1 shows that the sequence $\{\|\zeta^t-\zeta^*\|_{\mathbf{G}}^2\}_t$ converges linearly with the factor $\frac{1}{1+\delta'}$. When $\lambda_{\max}(\mathbf{I}-\mathbf{W})$ increases, both δ and δ' decrease. On the other hand, when $\hat{\lambda}_{\min}(\mathbf{I}-\mathbf{W})$ increases, δ' also increases. These observations indicate the impact of the network topology on the convergence speed of our proposed algorithm.

V. Numerical Experiments

We consider applying our proposed Newton tracking algorithm to solve a decentralized logistic regression problem of the form

$$x^*=\underset{x\in\mathbb{R}^p}{\operatorname{argmin}}\ \frac{\rho}{2}\|x\|^2+\sum_{i=1}^{n}\sum_{j=1}^{m_i}\ln\left(1+\exp\left(-\left(\mathbf{o}_{ij}^Tx\right)\mathbf{p}_{ij}\right)\right),$$

where node i has access to m_i training samples $(\mathbf{o}_{ij},\mathbf{p}_{ij})\in\mathbb{R}^p\times\{-1,+1\}$; $j=1,\ldots,m_i$. We add a regularization term $\frac{\rho}{2}\|x\|^2$ with $\rho>0$ to the loss function to avoid overfitting. In our experiments, we randomly generate the elements in $\mathbf{o}_{ij}$ according to the normal distribution and those in $\mathbf{p}_{ij}$ according to the uniform distribution. We randomly generate $\frac{\tau n(n-1)}{2}$ undirected edges for the network of n nodes, where $\tau\in(0,1]$ is the connectivity ratio. This ensures that the network is connected.

To evaluate the performance of the compared algorithms, the optimal logistic classifier x^* is pre-computed through centralized gradient descent. The performance metric is relative error, defined as $\|\mathbf{x}^t-\mathbf{x}^*\|/\|\mathbf{x}^0-\mathbf{x}^*\|$.

A. Comparison with Existing Methods

We compare Newton tracking with the first-order gradient tracking algorithm in [16] and the second-order algorithms NN-K [18], ESOM-K [22], and DQM [21]. In every iteration of NN-K and ESOM-K, the nodes need to execute a $(K+1)$-round inner loop to compute the inverse of a topology-dependent matrix in the forms of $\alpha\nabla^2 f(\mathbf{x})+(\mathbf{I}-\mathbf{W})$ and $\nabla^2 f(\mathbf{x})+\alpha(\mathbf{I}-\mathbf{W})+\epsilon\mathbf{I}$, respectively.

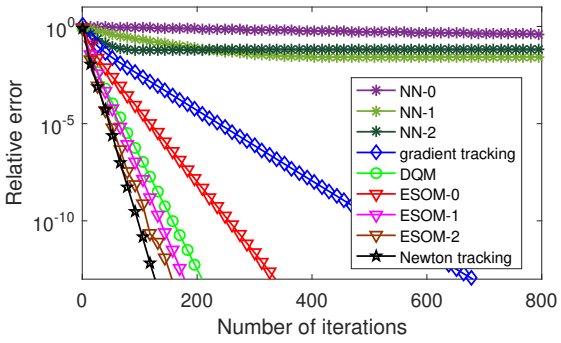

Fig. 1. Relative errors of Newton tracking, gradient tracking, DQM, NN-K, and ESOM-K versus number of iterations.

In the first experiment, we set the number of nodes as $n=10$ and the connectivity ratio as $\tau=0.5$. Each node holds 12 samples, i.e., $m_i=12$, for all i. The dimension of sample vectors $\mathbf{o}_{ij}$ is $p=8$. We set the regularization parameter as $\rho=0.001$.

We run gradient tracking, NN-K, ESOM-K, and DQM with fixed hand-optimized step sizes. The step sizes of gradient tracking and DQM are $\alpha=0.07$ and $\alpha=0.3$, respectively. The parameters of ESOM-0, ESOM-1, and ESOM-2 are $\alpha=3.3$ and $\epsilon=3$. For NN-K, a smaller step size improves accuracy but leads to slow convergence, while a larger step size accelerates the convergence at the cost of low accuracy. Therefore, for NN-0, NN-1, and NN-2 we set $\alpha=0.001$, $\alpha=0.008$, and $\alpha=0.02$, respectively. For Newton tracking, we use the same parameters as ESOM, i.e., $\alpha=3.3$ and $\epsilon=3$. Fig. 1 illustrates the relative error versus

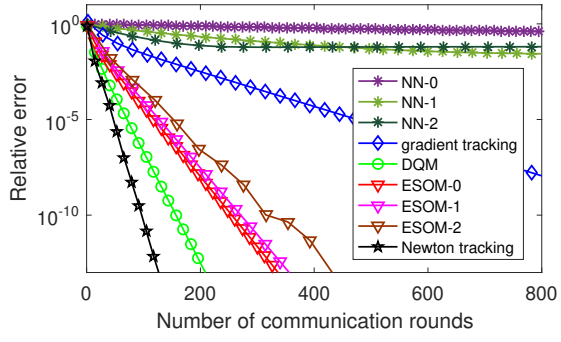

Fig. 2. Relative errors of Newton tracking, gradient tracking, NN-K, ESOM-K, and DQM versus number of communication rounds.

the number of iterations. Observe that NN-K converges to a neighborhood of the optimal solution. Among the exact

decentralized algorithms, the second-order algorithms Newton tracking, ESOM-K, and DQM outperform the first-order gradient tracking algorithm. The proposed Newton tracking algorithm has the best performance compared with the other algorithms and converges linearly, which validates the theoretical result in Theorem 1. Newton tracking and DQM

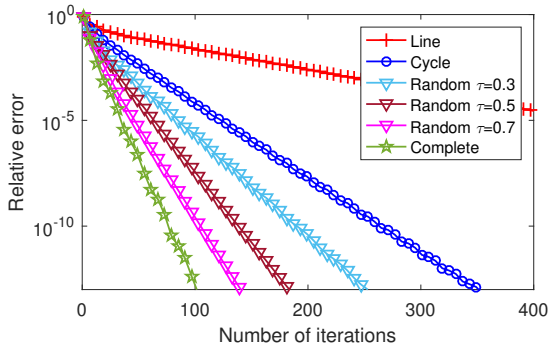

Fig. 3. Relative errors of Newton tracking versus number of iterations for line graph, cycle graph, random graphs with $\tau = \{0.3, 0.5, 0.7\}$, and complete graph.

require one round of communication per iteration. On the other hand, gradient tracking requires two rounds, while NN-K and ESOM-K require $K+1$ rounds. Fig. 2 illustrates the relative error versus the number of communication rounds. Observe that although ESOM-1 and ESOM-2 perform well as depicted in Fig. 1, their performance becomes worse in Fig. 2 because more rounds of communication are required in each iteration. In terms of communication cost, the proposed Newton tracking algorithm is still the best.

B. Effect of Network Topology

This section investigates the performance of Netwon tracking in four different topologies: line graph, cycle graph, random graphs with $\tau = \{0.3, 0.5, 0.7\}$, and complete graph. The parameters of Newton tracking are set as $\alpha = 2.3$ and $\epsilon = 2.4$. All other settings are the same as those in Section V-A.

Fig. 3 illustrates the relative errors versus the number of iterations. Observe that the proposed Newton tracking algorithm has linear convergence rates in all types of graphs. Among them, complete graph yields the fastest speed. This observation confirms the convergence rate developed in Theorem 1. To be specific, for line graph, cycle graph, random graphs with $\tau = \{0.3, 0.5, 0.7\}$, and complete graph, we have $\hat{\lambda}_{\min}(\mathbf{I}-\mathbf{W}) = \{0.03, 0.12, 0.17, 0.34, 0.43, 1.00\}$ and $\lambda_{\max}(\mathbf{I}-\mathbf{W}) = \{1.30, 1.33, 1.16, 1.15, 1.10, 1.00\}$, respectively. Since the complete graph has the largest $\hat{\lambda}_{\min}(\mathbf{I}-\mathbf{W})$ and the smallest $\lambda_{\max}(\mathbf{I}-\mathbf{W})$, it has the largest δ' according to (37). This, together with Thoerem 1, explains why our proposed algorithm has the fastest convergence speed when applied to the complete graph.

VI. Conclusions

In this paper we proposed a novel Newton tracking algorithm to solve the decentralized consensus optimization problem. Each node updates its local variable along a modified local Newton direction, which is calculated using neighboring and historical information. Newton tracking employs a fixed step size and yet can still be proven to converge to an exact solution. The connections between Newton tracking and several existing methods, including gradient tracking and second-order algorithms, were investigated. We proved that the proposed algorithm converges at a linear rate under the strong convexity assumption. Finally, we conducted numerical experiments to demonstrate the efficacy of Newton tracking and its superiority when compared with existing algorithms including gradient tracking, NN-K, ESOM-K, and DQM.

References

[1] Y. Liu, F. R. Yu, X. Li, H. Ji, and V. C. Leung, "Decentralized resource allocation for video transcoding and delivery in blockchain-based system with mobile edge computing," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 11, pp. 11 169–11 185, 2019.

[2] D.-T. Ta, K. Khawam, S. Lahoud, C. Adjih, and S. Martin, "LoRa-MAB: A flexible simulator for decentralized learning resource allocation in IoT networks," in *IFIP Wireless and Mobile Networking Conference*, 2019, pp. 55–62.

[3] E. Dall'Anese, H. Zhu, and G. B. Giannakis, "Distributed optimal power flow for smart microgrids," *IEEE Transactions on Smart Grid*, vol. 4, no. 3, pp. 1464–1475, 2013.

[4] H. J. Liu, W. Shi, and H. Zhu, "Hybrid voltage control in distribution networks under limited communication rates," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 2416–2427, 2019.

[5] A. Lalitha, S. Shekhar, T. Javidi, and F. Koushanfar, "Fully decentralized federated learning," in *Advances in Neural Information Processing Systems Workshop on Bayesian Deep Learning*, 2018.

[6] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *arXiv preprint arXiv:1908.07873*, 2019.

[7] X. Lian, C. Zhang, H. Zhang, C.-J. Hsieh, and J. Liu, "Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent," in *Advances in Neural Information Processing Systems*, 2017.

[8] A. Koppel, S. Paternain, C. Richard, and A. Ribeiro, "Decentralized online learning with kernels," *IEEE Transactions on Signal Processing*, vol. 66, no. 12, pp. 3240–3255, 2018.

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

[10] K. Yuan, Q. Ling, and W. Yin, "On the convergence of decentralized gradient descent," *SIAM Journal on Optimization*, vol. 26, no. 3, pp. 1835–1854, 2016.

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

[12] 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.

[13] K. Yuan, B. Ying, X. Zhao, and A. H. Sayed, "Exact diffusion for distributed optimization and learning – Part I: Algorithm development," *IEEE Transactions on Signal Processing*, vol. 67, no. 3, pp. 708–723, 2018.

[14] Z. Li, W. Shi, and M. Yan, "A decentralized proximal-gradient method with network independent step-sizes and separated convergence rates," *IEEE Transactions on Signal Processing*, vol. 67, no. 17, pp. 4494–4506, 2019.

[15] J. Xu, S. Zhu, Y. C. Soh, and L. Xie, "Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes," in *IEEE Conference on Decision and Control*, 2015, pp. 2055–2060.

[16] G. Qu and N. Li, "Harnessing smoothness to accelerate distributed optimization," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 3, pp. 1245–1260, 2017.

[17] Y. Sun, A. Daneshmand, and G. Scutari, "Convergence rate of distributed optimization algorithms based on gradient tracking," *arXiv preprint arXiv:1905.02637*, 2019.

[18] A. Mokhtari, Q. Ling, and A. Ribeiro, "Network Newton distributed optimization methods," *IEEE Transactions on Signal Processing*, vol. 65, no. 1, pp. 146–161, 2016.

[19] D. Bajovic, D. Jakovetic, N. Krejic, and N. K. Jerinkic, "Newton-like method with diagonal correction for distributed optimization," *SIAM Journal on Optimization*, vol. 27, no. 2, pp. 1171–1203, 2017.

[20] F. Mansoori and E. Wei, "A fast distributed asynchronous Newton-based optimization algorithm," *arXiv preprint arXiv:1901.01872*, 2019.

[21] A. Mokhtari, W. Shi, Q. Ling, and A. Ribeiro, "DQM: Decentralized quadratically approximated alternating direction method of multipliers," *IEEE Transactions on Signal Processing*, vol. 64, no. 19, pp. 5158–5173, 2016.

[22] ——, "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.

[23] M. Eisen, A. Mokhtari, and A. Ribeiro, "A primal-dual quasi-Newton method for exact consensus optimization," *IEEE Transactions on Signal Processing*, vol. 67, no. 23, pp. 5983–5997, 2019.

[24] S. Soori, K. Mischenko, A. Mokhtari, M. M. Dehnavi, and M. Gurbuzbalaban, "DAve-QN: A distributed averaged quasi-Newton method with local superlinear convergence rate," *arXiv preprint arXiv:1906.00506*, 2019.

[25] J. Zhang, K. You, and T. Başar, "Distributed adaptive Newton methods with globally superlinear convergence," *arXiv preprint arXiv:2002.07378*, 2020.

[26] S. Boyd, P. Diaconis, and L. Xiao, "Fastest mixing Markov chain on a graph," *SIAM Review*, vol. 46, no. 4, pp. 667–689, 2004.

[27] S. U. Pillai, T. Suel, and S. Cha, "The Perron-Frobenius theorem: Some of its applications," *IEEE Signal Processing Magazine*, vol. 22, no. 2, pp. 62–75, 2005.

[28] J. Zhang, Q. Ling, and A. M.-C. So, "A Newton tracking algorithm with exact linear convergence rate for decentralized consensus optimization," *arXiv preprint arXiv:2008.10157*, 2020.