Convergence and Applications of a Gossip-based Gauss-Newton Algorithm

Xiao Li, *Student Member, IEEE*, and Anna Scaglione, *Fellow, IEEE*

***Abstract*—The Gauss-Newton algorithm is a popular and efficient centralized method for solving non-linear least squares (NLLS) problems. In this paper, a multi-agent distributed version of this algorithm is proposed to solve general NLLS problems in a network, named Gossip-based Gauss-Newton (GGN) algorithm. Furthermore, we analyze and present sufficient conditions for its convergence and show numerically that the GGN algorithm achieves performance comparable to the centralized algorithm, with graceful degradation in case of network failures. More importantly, the GGN algorithm provides significant performance gains compared to other distributed first order methods.**

***Index Terms*—Gauss-Newton, gossip, distributed, convergence**

I. Introduction

Numerical algorithms for solving non-linear least squares (NLLS) problems are well studied and understood [1]–[3]. Popular methods include the so-called *Newton* and *Gauss-Newton* algorithms. Newton algorithms are second order methods that use the Hessian of the objective function to stabilize and accelerate local convergence [2], [4], while Gauss-Newton simplifies the computation of the Hessian particularly for NLLS problems by ignoring the higher order derivatives [3]. The Gauss-Newton algorithm is commonly used for power systems state estimation [5], localization [6], frequency estimation [7], Kalman filtering [8], medical imaging [9]. Given the fact that the data are acquired over a wide area for some of these problems, in this paper we are interested in the decentralized implementation of the Gauss-Newton algorithm in a network, via *gossiping*. Since their introduction [10], gossip algorithms have been extensively investigated [11], [12], as surveyed in [13]. Deterministic and randomized protocols for gossip algorithms with synchronous or asynchronous updates have been further studied [14], [15] and applied in different areas in networked control and distributed signal processing, such as distributed Kalman filtering [16] or convex optimization problems [17].

Our work is closely related with the recent developments in the area of distributed optimization via *network diffusion*, which evolved from the incremental methods in [18], [19] and gossip-based sub-gradient algorithms in [17] onto fully decentralized and randomized algorithms. The distributed algorithms analyzed in [20]–[24] tackle convex optimization problems through either synchronous or asynchronous communications. These techniques combine a local descent step with a network diffusion step. The convergence of these diffusion algorithms typically requires *convexity* and a diminishing step-size, which results in slow convergence in general [25]. Recently, [26] assumes *local strong convexity* and proposes a diffusion optimization scheme for general convex problems by using stochastic gradients with a constant step-size. Furthermore, the convergence analysis of network diffusion algorithms has also been developed for adaptive formulations using a constant step-size for linear filtering problems [27]–[29], or using a diminishing step-size for invertible non-linear systems [24]. Despite the simplicity of first order methods in diffusion algorithms, they generally suffer from slow convergence in contrast to Newton-type algorithms.

Recently, a gossip-based Newton method was derived in [30] to solve the special case of network utility maximization problems and later applied to power flow estimation [31]. The algorithm relies on the diagonal structure of the Hessian matrix and its convergence is proven under the hypothesis that the error of the computed Newton descent is bounded. In addition, the method is developed specifically for *strictly convex* problems, where furthermore the variables are completely separable for each distributed agent (i.e., its Hessian is block diagonal), while NLLS problems are oftentimes non-convex and non-separable. Although there have been some ad-hoc applications of the Gauss-Newton methods via network average consensus in sensor networks [32]–[34] or incremental methods in acoustic sources localization [35] that relax these assumptions, a thorough study of the algorithm performance in the general case is still missing.

Motivated by this background, in this paper, we propose and study the performance of the Gossip-based Gauss-Newton (GGN) algorithm, for general NLLS problems that are non-separable and non-convex. We also showcase its performance in power system state estimation (PSSE) [36], [37] for system monitoring and control. Recently, the development of distributed PSSE schemes has received considerable attention [38]–[47] to achieve wide area awareness in the expanding power grid. Most of these algorithms hierarchically aggregate the information from distributed control areas under the assumption that there are redundant measurements available at each area to uniquely identify the local state variables (i.e., local observability). Such condition is not required by the GGN algorithm in this paper, similar to the recent works in [48], [49]. In comparison, the proposed GGN algorithm is very different in terms of the network communications and algorithm convergence. The method in [48] is motivated by the diffusion algorithm in [24] (similar to [20] in an adaptive setting), which is a first order sub-gradient method. On the other hand, our approach converges much faster and

This work was supported by the TCIPG project sponsored by Department of Energy under the Award DE-OE0000097.

The authors are with the Department of Electrical and Computer Engineering, University of California, Davis, One Shields Avenue, Kemper Hall, Davis, California 95616-5294 (email : {eceli,ascaglione}@ucdavis.edu).

our communication model is more flexible and robust. The authors in [49] used the *Alternating Direction Method of Multipliers* (ADMM) to distribute the state estimation procedure by decomposing the state variables in different areas so that each agent estimates a local state. This is in contrast to the global state considered in this paper. Furthermore, the communications entailed by ADMM is constrained by the power grid topology, while the communication model considered in this paper is decoupled from the grid topology and more flexible in terms of network reconfigurations for random failures. Also, the numerical tests in [49] are based exclusively on a linear model using Phasor Measurement Unit (PMU) data, while the convergence in general is not discussed.

The challenge associated with PSSE is the presence of multiple stationary points due to the non-convexity of the NLLS objective. This fact confirms the importance of deriving the sufficient conditions for the convergence of the GGN, provided in this paper. These conditions indicate how close the algorithm needs to be initialized around the global minimizer in order to converge to it. The criterion has practical implications in the power grid application, since it can guide the design of measurement placement (see [50]) to improve convergence and performances. In the simulations, we show how the GGN algorithm performs compared to the diffusion algorithms in [48] and [24] in an adaptive setting with streaming data.

Synopsis: In Section II, we define the NLLS problems and provide the distributed NLLS formulation in a network. Then, the proposed GGN algorithm is described in detail in Section III and its convergence analysis follows in Section IV. We formulate the PSSE application in Section V as a NLLS problem and solve it using the proposed GGN algorithm. Finally, the convergence and performance of the GGN algorithm is demonstrated for PSSE in Section VI.

Notation: We denote vectors (matrices) by boldface lower-case (upper-case) symbols, and the set of real (complex) numbers by $\mathbb{R}$ ($\mathbb{C}$). The magnitude of a complex number x is denoted by $|x| = \sqrt{xx^*}$, where x^* is the conjugate of x. The transpose, conjugate transpose, and inverse of a non-singular matrix $\mathbf{X}$ are denoted by $\mathbf{X}^T$, $\mathbf{X}^H$ and $\mathbf{X}^{-1}$, respectively. The inner product between two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{C}^{N\times 1}$ is defined accordingly as $\langle \mathbf{x}, \mathbf{y}\rangle = \sum_{n=1}^N y_n^* x_n$. The $\mathbf{W}$-weighted Euclidean norm of a vector $\mathbf{x}$ is denoted by $\|\mathbf{x}\|_{\mathbf{W}} = \sqrt{\mathbf{x}^H\mathbf{W}\mathbf{x}}$, and the conventional Euclidean norm is written as $\|\mathbf{x}\|$. The 2-norm of a matrix $\mathbf{A}$ is denoted by $\|\mathbf{A}\|$ and the *Frobenius* norm is denoted by $\|\mathbf{A}\|_F$. Given a matrix $\mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_N]$ where $\mathbf{a}_n$ is a column vector, the vectorization operator is defined as $\mathrm{vec}(\mathbf{A}) = [\mathbf{a}_1^T, \cdots, \mathbf{a}_N^T]^T$. Finally. $\mathbf{I}_N$ is an $N \times N$ identity matrix and $\mathbf{1}_N$ is an N-dimensional all 1 vector.

II. Problem Statement

Let $\mathbf{x} \in \mathbb{R}^N$ be an unknown parameter vector associated with a specific network, belonging to a compact convex set $\mathbb{X}$. The network objective is described by a vector-valued continuously differentiable function $\mathbf{g}(\mathbf{x}) = [g_1(\mathbf{x}), \cdots, g_M(\mathbf{x})]^T$ with M outputs, defined as $g_m : \mathbb{R}^N \to \mathbb{R}$, $m = 1, \cdots, M$. Note that $\{g_m\}_{m=1}^M$ are not necessarily convex. Then, a non-linear least squares (NLLS) problem for the network is

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}\in\mathbb{X}} \ \|\mathbf{g}(\mathbf{x})\|^2. \tag{1}$$

Throughout this paper, we assume the following about (1):

Assumption 1.

1) *The vector function is continuous, differentiable, and bounded for* $\mathbf{x} \in \mathbb{X}$ *with*

$$\|\mathbf{g}(\mathbf{x})\| \le \epsilon_{\max}. \tag{2}$$

2) *The* $M \times N$ *Jacobian* $\mathbf{G}(\mathbf{x}) = \partial\mathbf{g}(\mathbf{x})/\partial\mathbf{x}^T$ *is full-column rank for all* $\mathbf{x} \in \mathbb{X}$. *Denote by* $\lambda_{\min}(\cdot)$ *and* $\lambda_{\max}(\cdot)$ *the minimum and maximum eigenvalues and let*

$$\sigma_{\min} = \min_{\mathbf{x}\in\mathbb{X}} \sqrt{\lambda_{\min}\left(\mathbf{G}^T(\mathbf{x})\mathbf{G}(\mathbf{x})\right)},$$
$$\sigma_{\max} = \max_{\mathbf{x}\in\mathbb{X}} \sqrt{\lambda_{\max}\left(\mathbf{G}^T(\mathbf{x})\mathbf{G}(\mathbf{x})\right)},$$

with $0 < \sigma_{\min} \le \sigma_{\max} < \infty$.

3) *The Jacobian* $\mathbf{G}(\mathbf{x})$ *satisfies the Lipschitz condition*

$$\|\mathbf{G}(\mathbf{x}) - \mathbf{G}(\mathbf{x}')\| \le \omega \|\mathbf{x} - \mathbf{x}'\|, \quad \mathbf{x}, \mathbf{x}' \in \mathbb{X},$$

where $\omega > 0$ *is a Lipschitz constant.*

A. Centralized Gauss-Newton Algorithm

When data and functions are available at a central point, the Gauss-Newton method starts from some initial point $\mathbf{x}^0$ and solves the NLLS problem iteratively [3]

$$\mathbf{x}^{k+1} = P_{\mathbb{X}}\left[\mathbf{x}^k - \alpha_k \mathbf{d}^k\right], \quad k = 1, 2, \cdots, \tag{3}$$

where α_k is the step-size in the k-th iteration and $P_{\mathbb{X}}[\cdot]$ is a projection onto the constrained set $\mathbb{X}$. According to Assumption 1, the Gauss-Newton Hessian matrix $\mathbf{G}^T(\mathbf{x}^k)\mathbf{G}(\mathbf{x}^k)$ is positive definite, hence the resulting $\mathbf{d}^k$ constitutes a descent direction of the objective function

$$\mathbf{d}^k = \left[\mathbf{G}^T(\mathbf{x}^k)\mathbf{G}(\mathbf{x}^k)\right]^{-1} \mathbf{G}^T(\mathbf{x}^k)\mathbf{g}(\mathbf{x}^k), \tag{4}$$

where $\mathbf{G}(\mathbf{x})$ is the $M \times N$ Jacobian matrix of $\mathbf{g}(\mathbf{x})$. In this paper, we assume that fixed points always exist for the update (3), which corresponds to the set of the stationary points of the cost function satisfying the first order condition

$$\mathbf{G}^T(\mathbf{x}^\star)\mathbf{g}(\mathbf{x}^\star) = \mathbf{0}, \quad \mathbf{x}^\star \in \mathbb{X}. \tag{5}$$

Note that if α_k is chosen differently at each iteration, the algorithm is called the *damped Gauss-Newton* method while $\alpha_k = \alpha$ corresponds to the *undamped Gauss-Newton* method. Under Assumption 1, it is well-known from [1], [3] that if the step-size α_k is chosen according to the Wolfe condition, the Gauss-Newton iteration converges to a stationary point of the cost function. Since many NLLS problems are non-convex by nature, thus $\widehat{\mathbf{x}}$ is one of the many stationary points $\mathbf{x}^\star$ and the focus in this paper is to study the local convergence property of the algorithm to a particular fixed point $\widehat{\mathbf{x}} \in \mathbb{X}$.

B. Distributed Formulation

Although the convergence of centralized Gauss-Newton algorithms is well studied [3] under Assumption 1, it is not

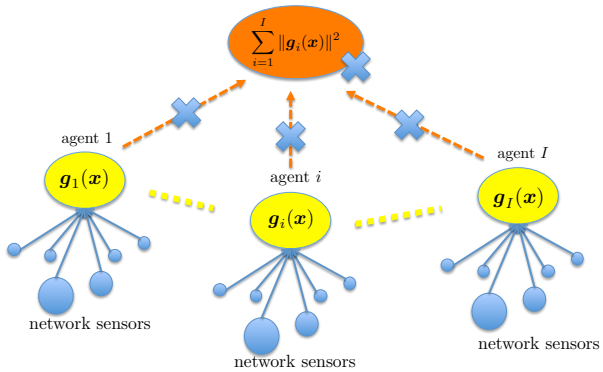


Fig. 1. Schematic of multi-agent computation structure.

immediately clear that similar local convergence properties can be maintained for the decentralized version. As shown in Fig. 1, suppose there are I distributed *agents*, and the i-th *agent* only knows a subset function $\mathbf{g}_i : \mathbb{R}^N \to \mathbb{R}^{M_i}$ from (1), i.e.

$$\mathbf{g}(\mathbf{x}) = [\mathbf{g}_1^T(\mathbf{x}), \ldots, \mathbf{g}_I^T(\mathbf{x})]^T \tag{6}$$

with $M = \sum_{i=1}^{I} M_i$. In this setting, the goal is to obtain

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}\in\mathbb{X}} \sum_{i=1}^{I} \|\mathbf{g}_i(\mathbf{x})\|^2 , \tag{7}$$

where each agent has only partial knowledge of the global cost function. Based on Assumption 1, we have the following results on the distributed formulation.

Corollary 1. *Let Assumption 1 hold. Given that the partial Jacobian $\mathbf{G}_i(\mathbf{x}) = \partial \mathbf{g}_i(\mathbf{x})/\partial \mathbf{x}^T$ is a sub-matrix of the full Jacobian $\mathbf{G}(\mathbf{x})$, then we have (cf. [51, Corollary 3.1.3])*

$$\|\mathbf{G}_i(\mathbf{x}) - \mathbf{G}_i(\mathbf{x}')\| \le \omega \|\mathbf{x} - \mathbf{x}'\| , \quad \mathbf{x}, \mathbf{x}' \in \mathbb{X}.$$

and furthermore the following conditions (cf. [52, Theorem 12.4]) for arbitrary $\mathbf{x}, \mathbf{x}' \in \mathbb{X}$

$$\left\|\mathbf{G}_i^T(\mathbf{x})\mathbf{g}_i(\mathbf{x}) - \mathbf{G}_i^T(\mathbf{x}')\mathbf{g}_i(\mathbf{x}')\right\| \le \nu_s \|\mathbf{x} - \mathbf{x}'\|$$
$$\left\|\mathbf{G}_i^T(\mathbf{x})\mathbf{G}_i(\mathbf{x}) - \mathbf{G}_i^T(\mathbf{x}')\mathbf{G}_i(\mathbf{x}')\right\| \le \nu_S \|\mathbf{x} - \mathbf{x}'\| ,$$

where $\nu_s \ge \omega(\epsilon_{\max} + \sigma_{\max})$ and $\nu_S \ge 2\sigma_{\max}\omega$ are the associated Lipschitz constants.

In the distributed setting, it is difficult to coordinate the step-size at different agents to satisfy the Wolfe condition [1] in a global sense. A variable step-size is also quite inconvenient, because of the difficulties of coordinating a change in the step-size across a network. As a result, we study the *undamped Gauss-Newton* case with a constant step-size $\alpha \in (0, 1]$ i.e.

$$\mathbf{x}_i^{k+1} = P_{\mathbb{X}}\left[\mathbf{x}_i^k - \alpha \mathbf{d}_i^k\right] , \tag{8}$$

where the exact decentralized descent is given by

$$\mathbf{d}_i^k = \left[\mathbf{G}^T(\mathbf{x}_i^k)\mathbf{G}(\mathbf{x}_i^k)\right]^{-1} \mathbf{G}^T(\mathbf{x}_i^k)\mathbf{g}(\mathbf{x}_i^k). \tag{9}$$

According to (9), each agent requires the computation of

$$\mathbf{G}^T(\mathbf{x}_i^k)\mathbf{G}(\mathbf{x}_i^k) = \sum_{j=1}^{I} \mathbf{G}_j^T(\mathbf{x}_i^k)\mathbf{G}_j(\mathbf{x}_i^k) \tag{10}$$

$$\mathbf{G}^T(\mathbf{x}_i^k)\mathbf{g}(\mathbf{x}_i^k) = \sum_{j=1}^{I} \mathbf{G}_j^T(\mathbf{x}_i^k)\mathbf{g}_j(\mathbf{x}_i^k), \tag{11}$$

while the i-th agent has only partial information available to compute $\mathbf{G}_i^T(\mathbf{x}_i^k)\mathbf{G}_i(\mathbf{x}_i^k)$ and $\mathbf{G}_i^T(\mathbf{x}_i^k)\mathbf{g}_i(\mathbf{x}_i^k)$. In the next section, we introduce the GGN algorithm.

III. Gossip-based Gauss-Newton (GGN) Algorithm

The proposed GGN algorithm implements the update in (4) in a fully distributed manner. There are two time scales in the GGN algorithm, one is the time for Gauss-Newton *update* and the other is the gossip *exchange* between every two Gauss-Newton updates. Throughout the rest of the paper, we consistently use *update* (denoted by "k") for the Gauss-Newton algorithm and *exchange* (denoted by "ℓ") for the network gossiping. All the network agents are assumed to have a synchronous clock that determines the time instants τ_k for the k-th algorithm update across the network. Between two updates $[\tau_k, \tau_{k+1})$, the agents exchange information via network gossiping at time $\tau_{k,\ell} \in [\tau_k, \tau_{k+1})$ for $\ell = 1, \cdots, \ell_k$.

Next, we describe the local update model for the GGN algorithm at each distributed agent in Section III-A, and introduce in Section III-B the gossip model for every exchange $\ell = 1, \cdots, \ell_k$ that takes place between every two updates.

A. Local Update Model

Let $\mathbf{x}_i^k$ be the local iterate at the i-th agent after the k-th update. For convenience, let

$$\mathbf{q}(\mathbf{x}_i^k) = \frac{1}{I}\sum_{j=1}^{I} \mathbf{G}_j^T(\mathbf{x}_i^k)\mathbf{g}_j(\mathbf{x}_i^k), \tag{12}$$

$$\mathbf{Q}(\mathbf{x}_i^k) = \frac{1}{I}\sum_{j=1}^{I} \mathbf{G}_j^T(\mathbf{x}_i^k)\mathbf{G}_j(\mathbf{x}_i^k). \tag{13}$$

The "exact descent" in (9), if it were to be computed at the i-th agent for the $(k+1)$-th update, would be

$$\mathbf{d}_i^k = \mathbf{Q}^{-1}(\mathbf{x}_i^k)\mathbf{q}(\mathbf{x}_i^k), \tag{14}$$

which is impossible to obtain in a distributed setting. This is because of the fact that agent j is not aware of the iterate $\mathbf{x}_i^k$ at other agents $i \ne j$ as well as that each node only knows its own mapping $\mathbf{g}_j$ and $\mathbf{G}_j$. In fact, the available information at the i-th agent after the k-th Gauss-Newton update is $\mathbf{G}_i^T(\mathbf{x}_i^k)\mathbf{g}_i(\mathbf{x}_i^k)$ and $\mathbf{G}_i^T(\mathbf{x}_i^k)\mathbf{G}_i(\mathbf{x}_i^k)$. Therefore, we propose to use an average surrogate for $\mathbf{q}(\mathbf{x}_i^k)$ and $\mathbf{Q}(\mathbf{x}_i^k)$

$$\bar{\mathbf{h}}_k = \frac{1}{I}\sum_{i=1}^{I} \mathbf{G}_i^T(\mathbf{x}_i^k)\mathbf{g}_i(\mathbf{x}_i^k), \tag{15}$$

$$\bar{\mathbf{H}}_k = \frac{1}{I}\sum_{i=1}^{I} \mathbf{G}_i^T(\mathbf{x}_i^k)\mathbf{G}_i(\mathbf{x}_i^k), \tag{16}$$

which can be obtained via network gossiping.

After the k-th update by each agent at τ_k, the network enters gossip exchange stage $[\tau_k, \tau_{k+1})$ to compute the surrogate $\bar{\mathbf{h}}_k$ and $\bar{\mathbf{H}}_k$. Define the length-$N_{\mathcal{H}}$ local information vector (i.e., $N_{\mathcal{H}} = N(N+1)$) at the i-th agent for the ℓ-th gossip exchange

$$\boldsymbol{\mathcal{H}}_{k,i}(\ell) = \begin{bmatrix} \mathbf{h}_{k,i}(\ell) \\ \text{vec}\left[\mathbf{H}_{k,i}(\ell)\right] \end{bmatrix}, \tag{17}$$

with the initial condition $\boldsymbol{\mathcal{H}}_{k,i}(0)$ given by

$$\mathbf{h}_{k,i}(0) \triangleq \mathbf{G}_i^T(\mathbf{x}_i^k)\mathbf{g}_i(\mathbf{x}_i^k) \tag{18}$$

$$\mathbf{H}_{k,i}(0) \triangleq \mathbf{G}_i^T(\mathbf{x}_i^k)\mathbf{G}_i(\mathbf{x}_i^k). \tag{19}$$

Clearly, the surrogates are the network averages of the initial conditions $\bar{\mathbf{h}}_k = \sum_{i=1}^I \mathbf{h}_{k,i}(0)/I$ and $\bar{\mathbf{H}}_k = \sum_{i=1}^I \mathbf{H}_{k,i}(0)/I$. Then, all the agents exchange their information $\boldsymbol{\mathcal{H}}_{k,i}(\ell) \rightarrow \boldsymbol{\mathcal{H}}_{k,i}(\ell+1)$ under the protocol described in Section III-B.

After ℓ_k exchanges, the "approximated descent" for the $(k+1)$-th update at the i-th agent is

$$\mathbf{d}_i^k(\ell_k) = \mathbf{H}_{k,i}^{-1}(\ell_k)\mathbf{h}_{k,i}(\ell_k) \tag{20}$$

and the local estimate is updated as

$$\mathbf{x}_i^{k+1} = P_{\mathbb{X}}\left[\mathbf{x}_i^k - \alpha \mathbf{d}_i^k(\ell_k)\right]. \tag{21}$$

B. Network Gossiping Model

Before describing the gossiping protocol, we first model the data exchange between different agents. We use the insights from [10], [20], [53] and impose some rules on the agent communications over time. For each exchange, an agent i communicates with its neighbor agent j during $[\tau_k, \tau_{k+1})$. This is captured by a time-varying network graph $\mathcal{G}_{k,\ell} = (\mathcal{I}, \mathcal{M}_{k,\ell})$ during $[\tau_{k,\ell}, \tau_{k,\ell+1})$ for every GN update k and gossip exchange ℓ. The node set $\mathcal{I} = \{1, \cdots, I\}$ refers to the set of *agents*, and the edge set $\mathcal{M}_{k,\ell}$ is formed by the communication links in that particular gossip exchange. Associated to the graph is the adjacency matrix $\mathbf{A}_k(\ell) = [A_{ij}^{(k,\ell)}]_{I\times I}$

$$A_{ij}^{(k,\ell)} = \begin{cases} 1, & \{i,j\} \in \mathcal{M}_{k,\ell} \\ 0, & \text{otherwise} \end{cases}. \tag{22}$$

Assumption 2. *The composite communication graph $\mathcal{G}_{k,\infty} = \{\mathcal{I}, \mathcal{M}_{k,\infty}\}$ for the k-th update is connected, where*

$$\mathcal{M}_{k,\infty} \triangleq \Big\{\{i,j\} : \{i,j\} \in \mathcal{M}_{k,\ell} \textit{ for infinitely many } \ell\Big\}.$$

There exists an integer $L \geq 1$ such that[1] *for any ℓ*

$$\{i,j\} \in \bigcup_{\ell'=0}^{L-1} \mathcal{M}_{k,\ell+\ell'}, \quad \forall\ \{i,j\} \in \mathcal{M}_{k,\infty}. \tag{23}$$

With the communication model in Assumption 2, each agent combines the information from its neighbors with certain weights. Define a weight matrix $\mathbf{W}_k(\ell) \triangleq [W_{ij}^k(\ell)]_{I\times I}$ for the network topology during $[\tau_{k,\ell}, \tau_{k,\ell+1})$, where the (i,j)-th entry $W_{ij}^k(\ell)$ of the matrix $\mathbf{W}_k(\ell)$ is the weight associated to the edge $\{i,j\}$, which is non-zero if and only if $\{i,j\} \in \mathcal{M}_{k,\ell}$.

[1]This is equivalent to the assumption that within a bounded communication interval of L, every agent pair $\{i,j\}$ in the composite graph communicates with each other at a frequency at least once every L network exchanges.

Assumption 3. *For all k and ℓ, the weight matrix $\mathbf{W}_k(\ell)$ is symmetric and doubly stochastic. There exists a scalar η with $0 < \eta < 1$ such that for all $i, j \in \mathcal{I}$*

1) *$W_{ii}^k(\ell) \geq \eta$ for all $k > 0$ and $\ell > 0$.*
2) *$W_{ij}^k(\ell) \geq \eta$ for all $k > 0$ and $\ell > 0$ if $\{i,j\} \in \mathcal{M}_{k,\ell}$.*
3) *$W_{ij}^k(\ell) = 0$ for all $k > 0$ and $\ell > 0$ if $\{i,j\} \notin \mathcal{M}_{k,\ell}$.*

The gossip exchange of each agent is local with its neighbors using this weight matrix $\mathbf{W}_k(\ell)$. By stacking the local information vectors $\boldsymbol{\mathcal{H}}_k(\ell) \triangleq [\boldsymbol{\mathcal{H}}_{k,1}^T(\ell), \cdots, \boldsymbol{\mathcal{H}}_{k,I}^T(\ell)]^T$, the exchange model can be written compactly as

$$\boldsymbol{\mathcal{H}}_k(\ell) = \left[\mathbf{W}_k(\ell) \otimes \mathbf{I}_{N_{\mathcal{H}}}\right] \boldsymbol{\mathcal{H}}_k(\ell-1), \quad 1 \leq \ell \leq \ell_k, \tag{24}$$

where ℓ_k is number of exchanges $[\tau_k, \tau_{k+1})$.

The gossip exchange model under Assumption 2 and 3 is a general model that includes time-varying network formations, where all agents form random communication links with their neighbors and advance their computations of the average of all local information vectors $\bar{\boldsymbol{\mathcal{H}}}_k = \sum_{i=1}^I \boldsymbol{\mathcal{H}}_{k,i}(0)/I$. With the prescribed communication model, we highlight the following two special cases which are often analyzed in consensus and gossiping literature [10], [13], [15], [17].

1) Coordinated Static Exchange (CSE) [13], [17]: In the CSE protocol, each agent combines the information from possible multiple neighbors, determined by the communication network $\mathbf{A}$, with a static weight matrix $\mathbf{W}$ for all updates and exchanges at $\tau_{k,\ell} \in [\tau_k, \tau_{k+1})$ for $\ell = 1, \cdots, \ell_k$. In particular, if the network is fully connected such that $\mathbf{A} = \mathbf{I}_I - \mathbf{1}_I\mathbf{1}_I^T$, the communication interval is simply $L = 1$ in which each agent talks to everybody in every exchange. There are multiple ways to choose the weight matrix in the CSE protocol, where one of the most popular choice is constructed according to the Laplacian $\mathbf{L} = \text{diag}(\mathbf{A}\mathbf{1}_I) - \mathbf{A}$ as $\mathbf{W} = \mathbf{I}_I - w\mathbf{L}$ with $w = \beta/\max(\mathbf{A}\mathbf{1}_I)$ for some $0 < \beta < 1$.

2) Uncoordinated Random Exchange (URE) [15]: For each exchange in the URE protocol during $[\tau_k, \tau_{k+1})$, a random agent i wakes up and chooses a neighbor agent j at random to communicate. Suppose agent $I_{k,\ell}$ wakes up at $\tau_{k,\ell} \in [\tau_k, \tau_{k+1})$ and $J_{k,\ell}$ is the node picked by node $I_{k,\ell}$ with probability $\gamma_{I_{k,\ell}, J_{k,\ell}}$. Then given some mixing parameter $0 < \beta < 1$, the weight matrix at this time is

$$\mathbf{W}_k(\ell) = \mathbf{I} - \beta\left(\mathbf{e}_{I_{k,\ell}} + \mathbf{e}_{J_{k,\ell}}\right)\left(\mathbf{e}_{I_{k,\ell}} + \mathbf{e}_{J_{k,\ell}}\right)^T, \tag{25}$$

where $\mathbf{e}_i$ is the I-dimensional canonical basis vector with 1 at the i-th entry and 0 otherwise. Note that the URE protocol does not necessarily satisfy Assumption 2, nevertheless numerical simulations indicate that its performance degrade moderately compared to the CSE protocol.

Lemma 1. *[20, Proposition 1] Let Assumption 2 and 3 hold. Given the minimum non-trivial weight η in Assumption 3, the entries of the matrix product $\prod_{\ell'=0}^{\ell} \mathbf{W}_k(\ell')$ converge with a geometric rate uniformly for all $i, j \in \mathcal{I}$ and k*

$$\left|\left[\prod_{\ell'=0}^{\ell} \mathbf{W}_k(\ell')\right]_{ij} - \frac{1}{I}\right| \leq 2\left(\frac{1+\eta^{-L_0}}{1-\eta^{L_0}}\right)\lambda_\eta^\ell, \tag{26}$$

Algorithm 1 Gossip-based Gauss-Newton (GGN) Algorithm

given initial variables $\mathbf{x}_i^0$ at all agents $i \in \mathcal{I}$.
set $k = 0$.
repeat
  set $k = k + 1$.
  initialization: For $i \in \mathcal{I}$, each agent i evaluates (18) and constructs $\boldsymbol{\mathcal{H}}_{k,i}(0)$ as (17);
  network gossiping: Each agent i exchanges information with neighbors via network gossiping as (24).
  local update: For $i \in \mathcal{I}$, each agent i updates its local variables as (21) and (20).
until $\left\|\mathbf{x}_i^{k+1} - \mathbf{x}_i^k\right\| \leq \epsilon$ or $k = K$.
set the local estimate as $\widehat{\mathbf{x}}_i = \mathbf{x}_i^k$.

with $L_0 = (I-1)L$ *and*

$$\lambda_\eta = (1 - \eta^{L_0})^{1/L_0} \in (0, 1). \tag{27}$$

It is clear from Lemma 1 that the limit of the weight matrix product exists $\lim_{\ell \to \infty} \prod_{\ell'=0}^{\ell} \mathbf{W}_k(\ell') = \mathbf{1}_N \mathbf{1}_N^T / I$ and thus

$$\lim_{\ell \to \infty} \boldsymbol{\mathcal{H}}_{k,i}(\ell) = \frac{1}{I} \sum_{i=1}^{I} \boldsymbol{\mathcal{H}}_{k,i}(0), \quad k = 1, 2, \cdots \tag{28}$$

which asymptotically leads to $\lim_{\ell \to \infty} \mathbf{d}_i^k(\ell) = \bar{\mathbf{H}}_k^{-1} \bar{\mathbf{h}}_k$.

IV. Convergence Analysis

In this section, we analyze the convergence of the GGN algorithm (summarized in Algorithm 1) by examining the recursion in (21). Note that the error made in the local descent (20) compared with the exact descent (14) stems from two sources, including the gossiping error resulting from a finite ℓ_k and the mismatch error by using the surrogates $\bar{\mathbf{h}}_k$ and $\bar{\mathbf{H}}_k$ instead of the exact quantities. In the following, we analyze the effect of this error in the convergence of the GGN algorithm.

A. Perturbed Recursion Analysis

At the $(k+1)$-th update, the error between the local estimate $\mathbf{x}_i^{k+1}$ and a fixed point in (5) satisfies the following recursion.

Lemma 2. *Let $\mathbb{X}$ be a compact convex set and Assumption 1 hold. The error $\|\mathbf{x}_i^{k+1} - \widehat{\mathbf{x}}\|$ between the local iterate $\mathbf{x}_i^k$ at each update* (21) *and the global solution $\widehat{\mathbf{x}}$ in* (5) *satisfies the following recursion*

$$\begin{aligned} &\left\|\mathbf{x}_i^{k+1} - \widehat{\mathbf{x}}\right\| \\ &\leq T_1 \left\|\mathbf{x}_i^k - \widehat{\mathbf{x}}\right\|^2 + T_2 \left\|\mathbf{x}_i^k - \widehat{\mathbf{x}}\right\| + \alpha \|\mathbf{d}_i^k(\ell_k) - \mathbf{d}_i^k\|, \end{aligned} \tag{29}$$

where

$$T_1 \triangleq \frac{\alpha\omega}{2\sigma_{\min}}, \quad \epsilon_{\min} \triangleq \|\mathbf{g}(\widehat{\mathbf{x}})\| \tag{30}$$

$$T_2 \triangleq (1-\alpha)\frac{\sigma_{\max}}{\sigma_{\min}} + \frac{\sqrt{2}\alpha\omega\epsilon_{\min}}{\sigma_{\min}^2}. \tag{31}$$

Proof: See Appendix A. ■

The error recursion is a perturbed version of the centralized recursion. The discrepancy between the distributed and centralized update is $\|\mathbf{d}_i^k(\ell_k) - \mathbf{d}_i^k\|$, and its convergence is analyzed in the following theorem.

Theorem 1. *(Convergence with Bounded Perturbation) Let Assumption 1 hold and $\mathbb{X}$ be a compact convex set. Given a step-size chosen as*

$$\max\left\{1 - \frac{3\sigma_{\min}}{\sigma_{\max}}, 0\right\} < \alpha \leq 1 \tag{32}$$

and the condition

$$\omega\epsilon_{\min} < \frac{\sigma_{\min}^2}{\sqrt{2}\alpha}\left[3 - (1-\alpha)\frac{\sigma_{\max}}{\sigma_{\min}}\right], \tag{33}$$

we define the following

$$\rho_{\min} = \frac{(1-T_2) - \sqrt{(1-T_2)^2 - 4\alpha T_1 \kappa}}{2T_1} \tag{34}$$

$$\rho_{\max} = \frac{(1-T_2) + \sqrt{(1-T_2)^2 - 4\alpha T_1 \kappa}}{2T_1} \tag{35}$$

where κ is a bounded perturbation with

$$0 < \kappa < \frac{(1-T_2)^2}{4\alpha T_1}. \tag{36}$$

If the $\|\mathbf{d}_i^k(\ell_k) - \mathbf{d}_i^k\| \leq \kappa$ is bounded for all k and $i \in \mathcal{I}$, then given any $\mathbf{x}_i^0$ that falls within the $\rho_{\max}$-neighborhood of $\widehat{\mathbf{x}} \in \mathbb{X}$

$$\left\|\mathbf{x}_i^0 - \widehat{\mathbf{x}}\right\| < \rho_{\max}, \tag{37}$$

the asymptotic error of the local iterate $\mathbf{x}_i^{k+1}$ at each agent with respect to $\widehat{\mathbf{x}}$ can be bounded as

$$\limsup_{k \to \infty} \left\|\mathbf{x}_i^{k+1} - \widehat{\mathbf{x}}\right\| \leq \rho_{\min}. \tag{38}$$

Proof: See Appendix B. ■

An intuition that can be drawn from the sufficient condition is that the smaller is the Lipschitz constant ω, the larger is the region of convergence around the global solution $\widehat{\mathbf{x}}$ one can start with. In other words, the smoother the cost function the better the convergence. If $\epsilon_{\min}$ in (30) is small (e.g., the fixed point is the minimizer $\widehat{\mathbf{x}}$ in (7)), then by letting $\alpha = 1$ and assuming $\kappa \to 0$, we have $\rho_{\max} \approx 2\sigma_{\min}/\omega - \kappa$ and the steady error is approximately $\rho_{\min} \approx \kappa$ with finite iterations, which scales with the gossiping error. Furthermore, when $\epsilon_{\min} = 0$ the convergence rate is quadratic, same as the Newton's method. Finally, when $\kappa = 0$, the result reduces to the centralized Gauss-Newton algorithm since there is no perturbation.

B. Perturbation Analysis of κ

If the perturbation κ is bounded, Theorem 1 is sufficient to guarantee convergence of the GGN algorithm. In the following, we analyze this perturbation and show that the bounded condition holds.

1) Gossiping error: Define at the ℓ-th exchange

$$\mathbf{h}_k(\ell) \triangleq [\mathbf{h}_{k,1}^T(\ell), \cdots, \mathbf{h}_{k,I}^T(\ell)]^T,$$
$$\mathbf{H}_k(\ell) \triangleq [\mathbf{H}_{k,1}^T(\ell), \cdots, \mathbf{H}_{k,I}^T(\ell)]^T$$

and their deviations from the exact averages $\bar{\mathbf{h}}_k$ and $\bar{\mathbf{H}}_k$ as

$$\mathbf{e}_k(\ell) = [\mathbf{e}_{k,1}^T(\ell), \cdots, \mathbf{e}_{k,I}^T(\ell)]^T,$$
$$\mathbf{E}_k(\ell) = [\mathbf{E}_{k,1}^T(\ell), \cdots, \mathbf{E}_{k,I}^T(\ell)]^T,$$

where $\mathbf{e}_{k,i}(\ell) = \mathbf{h}_{k,i}(\ell) - \bar{\mathbf{h}}_k$ and $\mathbf{E}_{k,i}(\ell) = \mathbf{H}_{k,i}(\ell) - \bar{\mathbf{H}}_k$. The gossip errors $\mathbf{e}_k(\ell_k)$ and $\mathbf{E}_k(\ell_k)$ are related to the properties of the weight matrices $\mathbf{W}_k(\ell)$ in Lemma 1.

Lemma 3. *Let Assumption 2 and 3 hold. The gossip error $\mathbf{e}_k(\ell_k)$ and $\mathbf{E}_k(\ell_k)$ after the k-th update can be bounded as*

$$\|\mathbf{e}_k(\ell_k)\| < C\lambda_\eta^{\ell_k}, \quad \|\mathbf{E}_k(\ell_k)\|_F < C\lambda_\eta^{\ell_k},$$

where

$$C \triangleq 2I\sigma_{\max}\sqrt{I(\epsilon_{\max}^2 + N\sigma_{\max}^2)}\left(\frac{1+\eta^{-L_0}}{1-\eta^{L_0}}\right). \tag{39}$$

Proof: See Appendix C. ■

2) Mismatch of surrogates: Define the errors between the surrogate $\bar{\mathbf{h}}_k$, $\bar{\mathbf{H}}_k$ and exact quantities $\mathbf{q}(\mathbf{x}_i^k)$ and $\mathbf{Q}(\mathbf{x}_i^k)$ as

$$\mathbf{s}_{k,i} = \bar{\mathbf{h}}_k - \mathbf{q}(\mathbf{x}_i^k) = \frac{1}{I}\sum_{j=1}^{I}[\mathbf{h}_{k,i}(\ell) - \mathbf{h}_{k,j}(\ell)] \tag{40}$$
$$\mathbf{S}_{k,i} = \bar{\mathbf{H}}_k - \mathbf{Q}(\mathbf{x}_i^k) = \frac{1}{I}\sum_{j=1}^{I}[\mathbf{H}_{k,i}(\ell) - \mathbf{H}_{k,j}(\ell)],$$

which thus leads to

$$\mathbf{h}_{k,i}(\ell) = \mathbf{q}(\mathbf{x}_i^k) + \mathbf{s}_{k,i} + \mathbf{e}_{k,i}(\ell), \tag{41}$$
$$\mathbf{H}_{k,i}(\ell) = \mathbf{Q}(\mathbf{x}_i^k) + \mathbf{S}_{k,i} + \mathbf{E}_{k,i}(\ell). \tag{42}$$

By (40) and Corollary 1, we have

$$\|\mathbf{s}_{k,i}\| \le \frac{\nu_s}{I}\sum_{j=1}^{I}\left\|\mathbf{x}_i^k - \mathbf{x}_j^k\right\| \tag{43}$$

$$\|\mathbf{S}_{k,i}\| \le \frac{\nu_S}{I}\sum_{j=1}^{I}\left\|\mathbf{x}_i^k - \mathbf{x}_j^k\right\|. \tag{44}$$

Clearly, this discrepancy depends on the disagreement $\left\|\mathbf{x}_i^k - \mathbf{x}_j^k\right\|$ for each pair of i-th and j-th agents, characterized by the mismatch $\mathbf{S}_{k,i}$ and $\mathbf{s}_{k,i}$ which originates from the gossip errors $\mathbf{E}_{k,i}(\ell_k)$ and $\mathbf{e}_{k,i}(\ell_k)$. Having analyzed the gossip error dynamics in Lemma 3, in the following we bound the disagreement $\left\|\mathbf{x}_i^k - \mathbf{x}_j^k\right\|$.

Assumption 4. *Denote by $\ell_{\min} = \min_k\{\ell_k\}$ the minimum exchange. We assume that $\{\ell_k\}_{k=0}^{\infty}$ are chosen to satisfy*[2]

$$\lambda_\infty \triangleq \lim_{K\to\infty}\sum_{k=0}^{K}\lambda_\eta^{(\ell_k - \ell_{\min})} < \infty.$$

For any $\xi \in (0, 1/2)$, the number $\ell_{\min}$ is chosen as

$$\ell_{\min} = \left\lceil \log\left(\frac{\xi}{4D}\right) / \log\lambda_\eta \right\rceil \tag{45}$$
$$D \triangleq CC_2(\nu\lambda_\infty C_1 C_2 + 1) \tag{46}$$

where C, λ_η are defined in (39) (27)*, $\nu = \max\{\nu_s, \nu_S\}$ and*

$$C_1 \triangleq 2\left(1 + \frac{\sigma_{\max}\epsilon_{\max}}{\sigma_{\min}^2}\right), \quad C_2 = \frac{I}{\sigma_{\min}^2} \tag{47}$$

with $\epsilon_{\max}$, $\sigma_{\min}$ and $\sigma_{\max}$ given by Assumption 1.

Lemma 4. *Let the minimum exchange $\ell_{\min}$ be chosen based on an arbitrary value $\xi \in (0, 1/2)$ using* (45)*. According to Lemma 1 under Assumption 1, 2, 3 and 4, then if the initializer is the same for all agents $\mathbf{x}_i^0 = \mathbf{x}^0$, the deviation $\left\|\mathbf{x}_i^K - \mathbf{x}_j^K\right\|$ for any i and j at the K-th update satisfies*

$$\left\|\mathbf{x}_i^K - \mathbf{x}_j^K\right\| \le \xi\left(\frac{CC_1C_2}{D}\right)\sum_{k=0}^{K-1}\lambda_\eta^{(\ell_k - \ell_{\min})}, \tag{48}$$

where C is the gossip error scale in (39)*, C_1, C_2 are defined in* (47) *and λ_η is the gossip convergence rate in* (27)*. Based on Assumption 4, this implies*

$$\left\|\mathbf{x}_i^K - \mathbf{x}_j^K\right\| \le 4CC_1C_2\sum_{k=0}^{K-1}\lambda_\eta^{\ell_k+1}.$$

Proof: See Appendix D. ■

Theorem 2. *Under Assumption 1, 2, 3 and 4, Given Lemma 1, 3 and 4, the discrepancy between the inexact and the exact descent is bounded for all i and k*

$$\left\|\mathbf{d}_i^k(\ell_k) - \mathbf{d}_i^k\right\| < \kappa \tag{49}$$

by the finite perturbation $\kappa \triangleq 4C_1D\lambda_\eta^{(\ell_{\min}+1)}$, whose magnitude vanishes exponentially with respect to the minimum number of gossip exchanges $\lim_{\ell_{\min}\to\infty}\kappa = 0$.

Proof: See Appendix E. ■

Theorem 1 and 2 indicate that if the exchanges ℓ_k's are large, then $\kappa \to 0$ and the recursion approaches the centralized version. Note that Theorem 1 and 2 are proven using very pessimistic bounds. In Section VI the numerical results show the algorithm behaves well even with link failures, in spite of not meeting all the conditions and assumption stated.

V. Application : Power System State Estimation

A power network is characterized by vertices (called *buses*) representing simple interconnections, generators or loads, denoted by the set $\mathcal{N} \triangleq \{1, \cdots, N\}$. Transmission lines connecting these buses constitute the power grid topology, denoted by the edge set $\mathcal{E}$ with cardinality $|\mathcal{E}| = L$. The electrical parameters of the grid are characterized by the admittance matrix $\mathbf{Y} = [-Y_{nm}]_{N\times N}$, where $Y_{nm} = G_{nm} + \mathrm{i}B_{nm}$, $\{n, m\} \in \mathcal{E}$ is the line admittance, and the shunt admittance $\bar{Y}_{nm} = \bar{G}_{nm} + \mathrm{i}\bar{B}_{nm}$ associated with the Π-model[3] of each transmission line $\{n, m\} \in \mathcal{E}$. Note that $Y_{nn} = -\sum_{l\neq n}(\bar{Y}_{nm} + Y_{nm})$ is defined as the self-admittance. The state of the power system corresponds to the voltage phasors at all buses, described by the voltage phase and magnitude $\mathbf{x} = [\mathbf{\Theta}^T, \mathbf{V}^T]^T$, where $\mathbf{\Theta} \triangleq [\theta_1, \cdots, \theta_N]^T$ is the phase vector with θ_1 being the slack bus phase, and $\mathbf{V} \triangleq [V_1, \cdots, V_N]^T$ contains the magnitude.

[2] A simple choice is $\ell_0 = \ell_{\min}$ and $\ell_k = \ell_{k-1} + 1$, then $\lambda_\infty = 1/(1-\lambda_\eta)$.

[3] The Π-model is a circuit equivalent of a transmission line by abstracting two electric buses as a two-port network in the shape of a Π connection [54].

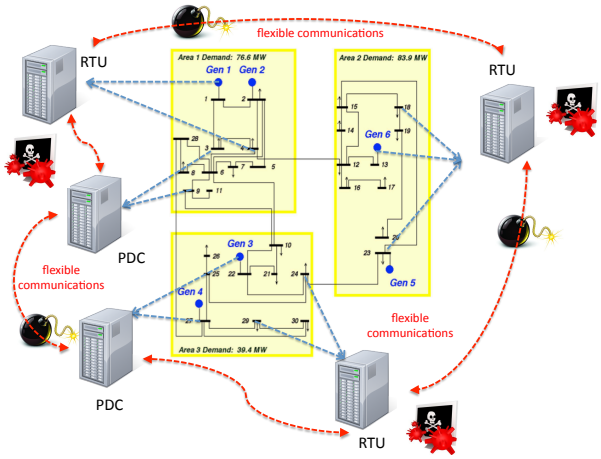


Fig. 2. Multi-site structure in IEEE-30 test case

A. Power Measurement Models

Power measurements include the *active/reactive power injection* (P_n, Q_n) for buses $n \in \mathcal{N}$, and the *active/reactive power flows* (P_{nm}, Q_{nm}) on transmission lines $(n,m) \in \mathcal{E}$. The ensemble of these quantities can be stacked into the length-$2N$ power injection vector $\mathbf{f}_{\mathcal{I}}(\mathbf{x})$, as well as the length-$4L$ line flow vector $\mathbf{f}_{\mathcal{F}}(\mathbf{x})$ respectively

$$\mathbf{f}_{\mathcal{I}}(\mathbf{x}) = [P_1(\mathbf{x}), \cdots, P_N(\mathbf{x}), Q_1(\mathbf{x}), \cdots, Q_N(\mathbf{x})]^T \quad (50)$$

$$\mathbf{f}_{\mathcal{F}}(\mathbf{x}) = [\cdots, P_{nm}(\mathbf{x}), \cdots, \cdots, Q_{nm}(\mathbf{x}), \cdots]^T, \quad (51)$$

which can be expressed in relation to the state $\mathbf{x}$ as in [54]

$$P_n(\mathbf{x}) = V_n \sum_{m \neq n}^{N} V_m \left(G_{nm} \cos\theta_{nm} + B_{nm} \sin\theta_{nm}\right)$$

$$Q_n(\mathbf{x}) = V_n \sum_{m \neq n}^{N} V_m \left(G_{nm} \sin\theta_{nm} - B_{nm} \cos\theta_{nm}\right)$$

$$P_{nm}(\mathbf{x}) = V_n^2 G_{nm} - V_n V_m \left(G_{nm} \cos\theta_{nm} + B_{nm} \sin\theta_{nm}\right)$$

$$Q_{nm}(\mathbf{x}) = -V_n^2 B_{nm} - V_n V_m \left(G_{nm} \sin\theta_{nm} - B_{nm} \cos\theta_{nm}\right),$$

where $\theta_{nm} = \theta_n - \theta_m$. By stacking the power flow equations and the measurements into vectors $\mathbf{f}(\mathbf{x}) \triangleq [\mathbf{f}_{\mathcal{I}}^T(\mathbf{x}), \mathbf{f}_{\mathcal{F}}^T(\mathbf{x})]^T$ and $\mathbf{z} \triangleq [\mathbf{z}_{\mathcal{I}}^T, \mathbf{z}_{\mathcal{F}}^T]^T$, the measurement ensemble in the presence of measurement error $\boldsymbol{\varepsilon} \triangleq [\boldsymbol{\varepsilon}_{\mathcal{I}}^T, \boldsymbol{\varepsilon}_{\mathcal{F}}^T]^T$ is

$$\mathbf{z} = \mathbf{f}(\bar{\mathbf{x}}) + \boldsymbol{\varepsilon}, \quad (52)$$

where $\bar{\mathbf{x}} = [\bar{\theta}_1, \cdots, \bar{\theta}_N, \bar{V}_1, \cdots, \bar{V}_N]^T$ is the true state.

B. Formulation and Solution for the PSSE

A reasonable abstraction of the data acquisition architecture in power systems is as an interconnected multi-site infrastructure, with I *sites* in which the i-th *site* covers a subset of buses $n \in \mathcal{N}_i$ satisfying $\mathcal{N}_j \bigcap \mathcal{N}_i = \varnothing$ and $\mathcal{N}_i, \mathcal{N}_j \subset \mathcal{N}$ (Fig. 2). The i-th *site* temporally aligns and aggregates a snapshot of M_i local measurements of $\{z_{i,m}\}_{m=1}^{M_i}$ within the site or on the lines that connect its own site with others. The i-th site's measurements are selected from the ensemble in (52) as

$$\mathbf{z}_i = \mathbf{T}_i \mathbf{z} = \mathbf{f}_i(\mathbf{x}) + \mathbf{T}_i \boldsymbol{\varepsilon}, \quad (53)$$

where $\mathbf{f}_i(\mathbf{x}) \triangleq \mathbf{T}_i \mathbf{f}(\mathbf{x})$, and $\mathbf{T}_i \triangleq \mathrm{diag}[\mathbf{T}_{i,\mathcal{I}}, \mathbf{T}_{i,\mathcal{F}}]$ is a block diagonal binary matrix selecting the corresponding measurements at the i-th site. Specifically, $\mathbf{T}_{i,\mathcal{I}} \in \{0,1\}^{M_{i,\mathcal{I}} \times 2N}$ and $\mathbf{T}_{i,\mathcal{F}} \in \{0,1\}^{M_{i,\mathcal{F}} \times 4L}$ are selection matrices with each row having only one "1" entry located at the index of the corresponding element in $\mathbf{f}(\mathbf{x})$ measured by field devices. The number of measurements recorded by each agent is $M_i = M_{i,\mathcal{I}} + M_{i,\mathcal{F}} = \mathrm{Tr}(\mathbf{T}_i^T \mathbf{T}_i)$.

The universally accepted problem formulation for static state estimation is to solve a weighted NLLS problem that fits the estimated state to the power measurements [36], [37]. Assuming $\mathbb{E}\{\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\} = \sigma^2 \mathbf{I}$, the state is estimated as

$$\widehat{\mathbf{x}} = \min_{\mathbf{x} \in \mathbb{X}} \sum_{i=1}^{I} \|\mathbf{z}_i - \mathbf{f}_i(\mathbf{x})\|^2 \quad (54)$$

where $\mathbb{X} \triangleq \{\theta_n \in [-\theta_{\max}, \theta_{\max}], V_n \in [0, V_{\max}], n \in \mathcal{N}\}$ with $\theta_{\max}$ and $V_{\max}$ being the phase angle and voltage limit. By letting $\mathbf{g}_i(\mathbf{x}) \triangleq \mathbf{z}_i - \mathbf{f}_i(\mathbf{x})$ and $\mathbf{G}_i(\mathbf{x}) \triangleq -\partial \mathbf{f}_i(\mathbf{x}) / \partial \mathbf{x}^T$, the problem can be solved using the proposed GGN algorithm.

In many practical scenarios [24], [27]–[29], many similar NLLS problems in a network take the form

$$\widehat{\mathbf{x}}[t] = \min_{\mathbf{x} \in \mathbb{X}} \sum_{i=1}^{I} \|\mathbf{z}_i[t] - \mathbf{f}_i(\mathbf{x})\|^2, \quad (55)$$

where $\mathbf{z}_i[t] \in \mathbb{R}^{M_i}$ is a snapshot of measurements taken at agent i at time t. In this scenario, the GGN algorithm can be readily applied to track the state by initializing $\mathbf{x}_i^0[t]$ with the previous local estimate $\widehat{\mathbf{x}}_i[t]$. In the following, we show the performance of the GGN algorithm in estimating and tracking the state of power systems using real-time measurements.

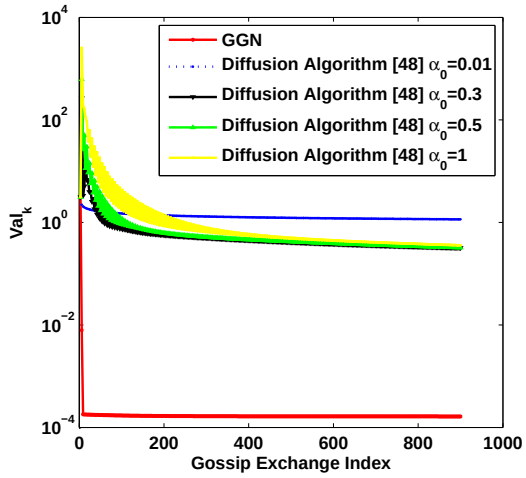
VI. Numerical Results

In this section, we compare the GGN algorithm cost in (54) and Mean Square Error (MSE) with that of the algorithms [48]. We also show numerically that the GGN algorithm can process measurements adaptively and compare it against the method in [24]. Given the distributed estimates $\{\widehat{V}_{i,n}^{(k)}, \widehat{\theta}_{i,n}^{(k)}\}$ at each update in each area from $\mathbf{v}_i^k = [\cdots, \widehat{\theta}_{i,n}^{(k)}, \cdots, \widehat{V}_{i,n}^{(k)} \cdots]^T$, the local MSE relative to $\bar{V}_n$'s and $\bar{\theta}_n$'s is

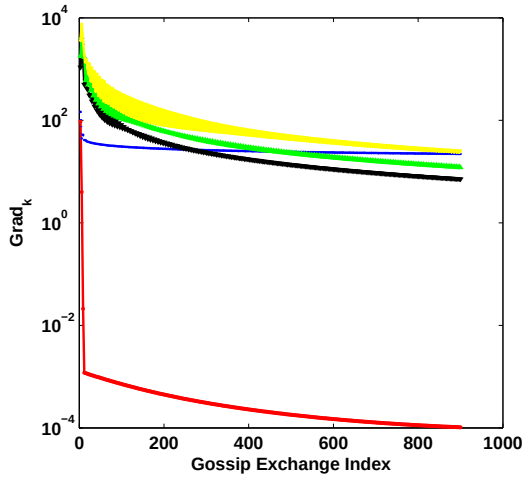
$$\mathrm{MSE}_{V,i}^{(k)} = \mathbb{E}\left\{\sum_{n=1}^{N} (\widehat{V}_{i,n}^{(k)} - \bar{V}_n)^2\right\} / N \quad (56)$$

$$\mathrm{MSE}_{\Theta,i}^{(k)} = \mathbb{E}\left\{\sum_{n=1}^{N} (\widehat{\theta}_{i,n}^{(k)} - \bar{\theta}_n)^2\right\} / N. \quad (57)$$

We considered the IEEE-30 bus ($N = 30$) model in MATPOWER 4.0. The initialization for the state is 1 for the voltage magnitude and 0 for the phase. We take one snapshot of the load profile from the UK National Grid load curve from [55] and scale the base load from MATPOWER on the load buses. Then we run the Optimal Power Flow (OPF) program to determine the generation dispatch for that snapshot. This gives us the true state $\bar{\mathbf{x}}$ and $\mathbf{f}(\bar{\mathbf{x}})$ in per unit (p.u.) values.



(a) Objective value



(b) Gradient norm

Fig. 3. Comparison between the GGN algorithm (CSE Protocol) and diffusion algorithm in [48] with $\ell_{\min} = 3$ exchanges

A. Comparison with Diffusion Algorithms under CSE Protocol

Here we evaluate the performance of the GGN algorithm against the diffusion algorithm for PSSE in [48], and its extension to adaptive processing in [24]. To make a fair comparison in terms of communication costs and accuracy, we exploit the CSE protocol used in [24], [48] for our method, where the exchange is coordinated and synchronous. Measurements $\{\mathbf{z}_i\}_{i=1}^I$ by are generated adding independent Gaussian errors $\varepsilon_{i,m} \sim \mathcal{N}(0, \sigma^2)$ with $\sigma^2 = 10^{-6}$. For simplicity, we divide the system into $I = 3$ sites as in Fig. 2 and the communication graph is fully connected, giving an adjacency matrix $\mathbf{A} = \mathbf{1}_I\mathbf{1}_I^T - \mathbf{I}$. The weight matrix is constructed according to the Laplacian $\mathbf{L} = \mathrm{diag}(\mathbf{A}\mathbf{1}_I) - \mathbf{A}$ as $\mathbf{W} = \mathbf{I}_I - w\mathbf{L}$ with $w = \beta/\max(\mathbf{A}\mathbf{1}_I)$ and $\beta = 0.3$. The step-size is $\alpha_{\mathrm{GGN}} = 0.5$ for the GGN algorithm while $\alpha_{\text{diff},\ell} = 0.01\ell^{-1}, 0.3\ell^{-1}, 0.5\ell^{-1}, \ell^{-1}$ for [24], [48]. The network diffusion algorithm takes place at each exchange ℓ, while the GGN algorithm runs $\ell_k = \ell_0 = \ell_{\min} = 3$ exchanges for each update. Therefore, the equivalent exchange index ℓ in the the figure for the diffusion algorithm is $\ell = \sum_{k=1}^K \ell_k$ ater the K-th updates, where $K = 1, \cdots, 3000$.

1) Estimation on Static Measurements: In this subsection, we show the comparison between our approach and that in [48] over 900 exchanges overall. In particular, the comparison is on the global objective (54) evaluated with local estimates

$$\mathrm{Val}_k = \sum_{i=1}^I \left\|\mathbf{z}_i - \mathbf{f}_i(\mathbf{x}_i^k)\right\|^2 \tag{58}$$

and the following term to evaluate the optimality in (5)

$$\mathrm{Grad}_k = \sum_{i=1}^I \left\|\mathbf{G}_i^T(\mathbf{x}_i^k)(\mathbf{z}_i - \mathbf{f}_i(\mathbf{x}_i^k))\right\|. \tag{59}$$

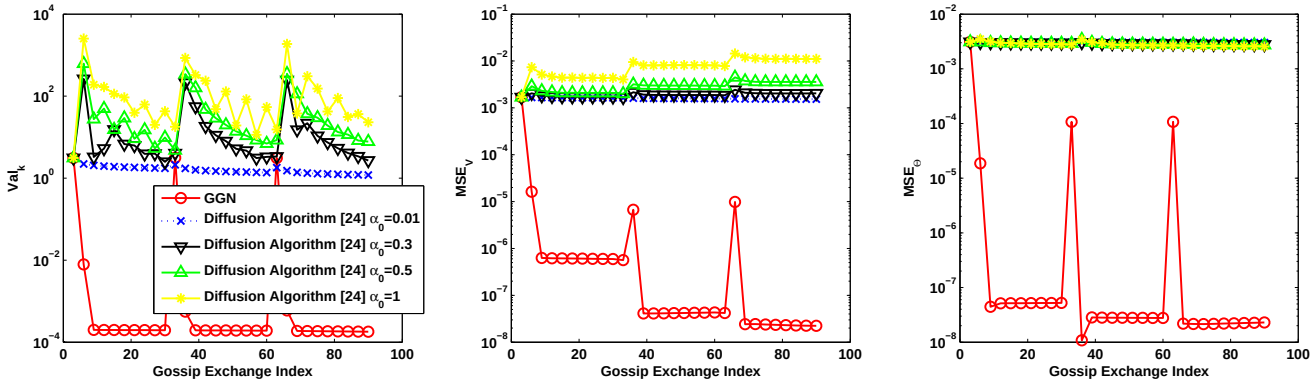
which are plotted against the total number of gossip exchanges so that the comparison is performed on the same time scale.

Clearly, the GGN algorithm converges much faster since it reaches the steady state error after $k = 10$ updates (i.e., $k\ell_{\min} = 30$ exchanges). It is observed in Fig. 3(a) to 3(b) that although the gossip exchange per update $\ell_{\min} = 3$ is small and does not satisfy Assumption 4, both Val_k and the Grad_k still decrease exponentially as the iterations progress. On the other hand, the objective value and the gradient norm of the diffusion algorithm in [48] decrease slowly. Furthermore, the update of diffusion algorithms exhibit more fluctuations especially in the beginning, while the GGN algorithm conditions the gradient by the GN Hessian and therefore the update tends to be smooth and continues to lie in the proximity of the desired solution with high accuracy. Furthermore, the performances of the diffusion algorithm are sensitive to the step-size $\alpha_{\text{diff},\ell}$, since $\alpha_{\text{diff},\ell} = 0.01\ell^{-1}$ is better initially due to less fluctuations as a result of the small step-size, while $0.3\ell^{-1}$ gradually outperforms $0.01\ell^{-1}$ due to the progress made by the larger step-size. However, when the step-size continues to increase, the performance starts to deteriorate from $\alpha_{\text{diff},\ell} = 0.5\ell^{-1}$ to ℓ^{-1}, and even diverges beyond a certain value.

2) Estimation via Adaptive Processing: Here we show numerically the applicability of the GGN algorithm to adaptive processing as described in (55) against the method proposed in [24] with the same network setting and step-sizes. Furthermore, we compare the global MSE performance of the GGN algorithm against the diffusion algorithm, given by

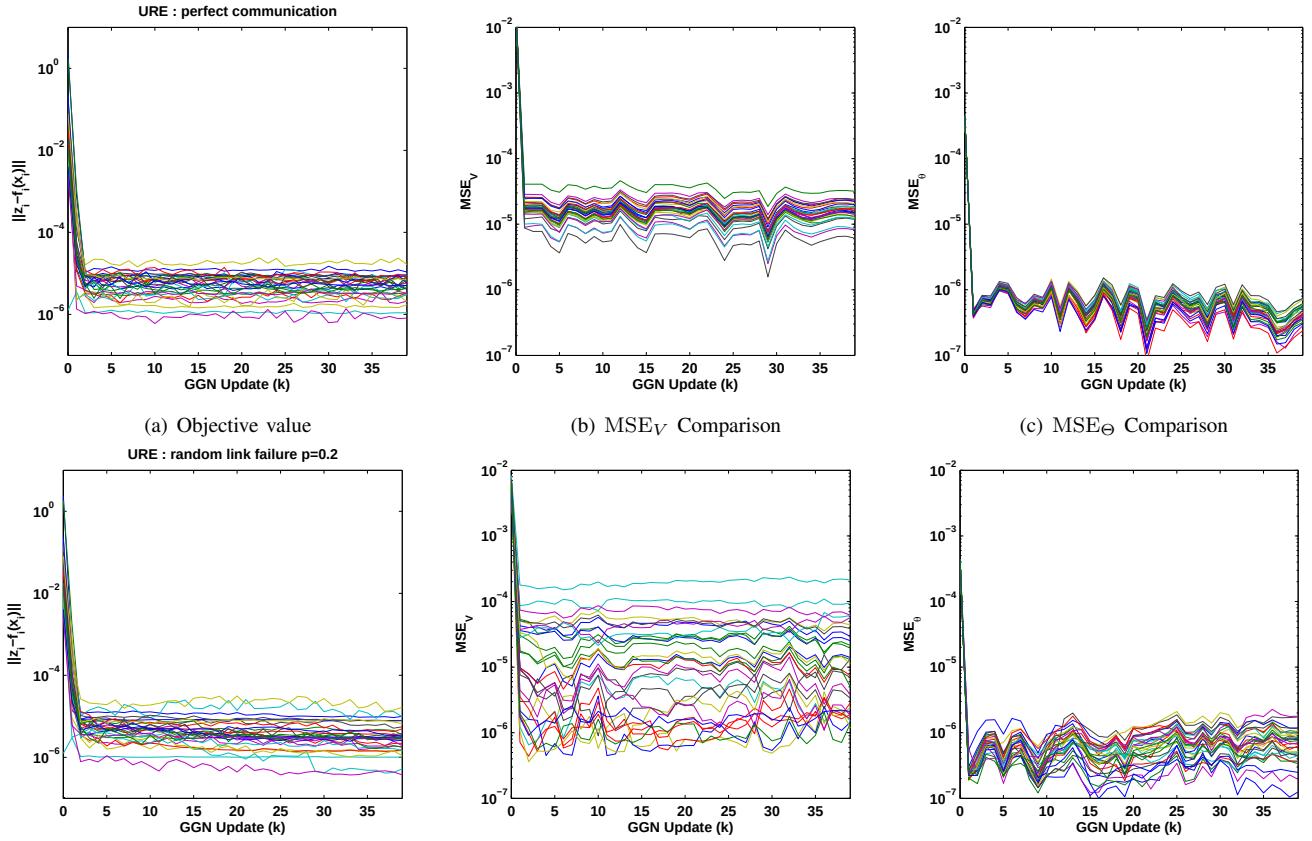
$$\mathrm{MSE}_V^{(k)} = \frac{1}{I}\sum_{i=1}^I \mathrm{MSE}_{V,i}^{(k)},\ \mathrm{MSE}_\Theta^{(k)} = \frac{1}{I}\sum_{i=1}^I \mathrm{MSE}_{\Theta,i}^{(k)}. \tag{60}$$

We generate 3 snapshots of measurements $\{\mathbf{z}_i[t]\}_{i=1}^I$ for $t = 1, \cdots, 3$ based on the same state $\bar{\mathbf{x}}[t] = \bar{\mathbf{x}}$ by adding independent Gaussian noise with variance $\sigma^2 = 10^{-6}$, similar to the adaptive setting considered in [24]. More specifically, we use $\ell_{\min} = 3$ gossip exchanges between every two algorithm updates until $k = 10$, thus leading to a total number of 30 exchanges per snapshot. It can be seen from Fig. 4(a) to 4(c) that the proposed GGN algorithm tracks the state accurately when new measurements stream in, where the spikes observed in the plots are caused by the new measurements. Since the number of gossip exchanges is limited, the diffusion algorithm in [48] and [24] convergence slowly and fail to track the state.



(a) Objective value (b) MSE_V Comparison (c) MSE_Θ Comparison

Fig. 4. Comparison between GGN Algorithm (CSE Protocol) and [24] against ℓ with $\ell_{\min} = 3$ exchanges for every update.



(a) Objective value (b) MSE_V Comparison (c) MSE_Θ Comparison

(d) Objective value (e) MSE_V Comparison (f) MSE_Θ Comparison

Fig. 5. MSE performance of GGN (URE Protocol) in IEEE-30 bus system with $I = N = 30$ agents and $\mathcal{O}(N)$ pair-wise gossip exchanges. (top) : perfect communication (bottom) : $p = 0.3$ random link failures (each line corresponds to one agent).

B. MSE Performance under URE Protocol with Link Failures

In this section, we examine the MSE performance of the GGN algorithm under the URE protocol with a fixed number of algorithm updates $K = 40$. The performance is evaluated with a demanding setting, where we divide the N-bus system into $I = N$ sites and each site only communicates with each of its neighbors 10 times on average. The network-wide communication volume in this scenario is on the order of the network diameter $\mathcal{O}(N)$, which implies the number of transmissions in the centralized scheme as if the local measurements are relayed and routed through the entire network. For simplicity, we simulate that at each exchange, the i-th distributed agents wakes up with uniform probability $1/I$ and picks a neighbor with equal probability $1/I$.

In order to show the robustness of the proposed algorithm, we examine the performance of the GGN algorithm for cases with random link failures, where any established link $\{i, j\} \in \mathcal{M}$ fails with probability $p = 0.3$ independently. It is clear that this communication model with link failures may not satisfy Assumption 2, but it is shown below that

our approach is robust to the random setting and degrades gracefully with the probability of failures. In Fig. 5, we track both the individual objective $\mathrm{Val}_i^{(k)} = \|\mathbf{z}_i - \mathbf{f}_i(\mathbf{x}_i^k)\|^2$ as well as the individual $\mathrm{MSE}_{V,i}^{(k)}$ and $\mathrm{MSE}_{\Theta,i}^{(k)}$ defined in (56) and (57). It can be observed from the figure that the MSE curves of state estimates of different sites are highly consistent and they all converge asymptotically when there is no link failures. Similar behaviors can be observed for the case with random link failures, where the local estimate at each site is not in perfect consistence with the others, but the accuracy remains satisfactory compared to the perfect case and degrades gracefully with the probability of link failures.

VII. Conclusions

In this paper, we study the convergence and performance of the GGN algorithm and discuss its application in power system state estimation. The numerical results suggest that that the proposed algorithm leads to accurate state estimates across the distributed areas, is robust to link/node failures, with polinomial communication and computation cost.

VIII. Acknowledgements

We wish to thank the Associate Editor the anonymous Reviewers for their comments. Their suggestions helped improve this article significantly.

Appendix A
Proof of Lemma 2

To study the convergence of the GGN algorithm, we examine the update in (21) and re-write it with respect to the exact descent $\mathbf{d}_i^k$ in (14)

$$\mathbf{x}_i^{k+1} = P_{\mathbb{X}}\left[\mathbf{x}_i^k - \alpha\mathbf{d}_i^k + \alpha\left(\mathbf{d}_i^k - \mathbf{d}_i^k(\ell_k)\right)\right]. \tag{61}$$

By subtracting the fixed point $\widehat{\mathbf{x}}$ and using the non-expansive property of the operator $P_{\mathbb{X}}(\cdot)$ on the closed convex set $\mathbb{X}$, we have the following recursion

$$\left\|\mathbf{x}_i^{k+1} - \widehat{\mathbf{x}}\right\| \le \left\|\mathbf{x}_i^k - \widehat{\mathbf{x}} - \alpha\mathbf{d}_i^k\right\| + \alpha\left\|\mathbf{d}_i^k(\ell_k) - \mathbf{d}_i^k\right\|.$$

For convenience, we denote $\mathbf{G}^\dagger(\cdot)$ as the pseudo-inverse of $\mathbf{G}(\cdot)$. For any fixed point $\widehat{\mathbf{x}} \in \mathbb{X}$ in (5) such that $\mathbf{G}^\dagger(\widehat{\mathbf{x}})\mathbf{g}(\widehat{\mathbf{x}}) = \mathbf{0}$, the first term can be equivalently written by substituting (9) as follows

$$\begin{aligned}\mathbf{x}_i^k - \widehat{\mathbf{x}} - \alpha\mathbf{d}_i^k &= \mathbf{x}_i^k - \widehat{\mathbf{x}} \\ &\quad - \alpha\mathbf{G}^\dagger(\mathbf{x}_i^k)\mathbf{g}(\mathbf{x}_i^k) + \alpha\mathbf{G}^\dagger(\widehat{\mathbf{x}})\mathbf{g}(\widehat{\mathbf{x}}).\end{aligned} \tag{62}$$

Using (4) together with the invertibility condition of $\mathbf{G}(\mathbf{x})$ over $\mathbf{x} \in \mathbb{X}$ in Assumption 1, we have

$$\mathbf{x}_i^k - \widehat{\mathbf{x}} = \mathbf{G}^\dagger(\mathbf{x}_i^k)\mathbf{G}(\mathbf{x}_i^k)\left(\mathbf{x}_i^k - \widehat{\mathbf{x}}\right). \tag{63}$$

Then by substituting (63) into (62), and meanwhile adding and subtracting simultaneously a term $\alpha\mathbf{G}^\dagger(\mathbf{x}_i^k)\mathbf{g}(\widehat{\mathbf{x}})$, we have the following expression

$$\mathbf{x}_i^k - \widehat{\mathbf{x}} - \alpha\mathbf{d}_i^k \tag{64}$$

$$= \mathbf{G}^\dagger(\mathbf{x}_i^k)\left[\mathbf{G}(\mathbf{x}_i^k)\left(\mathbf{x}_i^k - \widehat{\mathbf{x}}\right) - \alpha\mathbf{g}(\mathbf{x}_i^k) + \alpha\mathbf{g}(\widehat{\mathbf{x}})\right] \tag{65}$$

$$+ \alpha\left[\mathbf{G}^\dagger(\widehat{\mathbf{x}}) - \mathbf{G}^\dagger(\mathbf{x}_i^k)\right]\mathbf{g}(\widehat{\mathbf{x}}). \tag{66}$$

The expression in the first term can be re-written with the mean-value theorem as follows

$$\begin{aligned}&\alpha\mathbf{g}(\widehat{\mathbf{x}}) - \alpha\mathbf{g}(\mathbf{x}) - \mathbf{G}(\mathbf{x})(\widehat{\mathbf{x}} - \mathbf{x}) \\ &= \alpha\left[\int_0^1 \mathbf{G}(\mathbf{x} + t(\widehat{\mathbf{x}} - \mathbf{x}))(\widehat{\mathbf{x}} - \mathbf{x})\mathrm{d}t\right] - \mathbf{G}(\mathbf{x})(\widehat{\mathbf{x}} - \mathbf{x}) \\ &= \alpha\left(\int_0^1 \left[\mathbf{G}(\mathbf{x} + t(\widehat{\mathbf{x}} - \mathbf{x})) - \mathbf{G}(\mathbf{x})\right](\widehat{\mathbf{x}} - \mathbf{x})\mathrm{d}t\right) \\ &\quad - (1-\alpha)\mathbf{G}(\mathbf{x})(\widehat{\mathbf{x}} - \mathbf{x}),\end{aligned} \tag{67}$$

whose norm can be bounded by using Assumption 1 as

$$\begin{aligned}&\|\alpha\mathbf{g}(\widehat{\mathbf{x}}) - \alpha\mathbf{g}(\mathbf{x}) - \mathbf{G}(\mathbf{x})(\widehat{\mathbf{x}} - \mathbf{x})\| \\ &\le \alpha\left[\int_0^1 \|\mathbf{G}(\mathbf{x} + t(\widehat{\mathbf{x}} - \mathbf{x})) - \mathbf{G}(\mathbf{x})\|\,\mathrm{d}t\right]\|\mathbf{x} - \widehat{\mathbf{x}}\| \\ &\quad + (1-\alpha)\sigma_{\max}\|\mathbf{x} - \widehat{\mathbf{x}}\|.\end{aligned} \tag{68}$$

From the Lipschitz condition in Assumption 1, we have

$$\int_0^1 \|\mathbf{G}(\mathbf{x} + t(\widehat{\mathbf{x}} - \mathbf{x})) - \mathbf{G}(\mathbf{x})\|\,\mathrm{d}t \le \omega\|\mathbf{x} - \widehat{\mathbf{x}}\|\int_0^1 t\mathrm{d}t.$$

Thus, if condition (3) of Assumption 1 holds, we have

$$\begin{aligned}&\|\alpha\mathbf{g}(\widehat{\mathbf{x}}) - \alpha\mathbf{g}(\mathbf{x}) - \mathbf{G}(\mathbf{x})(\widehat{\mathbf{x}} - \mathbf{x})\| \\ &\le \frac{\alpha\omega}{2}\|\mathbf{x} - \widehat{\mathbf{x}}\|^2 + (1-\alpha)\sigma_{\max}\|\mathbf{x} - \widehat{\mathbf{x}}\|,\end{aligned} \tag{69}$$

and finally according to [56, Lemma 1], we have

$$\begin{aligned}\|\mathbf{G}^\dagger(\mathbf{x}) - \mathbf{G}^\dagger(\widehat{\mathbf{x}})\| &\le \sqrt{2}\|\mathbf{G}^\dagger(\mathbf{x})\|\|\mathbf{G}^\dagger(\widehat{\mathbf{x}})\|\|\mathbf{G}(\mathbf{x}) - \mathbf{G}(\widehat{\mathbf{x}})\| \\ &\le \frac{\sqrt{2}\omega}{\sigma_{\min}^2}\|\mathbf{x} - \widehat{\mathbf{x}}\|.\end{aligned} \tag{70}$$

By definition we have $\|\mathbf{G}^\dagger(\boldsymbol{x})\|^2 = \|\left(\mathbf{G}^T(\boldsymbol{x})\mathbf{G}(\boldsymbol{x})\right)^{-1}\|$. Also Assumption 1 implies $\|\mathbf{G}^\dagger(\boldsymbol{x})\| \le 1/\sigma_{\min}$.

For convenience, we let $\epsilon_{\min} \triangleq \|\mathbf{g}(\widehat{\mathbf{x}})\|$ be the goodness of fit at $\widehat{\mathbf{x}}$ and define the following constants

$$T_1 \triangleq \frac{\alpha\omega}{2\sigma_{\min}}, \quad T_2 \triangleq (1-\alpha)\frac{\sigma_{\max}}{\sigma_{\min}} + \frac{\sqrt{2}\alpha\omega\epsilon_{\min}}{\sigma_{\min}^2}. \tag{71}$$

Then, substituting $\|\mathbf{G}^\dagger(\mathbf{x})\| \le 1/\sigma_{\min}$, (69) and (70) back into (64) and using (30), we have

$$\left\|\mathbf{x}_i^k - \widehat{\mathbf{x}} - \alpha\mathbf{d}_i^k\right\| \le T_1\left\|\mathbf{x}_i^k - \widehat{\mathbf{x}}\right\|^2 + T_2\left\|\mathbf{x}_i^k - \widehat{\mathbf{x}}\right\|,$$

Therefore, we have the error recursion (29).

Appendix B
Proof of Theorem 1

If the discrepancy error is upper bounded by a constant $\kappa \ge 0$ such that $\|\mathbf{d}_i^k(\ell_k) - \mathbf{d}_i^k\| \le \kappa$, then from Lemma 2, the recursion can be simplified as

$$\left\|\mathbf{x}_i^{k+1} - \mathbf{x}^\star\right\| \le T_1\left\|\mathbf{x}_i^k - \mathbf{x}^\star\right\|^2 + T_2\left\|\mathbf{x}_i^k - \mathbf{x}^\star\right\| + \alpha\kappa. \tag{72}$$

where T_1 and T_2 are given in (30). Let $\zeta_{i,k} = \left\|\mathbf{x}_i^k - \mathbf{x}^\star\right\|$, then the error recursion can be expressed as a dynamical system as

$$\zeta_{i,k+1} \le T_1\zeta_{i,k}^2 + T_2\zeta_{i,k} + \alpha\kappa, \quad \zeta_{i,k} > 0. \tag{73}$$

Since $\zeta_{i,k}$ is non-negative, this error dynamic can be upper bounded by the dynamical system of $\rho_{k+1} = \psi(\rho_k)$ with

$$\psi(\rho_k) = T_1\rho_k^2 + T_2\rho_k + \alpha\kappa, \quad \rho_k > 0, \tag{74}$$

whose equilibrium points are obtained by solving

$$\rho = T_1\rho^2 + T_2\rho + \alpha\kappa. \tag{75}$$

When κ satisfies

$$(1-T_2)^2 - 4\alpha T_1\kappa \geq 0, \tag{76}$$

the equilibrium points of (75) exist and are obtained as (34).

Now let $\dot{\psi}(\rho) \triangleq \mathrm{d}\psi(\rho)/\mathrm{d}\rho$ be the first order derivative of the dynamics. According to [57, cf. Proposition 1.9], an equilibrium point is a stable sink if $|\dot{\psi}(\cdot)| < 1$ and unstable otherwise. Thus, the equilibrium point $\rho_{\max}$ is unstable since the following is always true

$$\left|\dot{\psi}(\rho_{\max})\right| = |2T_1\rho_{\max} + T_2| \tag{77}$$

$$= \left|1 + \sqrt{(1-T_2)^2 - 4\alpha T_1\kappa}\right| > 1, \tag{78}$$

while the point $\rho_{\min}$ is a sink if

$$\left|\dot{\psi}(\rho_{\min})\right| = \left|1 - \sqrt{(1-T_2)^2 - 4\alpha T_1\kappa}\right| < 1. \tag{79}$$

To guarantee $\left|\dot{\psi}(\rho_{\min})\right| < 1$, it requires

$$0 < (1-T_2)^2 - 4\alpha T_1\kappa < 4, \tag{80}$$

which together with (76) leads to the following condition on the bounded perturbation κ

$$\frac{T_2^2 - 2T_2 - 3}{4\alpha T_1} < \kappa < \frac{T_2^2 - 2T_2 + 1}{4\alpha T_1}. \tag{81}$$

Clearly, given an arbitrary $\alpha \in (0, 1]$, the lower bound on κ is unrealistic if $T_2^2 - 2T_2 - 3 > 0$ since the lower bound could approach infinity as $\alpha \to 0$. Therefore, to ensure convergence with an arbitrarily small perturbation, it is sufficient to have

$$T_2^2 - 2T_2 - 3 < 0 \Longrightarrow -1 < T_2 < 3. \tag{82}$$

Since $T_2 \geq 0$ by definition (30), the condition becomes

$$0 \leq (1-\alpha)\frac{\sigma_{\max}}{\sigma_{\min}} + \frac{\sqrt{2}\alpha\omega\epsilon_{\min}}{\sigma_{\min}^2} < 3. \tag{83}$$

By re-arranging the terms, this condition is equivalent to

$$\begin{cases} \dfrac{\sqrt{2}\alpha\omega\epsilon_{\min}}{\sigma_{\min}^2} < 3 - (1-\alpha)\dfrac{\sigma_{\max}}{\sigma_{\min}} \\ 3 - (1-\alpha)\dfrac{\sigma_{\max}}{\sigma_{\min}} > 0 \end{cases}, \tag{84}$$

which can be simplified as

$$\begin{cases} \omega\epsilon_{\min} < \dfrac{\sigma_{\min}^2}{\sqrt{2}\alpha}\left[3 - (1-\alpha)\dfrac{\sigma_{\max}}{\sigma_{\min}}\right] \\ \max\left\{1 - \dfrac{3\sigma_{\min}}{\sigma_{\max}}, 0\right\} < \alpha \leq 1 \end{cases}. \tag{85}$$

Thus, if the initial error $\zeta_{i,0} > \rho_{\max}$, the error keeps growing. On the other hand, if the errors are bounded by $0 < \zeta_{i,k} < \rho_{\max}$ for all i's and k's, the algorithm reaches the equilibrium error floor $\rho_{\min}$. Thus, as long as the initialization error $\zeta_{i,0}$ satisfies $0 < \zeta_{i,0} < \rho_{\max}$ for $i = 1, \cdots, I$, the algorithm progresses with contracting error until reaching the error floor $\rho_{\min}$ due to the constant bounded perturbation κ.

As a result, as long as the initial condition $\mathbf{x}_i^0$ satisfies $\left\|\mathbf{x}_i^0 - \mathbf{x}^\star\right\| < \rho_{\max}$ with respect to a certain fixed point $\mathbf{x}^\star$, the error norm is upper bounded by

$$\limsup_{k\to\infty} \left\|\mathbf{x}_i^k - \mathbf{x}^\star\right\| \leq \rho_{\min}.$$

Appendix C
Proof of Lemma 3

Using (24), we evaluate the deviation of $\boldsymbol{\mathcal{H}}_k(\ell)$ from the average $\bar{\boldsymbol{\mathcal{H}}}_k = \left[\mathbf{1}^T \otimes \mathbf{I}_{N_\mathcal{H}}\right]\boldsymbol{\mathcal{H}}_k(0)/I$ for a finite ℓ. By subtracting the average $\bar{\boldsymbol{\mathcal{H}}}_k$ on both sides of (24), we have

$$\begin{aligned} &\boldsymbol{\mathcal{H}}_k(\ell) - \bar{\boldsymbol{\mathcal{H}}}_k \\ &= \left[\mathbf{W}_k(\ell) \otimes \mathbf{I}_{N_\mathcal{H}}\right]\boldsymbol{\mathcal{H}}_k(\ell-1) - \frac{\mathbf{1}_N\mathbf{1}_N^T \otimes \mathbf{I}_{N_\mathcal{H}}}{I}\boldsymbol{\mathcal{H}}_k(0) \\ &= \left[\prod_{\ell'=0}^{\ell} \mathbf{W}_k(\ell') \otimes \mathbf{I}_{N_\mathcal{H}}\right]\boldsymbol{\mathcal{H}}_k(0) - \frac{\mathbf{1}_N\mathbf{1}_N^T \otimes \mathbf{I}_{N_\mathcal{H}}}{I}\boldsymbol{\mathcal{H}}_k(0) \\ &= \left[\left(\prod_{\ell'=0}^{\ell} \mathbf{W}_k(\ell') - \frac{\mathbf{1}_N\mathbf{1}_N^T}{I}\right) \otimes \mathbf{I}_{N_\mathcal{H}}\right]\boldsymbol{\mathcal{H}}_k(0). \end{aligned}$$

Then, we bound the norms of the above equation as

$$\left\|\boldsymbol{\mathcal{H}}_k(\ell) - \bar{\boldsymbol{\mathcal{H}}}_k\right\| \leq \left\|\prod_{\ell'=0}^{\ell} \mathbf{W}_k(\ell') - \frac{\mathbf{1}_N\mathbf{1}_N^T}{I}\right\| \left\|\boldsymbol{\mathcal{H}}_k(0)\right\|. \tag{86}$$

Using Lemma 1 and the norm inequality $\|\cdot\| \leq \|\cdot\|_F$, we have

$$\begin{aligned} \left\|\boldsymbol{\mathcal{H}}_k(\ell) - \bar{\boldsymbol{\mathcal{H}}}_k\right\| &\leq \left\|\prod_{\ell'=0}^{\ell} \mathbf{W}_k(\ell') - \frac{\mathbf{1}_N\mathbf{1}_N^T}{I}\right\|_F \left\|\boldsymbol{\mathcal{H}}_k(0)\right\| \\ &\leq \left[2I\left(\frac{1+\eta^{-L_0}}{1-\eta^{L_0}}\right)\lambda_\eta^\ell\right]\left\|\boldsymbol{\mathcal{H}}_k(0)\right\|. \end{aligned}$$

The quantity $\|\boldsymbol{\mathcal{H}}_k(0)\|$ is by definition (17) determined as

$$\begin{aligned} \|\boldsymbol{\mathcal{H}}_k(0)\|^2 &= \sum_{i=1}^{I}\|\mathbf{h}_{k,i}(0)\|^2 + \sum_{i=1}^{I}\|\mathbf{H}_{k,i}(0)\|_F^2 \\ &= \sum_{i=1}^{I}\left(\left\|\mathbf{G}_i^T(\mathbf{x}_i^k)\mathbf{g}_i(\mathbf{x}_i^k)\right\|^2 + \left\|\mathbf{G}_i^T(\mathbf{x}_i^k)\mathbf{G}_i(\mathbf{x}_i^k)\right\|_F^2\right) \\ &\leq I\sigma_{\max}^2(\epsilon_{\max}^2 + N\sigma_{\max}^2), \end{aligned} \tag{87}$$

where the norm inequality is used

$$\|\mathbf{G}_i^T(\mathbf{x})\mathbf{G}_i(\mathbf{x})\|_F^2 \leq N\|\mathbf{G}^T(\mathbf{x})\mathbf{G}(\mathbf{x})\|_2^2 = N\sigma_{\max}^4.$$

Letting $C = 2I\sqrt{I\sigma_{\max}^2(\epsilon_{\max}^2 + N\sigma_{\max}^2)}\left(\frac{1+\eta^{-L_0}}{1-\eta^{L_0}}\right)$, then the error is bounded as $\left\|\boldsymbol{\mathcal{H}}_k(\ell) - \bar{\boldsymbol{\mathcal{H}}}_k\right\| \leq C\lambda_\eta^\ell$.

By definition of $\mathbf{e}_{k,i}(\ell)$ and $\mathbf{E}_{k,i}(\ell)$, we have

$$\boldsymbol{\mathcal{H}}_k(\ell) - \bar{\boldsymbol{\mathcal{H}}}_k = \begin{bmatrix} \mathbf{e}_{k,1}(\ell) \\ \mathrm{vec}\left[\mathbf{E}_{k,1}(\ell)\right] \\ \vdots \\ \mathbf{e}_{k,I}(\ell) \\ \mathrm{vec}\left[\mathbf{E}_{k,I}(\ell)\right] \end{bmatrix}, \tag{88}$$

and hence the norm of each component is bounded by the total norm $\|\mathbf{e}_k(\ell)\| < C\lambda_\eta^\ell$ and $\|\mathbf{E}_k(\ell)\|_F < C\lambda_\eta^\ell$.

Appendix D
Proof of Lemma 4

We prove this result by mathematical induction. We will repetitively use matrix expansion [51] for any $\mathbf{Z}$ and $\delta\mathbf{Z}$,

$$(\mathbf{Z}+\delta\mathbf{Z})^{-1} = \sum_{q=0}^{\infty}(-1)^q \left(\mathbf{Z}^{-1}\delta\mathbf{Z}\right)^q \mathbf{Z}^{-1} \tag{89}$$

as long as $\left\|\mathbf{Z}^{-1}\delta\mathbf{Z}\right\| < 1$

A. Initial Case: $k=1$

Given $\mathbf{x}_i^0 = \mathbf{x}^0$ for all i, then for any $i \neq j$ we have

$$\left\|\mathbf{x}_i^1 - \mathbf{x}_j^1\right\| \leq \left\|\mathbf{d}_i^0(\ell_0) - \mathbf{d}_j^0(\ell_0)\right\|, \tag{90}$$

where the discrepancy is expressed explicitly as

$$\begin{aligned}\mathbf{d}_i^0(\ell_0) - \mathbf{d}_j^0(\ell_0) &= \left[\bar{\mathbf{H}}_0 + \mathbf{E}_{0,i}(\ell_0)\right]^{-1}\left[\bar{\mathbf{h}}_0 + \mathbf{e}_{0,i}(\ell_0)\right] \\ &\quad - \left[\bar{\mathbf{H}}_0 + \mathbf{E}_{0,j}(\ell_0)\right]^{-1}\left[\bar{\mathbf{h}}_0 + \mathbf{e}_{0,j}(\ell_0)\right].\end{aligned}$$

Thus, if $\mathbf{E}_{0,i}(\ell_0)$, $\mathbf{E}_{0,j}(\ell_0)$ are small enough, the expansion in (89) can be applied here to simplify the expression.

1) Matrix series expansion: Since $\mathbf{x}_i^0 = \mathbf{x}^0$ for all i such that $\bar{\mathbf{H}}_0 = \mathbf{Q}(\mathbf{x}_i^0)$ and $\bar{\mathbf{h}}_0 = \mathbf{q}(\mathbf{x}_i^0)$, they can be bounded based on Assumption 1 as follows

$$\begin{aligned}\left\|\bar{\mathbf{h}}_0\right\| &= \left\|\mathbf{q}(\mathbf{x}_i^0)\right\| \\ &= \frac{1}{I}\left\|\mathbf{G}^T(\mathbf{x}_i^0)\mathbf{g}(\mathbf{x}_i^0)\right\| \leq \frac{\sigma_{\max}\epsilon_{\max}}{I}, \\ \left\|\bar{\mathbf{H}}_0^{-1}\right\| &= \left\|\mathbf{Q}^{-1}(\mathbf{x}_i^0)\right\| \\ &= I\left\|\left(\mathbf{G}^T(\mathbf{x}_i^0)\mathbf{G}(\mathbf{x}_i^0)\right)^{-1}\right\| \leq \frac{I}{\sigma_{\min}^2}.\end{aligned} \tag{91}$$

Note that from the norm inequality of sub-matrices

$$\begin{aligned}\left\|\bar{\mathbf{H}}_0^{-1}\mathbf{E}_{0,i}(\ell_0)\right\| &\leq \left\|\bar{\mathbf{H}}_0^{-1}\right\|\left\|\mathbf{E}_{0,i}(\ell_0)\right\| \leq \left\|\bar{\mathbf{H}}_0^{-1}\right\|\left\|\mathbf{E}_0(\ell_0)\right\|_F \\ \left\|\bar{\mathbf{H}}_0^{-1}\mathbf{E}_{0,j}(\ell_0)\right\| &\leq \left\|\bar{\mathbf{H}}_0^{-1}\right\|\left\|\mathbf{E}_{0,j}(\ell_0)\right\| \leq \left\|\bar{\mathbf{H}}_0^{-1}\right\|\left\|\mathbf{E}_0(\ell_0)\right\|_F,\end{aligned}$$

and by Assumption 4 we have $\ell_0 \geq \ell_{\min}$. From Lemma 3 and Assumption 4, the above inequalities can be bounded as

$$\left\|\bar{\mathbf{H}}_0^{-1}\right\|\left\|\mathbf{E}_0(\ell_0)\right\|_F \leq \frac{I}{\sigma_{\min}^2}C\lambda_\eta^{\ell_0} = \lambda_\eta^{(\ell_0-\ell_{\min})}\frac{IC}{\sigma_{\min}^2}\lambda_\eta^{\ell_{\min}}. \tag{92}$$

Choosing $\ell_{\min}$ according to (45), we have $\lambda_\eta^{(\ell_0-\ell_{\min})} < 1$ and

$$\ell_{\min} > \log\left(\frac{\xi}{4D}\right)/\log\lambda_\eta \quad\Longrightarrow\quad \frac{IC}{\sigma_{\min}^2}\lambda_\eta^{\ell_{\min}} < \frac{IC}{4\sigma_{\min}^2 D}\xi.$$

For notation convenience, we define

$$\tilde{\xi} = \frac{IC}{\sigma_{\min}^2 D}\xi, \tag{93}$$

and clearly, we have $0 < \tilde{\xi} < \xi < 1/2$ according to the definition of D in (45) by Assumption 4. Therefore, letting $\delta\mathbf{Z} = \mathbf{E}_0(\ell_0)$ and $\mathbf{Z} = \bar{\mathbf{H}}_0$, it follows from Lemma 3 that

$$\left\|\mathbf{Z}^{-1}\delta\mathbf{Z}\right\| = \left\|\bar{\mathbf{H}}_0^{-1}\right\|\left\|\mathbf{E}_0(\ell_0)\right\|_F \leq \frac{1}{4}\lambda_\eta^{(\ell_0-\ell_{\min})}\tilde{\xi} < \frac{1}{8}. \tag{94}$$

and the expansion holds. By the matrix series expansion and grouping all the high order terms $q \geq 1$, we have

$$\begin{aligned}&\mathbf{d}_i^0(\ell_0) - \mathbf{d}_j^0(\ell_0) \\ &= \left[\bar{\mathbf{H}}_0^{-1} - \sum_{q=1}^{\infty}(-1)^q\left(\bar{\mathbf{H}}_0^{-1}\mathbf{E}_{0,i}(\ell_0)\right)^q\bar{\mathbf{H}}_0^{-1}\right]\left[\bar{\mathbf{h}}_0 + \mathbf{e}_{0,i}(\ell_0)\right] \\ &\quad - \left[\bar{\mathbf{H}}_0^{-1} - \sum_{q=1}^{\infty}(-1)^q\left(\bar{\mathbf{H}}_0^{-1}\mathbf{E}_{0,j}(\ell_0)\right)^q\bar{\mathbf{H}}_0^{-1}\right]\left[\bar{\mathbf{h}}_0 + \mathbf{e}_{0,j}(\ell_0)\right].\end{aligned} \tag{95}$$

To simplify the above expression, we write it in three terms $\mathbf{D}_1(\ell_0)$, $\mathbf{D}_2(\ell_0)$ and $\mathbf{D}_3(\ell_0)$ as follows

$$\mathbf{d}_i^0(\ell_0) - \mathbf{d}_j^0(\ell_0) = \mathbf{D}_1(\ell_0) + \mathbf{D}_2(\ell_0) + \mathbf{D}_3(\ell_0),$$

where $\mathbf{D}_1(\ell_0) \triangleq \bar{\mathbf{H}}_0^{-1}\left[\mathbf{e}_{0,i}(\ell_0) - \mathbf{e}_{0,j}(\ell_0)\right]$ and

$$\begin{aligned}\mathbf{D}_2(\ell_0) &\triangleq \sum_{q=1}^{\infty}(-1)^q\left(\bar{\mathbf{H}}_0^{-1}\mathbf{E}_{0,j}(\ell_0)\right)^q\bar{\mathbf{H}}_0^{-1}\bar{\mathbf{h}}_0 \\ &\quad - \sum_{q=1}^{\infty}(-1)^q\left(\bar{\mathbf{H}}_0^{-1}\mathbf{E}_{0,i}(\ell_0)\right)^q\bar{\mathbf{H}}_0^{-1}\bar{\mathbf{h}}_0 \\ \mathbf{D}_3(\ell_0) &\triangleq \sum_{q=1}^{\infty}(-1)^q\left(\bar{\mathbf{H}}_0^{-1}\mathbf{E}_{0,j}(\ell_0)\right)^q\bar{\mathbf{H}}_0^{-1}\mathbf{e}_{0,j}(\ell_0) \\ &\quad - \sum_{q=1}^{\infty}(-1)^q\left(\bar{\mathbf{H}}_0^{-1}\mathbf{E}_{0,i}(\ell_0)\right)^q\bar{\mathbf{H}}_0^{-1}\mathbf{e}_{0,i}(\ell_0).\end{aligned}$$

2) Proof of success when $k=1$: According to the triangular inequality for norms, we can bound

$$\begin{aligned}\left\|\mathbf{e}_{0,i}(\ell_0) - \mathbf{e}_{0,j}(\ell_0)\right\| &\leq 2\left\|\mathbf{e}_0(\ell_0)\right\| \\ \left\|\mathbf{E}_{0,i}(\ell_0) - \mathbf{E}_{0,j}(\ell_0)\right\| &\leq 2\left\|\mathbf{E}_0(\ell_0)\right\|_F.\end{aligned}$$

Using (94), we can bound the norm of the first term as

$$\left\|\mathbf{D}_1(\ell_0)\right\| \leq 2\left\|\bar{\mathbf{H}}_0^{-1}\right\|\left\|\mathbf{e}_0(\ell_0)\right\| \leq \frac{1}{2}\tilde{\xi}\lambda_\eta^{(\ell_0-\ell_{\min})}. \tag{96}$$

Similarly, the infinite sum in the second term is bounded as

$$\begin{aligned}&\left\|\sum_{q=1}^{\infty}(-1)^q\left[\left(\bar{\mathbf{H}}_0^{-1}\mathbf{E}_{0,j}(\ell_0)\right)^q - \left(\bar{\mathbf{H}}_0^{-1}\mathbf{E}_{0,i}(\ell_0)\right)^q\right]\right\| \\ &\leq 2\sum_{q=1}^{\infty}\left(\left\|\bar{\mathbf{H}}_0^{-1}\right\|\left\|\mathbf{E}_0(\ell_0)\right\|_F\right)^q \\ &\leq 2\sum_{q=1}^{\infty}\left(\frac{1}{4}\tilde{\xi}\lambda_\eta^{(\ell_0-\ell_{\min})}\right)^q = \frac{1}{2}\frac{\tilde{\xi}\lambda_\eta^{(\ell_0-\ell_{\min})}}{\left(1-\frac{1}{4}\tilde{\xi}\lambda_\eta^{(\ell_0-\ell_{\min})}\right)},\end{aligned} \tag{97}$$

where the last equality comes from the convergence of geometric series $\lim_{K\to\infty}\sum_{k=1}^{K}a^k = a/(1-a)$ for any $|a|<1$. Since $0 < \tilde{\xi} < \xi < 1/2$ and $\lambda_\eta^{(\ell_0-\ell_{\min})} \leq 1$, then

$$\frac{\tilde{\xi}\lambda_\eta^{(\ell_0-\ell_{\min})}}{\left(1-\frac{1}{4}\tilde{\xi}\lambda_\eta^{(\ell_0-\ell_{\min})}\right)} < 2\tilde{\xi}\lambda_\eta^{(\ell_0-\ell_{\min})} \tag{98}$$

and thus the norm of the second term is bounded as

$$\begin{aligned}\left\|\mathbf{D}_2(\ell_0)\right\| &\leq \tilde{\xi}\lambda_\eta^{(\ell_0-\ell_{\min})}\left\|\bar{\mathbf{H}}_0^{-1}\right\|\left\|\bar{\mathbf{h}}_0\right\| \\ &\leq \frac{\sigma_{\max}\epsilon_{\max}}{\sigma_{\min}^2}\tilde{\xi}\lambda_\eta^{(\ell_0-\ell_{\min})},\end{aligned}$$

where the last inequality comes from (91).

Following the same rationale, the norm of the third term can be bounded as

$$\|\mathbf{D}_3(\ell_0)\| \le 2\sum_{q=1}^{\infty}\left(\|\bar{\mathbf{H}}_0^{-1}\|\,\|\mathbf{E}_0(\ell_0)\|_F\right)^q \|\bar{\mathbf{H}}_0^{-1}\mathbf{e}_0(\ell_0)\|$$
$$\le 2\sum_{q=1}^{\infty}\left(\|\bar{\mathbf{H}}_0^{-1}\|\,\|\mathbf{E}_0(\ell_0)\|_F\right)^{q+1} \tag{99}$$

which leads to

$$\|\mathbf{D}_3(\ell_0)\| \le 2\sum_{q=1}^{\infty}\left(\frac{1}{4}\tilde{\xi}\lambda_\eta^{(\ell_0-\ell_{\min})}\right)^{q+1} \tag{100}$$
$$= \frac{\tilde{\xi}\lambda_\eta^{(\ell_0-\ell_{\min})}}{\left(1-\frac{1}{4}\tilde{\xi}\lambda_\eta^{(\ell_0-\ell_{\min})}\right)}\cdot\frac{1}{8}\tilde{\xi}\lambda_\eta^{(\ell_0-\ell_{\min})} \tag{101}$$
$$< \frac{1}{2}\tilde{\xi}\lambda_\eta^{(\ell_0-\ell_{\min})}.$$

where the last inequality has used the results in (98). Note that this bound is very loose since we bound a second order term with the first order term.

Substituting $\tilde{\xi} = IC\xi/(\sigma_{\min}^2 D)$ defined in (93) back to (96), (97), (100) and summing them up, we have

$$\|\mathbf{d}_i^0(\ell_0) - \mathbf{d}_j^0(\ell_0)\| \le \left(1+\frac{\sigma_{\max}\epsilon_{\max}}{\sigma_{\min}^2}\right)\frac{IC}{\sigma_{\min}^2 D}\xi\lambda_\eta^{(\ell_0-\ell_{\min})}$$

Introducing the constants C_1 and C_2 defined in (47) and the inequality in (90), we have

$$\|\mathbf{x}_i^1 - \mathbf{x}_j^1\| \le \xi\left(\frac{CC_1C_2}{D}\right)\lambda_\eta^{(\ell_0-\ell_{\min})}. \tag{102}$$

and therefore the result holds for $k=1$.

B. Induction: $k=K$ and $k=K+1$

Let the error bound hold for $k=K$ such that for any $i\neq j$

$$\|\mathbf{x}_i^K - \mathbf{x}_j^K\| \le \xi\left(\frac{CC_1C_2}{D}\right)\sum_{k=0}^{K}\lambda_\eta^{(\ell_k-\ell_{\min})}. \tag{103}$$

with C_1, C_2 given in (47). The inequality below holds

$$\|\mathbf{x}_i^{K+1} - \mathbf{x}_j^{K+1}\| \le \|\mathbf{x}_i^K - \mathbf{x}_j^K\| + \|\mathbf{d}_i^K(\ell_K) - \mathbf{d}_j^K(\ell_K)\|,$$

where

$$\mathbf{d}_i^K(\ell_K) - \mathbf{d}_j^K(\ell_K) = \left[\bar{\mathbf{H}}_K + \mathbf{E}_{K,i}(\ell_K)\right]^{-1}\left[\bar{\mathbf{h}}_K + \mathbf{e}_{K,i}(\ell_K)\right]$$
$$- \left[\bar{\mathbf{H}}_K + \mathbf{E}_{K,j}(\ell_K)\right]^{-1}\left[\bar{\mathbf{h}}_K + \mathbf{e}_{K,j}(\ell_K)\right]. \tag{104}$$

Similar to the case when $k=1$, if the perturbations $\mathbf{E}_{K,i}(\ell_K)$, $\mathbf{E}_{K,j}(\ell_K)$ are small enough, the expansion in (89) can be applied here to simplify the expression.

1) Matrix series expansion: By definition (40), we have

$$\|\bar{\mathbf{H}}_K^{-1}\| = \left\|\left[\mathbf{Q}(\mathbf{x}_i^K) + \mathbf{S}_{K,i}\right]^{-1}\right\|, \tag{105}$$

which is another perturbed inverse. Thus we first examine whether this inverse can be expanded using the series expansion in (89). From (43) and (103), we have

$$\|\mathbf{S}_{K,i}\| < \nu_S\xi\left(\frac{CC_1C_2}{D}\right)\sum_{k=0}^{K}\lambda_\eta^{(\ell_k-\ell_{\min})} \tag{106}$$
$$< \xi\left(\frac{\nu_S CC_1C_2}{D}\right)\lambda_\infty, \tag{107}$$

where the last inequality comes from the non-negativity of λ_η (i.e., $\lambda_\infty > \sum_{k=0}^{K}\lambda_\eta^{(\ell_k-\ell_{\min})}$ for all finite K). By the definition of D in (45) in Assumption 4, we have

$$\|\mathbf{Q}^{-1}(\mathbf{x}_i^K)\mathbf{S}_{K,i}\| \le \|\mathbf{Q}^{-1}(\mathbf{x}_i^K)\|\,\|\mathbf{S}_{K,i}\| \tag{108}$$
$$\le \underbrace{\frac{I}{\sigma_{\min}^2}\left(\frac{\nu_S CC_1C_2}{D}\right)\lambda_\infty}_{<1,\ \text{from (45)}}\xi < \xi < 1/2,$$

where we have used the fact that $\|\mathbf{Q}^{-1}(\mathbf{x}_i^K)\| \le I/\sigma_{\min}^2$ (see (91)). Therefore, the matrix series expansion holds for (105). Then using the above calculations, we have

$$\|\bar{\mathbf{H}}_K^{-1}\| \le \|\mathbf{Q}^{-1}(\mathbf{x}_i^K)\| \tag{109}$$
$$+\sum_{q=1}^{\infty}\left(\|\mathbf{Q}^{-1}(\mathbf{x}_i^K)\|\,\|\mathbf{S}_{K,i}\|\right)^q\|\mathbf{Q}^{-1}(\mathbf{x}_i^K)\|$$
$$\le \frac{I}{\sigma_{\min}^2} + \frac{I}{\sigma_{\min}^2}\sum_{q=1}^{\infty}\xi^q$$
$$= \frac{I}{\sigma_{\min}^2}\left(1+\frac{\xi}{1-\xi}\right) = \frac{I}{\sigma_{\min}^2}\frac{1}{1-\xi} < \frac{2I}{\sigma_{\min}^2}.$$

Similar to the case with $k=1$, we have

$$\|\bar{\mathbf{H}}_K^{-1}\mathbf{E}_{K,i}(\ell_K)\| \le \|\bar{\mathbf{H}}_K^{-1}\|\,\|\mathbf{E}_{K,i}(\ell_K)\| \le \|\bar{\mathbf{H}}_K^{-1}\|\,\|\mathbf{E}_K(\ell_K)\|_F$$
$$\|\bar{\mathbf{H}}_K^{-1}\mathbf{E}_{K,j}(\ell_K)\| \le \|\bar{\mathbf{H}}_K^{-1}\|\,\|\mathbf{E}_{K,j}(\ell_K)\| \le \|\bar{\mathbf{H}}_K^{-1}\|\,\|\mathbf{E}_K(\ell_K)\|_F.$$

From Lemma 3 and Assumption 4, the above bound can be further bounded using (109) as

$$\|\bar{\mathbf{H}}_K^{-1}\|\,\|\mathbf{E}_K(\ell_K)\|_F \le \frac{2I}{\sigma_{\min}^2}C\lambda_\eta^{\ell_K} = \lambda_\eta^{(\ell_K-\ell_{\min})}\frac{IC}{2\sigma_{\min}^2 D}\xi.$$

For notation convenience, we again let $\tilde{\xi} = IC\xi/(\sigma_{\min}^2 D)$ in (93) with $\tilde{\xi} < \xi < 1/2$ and let $\delta\mathbf{Z} = \mathbf{E}_{K,i}(\ell_K)$ or $\mathbf{E}_{K,j}(\ell_K)$ and $\mathbf{Z} = \bar{\mathbf{H}}_K$. As a result, we have

$$\|\mathbf{Z}^{-1}\delta\mathbf{Z}\| = \|\bar{\mathbf{H}}_K^{-1}\|\,\|\mathbf{E}_K(\ell_K)\|_F \tag{110}$$
$$\le \frac{1}{2}\lambda_\eta^{(\ell_K-\ell_{\min})}\tilde{\xi} < \frac{1}{4}. \tag{111}$$

Therefore, the matrix expansion holds. By grouping all the high order terms $q\ge 1$ in the matrix expansion, we have

$$\mathbf{d}_i^K(\ell_K) - \mathbf{d}_j^K(\ell_K)$$
$$= \left[\bar{\mathbf{H}}_K^{-1} - \sum_{q=1}^{\infty}(-1)^q\left(\bar{\mathbf{H}}_K^{-1}\mathbf{E}_{K,i}(\ell_K)\right)^q\bar{\mathbf{H}}_K^{-1}\right]\left[\bar{\mathbf{h}}_K + \mathbf{e}_{K,i}(\ell_K)\right]$$
$$- \left[\bar{\mathbf{H}}_K^{-1} - \sum_{q=1}^{\infty}(-1)^q\left(\bar{\mathbf{H}}_K^{-1}\mathbf{E}_{K,j}(\ell_K)\right)^q\bar{\mathbf{H}}_K^{-1}\right]\left[\bar{\mathbf{h}}_K + \mathbf{e}_{K,j}(\ell_K)\right].$$

To simplify the above expression, we write it in three terms $\mathbf{D}_1(\ell_K)$, $\mathbf{D}_2(\ell_K)$ and $\mathbf{D}_3(\ell_K)$ as follows

$$\mathbf{d}_i^K(\ell_K) - \mathbf{d}_j^K(\ell_K) = \mathbf{D}_1(\ell_K) + \mathbf{D}_2(\ell_K) + \mathbf{D}_3(\ell_K), \quad (112)$$

where $\mathbf{D}_1(\ell_K) \triangleq \bar{\mathbf{H}}_K^{-1}\left[\mathbf{e}_{K,i}(\ell_K) - \mathbf{e}_{K,j}(\ell_K)\right]$ and

$$\begin{aligned}\mathbf{D}_2(\ell_K) &\triangleq \sum_{q=1}^{\infty}(-1)^q\left(\bar{\mathbf{H}}_K^{-1}\mathbf{E}_{K,j}(\ell_K)\right)^q\bar{\mathbf{H}}_K^{-1}\bar{\mathbf{h}}_K \\ &\quad - \sum_{q=1}^{\infty}(-1)^q\left(\bar{\mathbf{H}}_K^{-1}\mathbf{E}_{K,i}(\ell_K)\right)^q\bar{\mathbf{H}}_K^{-1}\bar{\mathbf{h}}_K \\ \mathbf{D}_3(\ell_K) &\triangleq \sum_{q=1}^{\infty}(-1)^q\left(\bar{\mathbf{H}}_K^{-1}\mathbf{E}_{K,j}(\ell_K)\right)^q\bar{\mathbf{H}}_K^{-1}\mathbf{e}_{K,j}(\ell_K) \\ &\quad - \sum_{q=1}^{\infty}(-1)^q\left(\bar{\mathbf{H}}_K^{-1}\mathbf{E}_{K,i}(\ell_K)\right)^q\bar{\mathbf{H}}_K^{-1}\mathbf{e}_{K,i}(\ell_K).\end{aligned}$$

2) Proof of success when $k = K+1$: According to the triangular inequality for norms, we can bound

$$\begin{aligned}\|\mathbf{e}_{K,i}(\ell_K) - \mathbf{e}_{K,j}(\ell_K)\| &\le 2\|\mathbf{e}_K(\ell_K)\| \\ \|\mathbf{E}_{K,i}(\ell_K) - \mathbf{E}_{K,j}(\ell_K)\| &\le 2\|\mathbf{E}_K(\ell_K)\|_F\end{aligned}$$

Using (110), we can bound the norm of the first term as

$$\|\mathbf{D}_1(\ell_K)\| \le 2\left\|\bar{\mathbf{H}}_K^{-1}\right\|\|\mathbf{e}_K(\ell_K)\| \le \tilde{\xi}\lambda_\eta^{(\ell_K-\ell_{\min})}. \quad (113)$$

Similarly, the infinite sum in the second term is bounded as

$$\begin{aligned}&\left\|\sum_{q=1}^{\infty}(-1)^q\left[\left(\bar{\mathbf{H}}_K^{-1}\mathbf{E}_{K,j}(\ell_K)\right)^q - \left(\bar{\mathbf{H}}_K^{-1}\mathbf{E}_{K,i}(\ell_K)\right)^q\right]\right\| \\ &\le 2\sum_{q=1}^{\infty}\left(\left\|\bar{\mathbf{H}}_K^{-1}\right\|\|\mathbf{E}_K(\ell_K)\|_F\right)^q \quad (114) \\ &\le 2\sum_{q=1}^{\infty}\left(\frac{1}{2}\tilde{\xi}\lambda_\eta^{(\ell_K-\ell_{\min})}\right)^q = \frac{\tilde{\xi}\lambda_\eta^{(\ell_K-\ell_{\min})}}{\left(1-\frac{1}{2}\tilde{\xi}\lambda_\eta^{(\ell_K-\ell_{\min})}\right)},\end{aligned}$$

where the last equality comes from the convergence of geometric series $\lim_{K\to\infty}\sum_{k=1}^K a^k = a/(1-a)$ for any $|a|<1$. Since $0<\tilde{\xi}<\xi<1/2$ and $\lambda_\eta^{(\ell_K-\ell_{\min})}<1$, then

$$\frac{\tilde{\xi}\lambda_\eta^{(\ell_K-\ell_{\min})}}{\left(1-\frac{1}{2}\tilde{\xi}\lambda_\eta^{(\ell_K-\ell_{\min})}\right)} < 2\tilde{\xi}\lambda_\eta^{(\ell_K-\ell_{\min})} \quad (115)$$

and thus the norm of the second term is bounded as

$$\begin{aligned}\|\mathbf{D}_2(\ell_K)\| &\le 2\tilde{\xi}\lambda_\eta^{(\ell_K-\ell_{\min})}\left\|\bar{\mathbf{H}}_K^{-1}\right\|\left\|\bar{\mathbf{h}}_K\right\| \quad (116) \\ &\le \frac{2\sigma_{\max}\epsilon_{\max}}{\sigma_{\min}^2}\tilde{\xi}\lambda_\eta^{(\ell_K-\ell_{\min})}, \quad (117)\end{aligned}$$

where the last inequality comes from (91). Following the same rationale, the norm of the third term can be bounded as

$$\begin{aligned}\|\mathbf{D}_3(\ell_K)\| &\le 2\sum_{q=1}^{\infty}\left(\left\|\bar{\mathbf{H}}_K^{-1}\right\|\|\mathbf{E}_K(\ell_K)\|_F\right)^q\left\|\bar{\mathbf{H}}_K^{-1}\mathbf{e}_K(\ell_K)\right\| \\ &\le 2\sum_{q=1}^{\infty}\left(\left\|\bar{\mathbf{H}}_K^{-1}\right\|\|\mathbf{E}_K(\ell_K)\|_F\right)^{q+1} \\ &\le 2\sum_{q=1}^{\infty}\left(\frac{1}{2}\tilde{\xi}\lambda_\eta^{(\ell_K-\ell_{\min})}\right)^{q+1} \\ &= \frac{\tilde{\xi}\lambda_\eta^{(\ell_K-\ell_{\min})}}{\left(1-\frac{1}{2}\tilde{\xi}\lambda_\eta^{(\ell_K-\ell_{\min})}\right)}\cdot\frac{1}{2}\tilde{\xi}\lambda_\eta^{(\ell_K-\ell_{\min})} < \tilde{\xi}\lambda_\eta^{(\ell_K-\ell_{\min})}.\end{aligned}$$

where the last inequality has used the results in (115). Note that this is again a very loose bound.

Substituting $\tilde{\xi} = IC\xi/(\sigma_{\min}^2 D)$ in (93) back to (113), (114) and (118) and using the constants C_1 and C_2, we have

$$\left\|\mathbf{d}_i^K(\ell_K) - \mathbf{d}_j^K(\ell_K)\right\| \le \xi\left(\frac{CC_1C_2}{D}\right)\lambda_\eta^{(\ell_K-\ell_{\min})}$$

Similarly, based on (47) and (90), we have

$$\begin{aligned}&\left\|\mathbf{x}_i^{K+1} - \mathbf{x}_j^{K+1}\right\| \\ &\le \left\|\mathbf{x}_i^K - \mathbf{x}_j^K\right\| + \left\|\mathbf{d}_i^K(\ell_K) - \mathbf{d}_j^K(\ell_K)\right\| \\ &\le \xi\left(\frac{CC_1C_2}{D}\right)\sum_{k=0}^{K-1}\lambda_\eta^{(\ell_k-\ell_{\min})} + \xi\left(\frac{CC_1C_2}{D}\right)\lambda_\eta^{(\ell_K-\ell_{\min})} \\ &= \xi\left(\frac{CC_1C_2}{D}\right)\sum_{k=0}^{K}\lambda_\eta^{(\ell_k-\ell_{\min})},\end{aligned}$$

and therefore given that the recursion holds for $k=K$, it holds true for $k=K+1$. The induction is complete. Given (45), we have $\xi \le 4D\lambda_\eta^{(\ell_{\min}+1)}$, and

$$\left\|\mathbf{x}_i^{K+1} - \mathbf{x}_j^{K+1}\right\| \le 4CC_1C_2\sum_{k=0}^{K}\lambda_\eta^{\ell_k+1}. \quad (118)$$

Appendix E
Proof of Theorem 2

By the decomposition in (41), we have

$$\begin{aligned}&\mathbf{d}_i^k(\ell_k) - \mathbf{d}_i^k \quad (119) \\ &= \left[\mathbf{Q}(\mathbf{x}_i^k) + \mathbf{S}_{k,i} + \mathbf{E}_{k,i}(\ell_k)\right]^{-1}\left[\mathbf{q}(\mathbf{x}_i^k) + \mathbf{s}_{k,i} + \mathbf{e}_{k,i}(\ell_k)\right] \\ &\quad - \mathbf{Q}(\mathbf{x}_i^k)^{-1}\mathbf{q}(\mathbf{x}_i^k).\end{aligned}$$

Now that we verify that the matrix series expansion holds for similar approximations. First of all, from Lemma 4 and in particular (106), we have $\|\mathbf{S}_{k,i}(\ell_k)\| \le \nu_S C_1\lambda_\infty\xi/D$. The expansion depends on the quantity

$$\begin{aligned}&\left\|\mathbf{Q}^{-1}(\mathbf{x}_i^k)\left(\mathbf{S}_{k,i} + \mathbf{E}_{k,i}(\ell_k)\right)\right\| \\ &\le \left\|\mathbf{Q}^{-1}(\mathbf{x}_i^k)\right\|\|\mathbf{S}_{k,i}\| + \left\|\mathbf{Q}^{-1}(\mathbf{x}_i^k)\right\|\|\mathbf{E}_{K,i}(\ell_k)\|.\end{aligned}$$

Using the derivation in (108) and C_1, C_2 in (47), we have

$$\begin{aligned}
&\left\|\mathbf{Q}^{-1}(\mathbf{x}_i^k)\left(\mathbf{S}_{k,i}+\mathbf{E}_{k,i}(\ell_k)\right)\right\| \\
&< C_2\left(\frac{\nu_S C C_1 C_2}{D}\right)\lambda_\infty \xi + \left(\frac{CC_2}{4D}\right)\xi\lambda_\eta^{(\ell_k-\ell_0)} \\
&= \frac{CC_2}{D}\left(\nu_S C_1 C_2 \lambda_\infty + \frac{1}{4}\lambda_\eta^{(\ell_k-\ell_0)}\right)\xi \\
&< \underbrace{\frac{CC_2\left(\nu\lambda_\infty C_1 C_2+1\right)}{D}}_{=1,\ \text{from (45)}}\xi < \xi < \frac{1}{2},
\end{aligned} \tag{120}$$

where the last inequality is by the definition of D in (45). Then (119) can be re-written as

$$\begin{aligned}
&\mathbf{d}_i^k(\ell_k) - \mathbf{d}_i^k \\
&= \left[\mathbf{Q}^{-1}(\mathbf{x}_i^k) - \sum_{q=1}^{\infty}\left(\mathbf{Q}^{-1}(\mathbf{x}_i^k)(\mathbf{S}_{k,i}+\mathbf{E}_{k,i}(\ell_k))\right)^q \mathbf{Q}^{-1}(\mathbf{x}_i^k)\right] \\
&\quad \times\left[\mathbf{q}(\mathbf{x}_i^k)+\mathbf{s}_{k,i}+\mathbf{e}_{k,i}(\ell_k)\right] - \mathbf{Q}^{-1}(\mathbf{x}_i^k)\mathbf{q}(\mathbf{x}_i^k) \\
&= \mathbf{Q}^{-1}(\mathbf{x}_i^k)\left[\mathbf{s}_{k,i}+\mathbf{e}_{k,i}(\ell_k)\right] \\
&\quad - \sum_{q=1}^{\infty}\left(\mathbf{Q}^{-1}(\mathbf{x}_i^k)(\mathbf{S}_{k,i}+\mathbf{E}_{k,i}(\ell_k))\right)^q \mathbf{Q}^{-1}(\mathbf{x}_i^k)\mathbf{q}(\mathbf{x}_i^k) \\
&\quad - \sum_{q=1}^{\infty}\left(\mathbf{Q}^{-1}(\mathbf{x}_i^k)(\mathbf{S}_{k,i}+\mathbf{E}_{k,i}(\ell_k))\right)^q \mathbf{Q}^{-1}(\mathbf{x}_i^k)\left[\mathbf{s}_{k,i}+\mathbf{e}_{k,i}(\ell_k)\right]
\end{aligned}$$

According to Lemma 4 and Assumption 4, we have $\|\mathbf{s}_{k,i}(\ell_k)\| \le \nu_s C C_1 C_2 \lambda_\infty \xi / D$, and the norm of the first term above can be bounded similarly as (120)

$$\left\|\mathbf{Q}^{-1}(\mathbf{x}_i^k)\left[\mathbf{s}_{k,i}+\mathbf{e}_{k,i}(\ell_k)\right]\right\| < \xi. \tag{121}$$

Likewise, the norm of the second term is bounded as

$$\begin{aligned}
&\left\|\sum_{q=1}^{\infty}\left(\mathbf{Q}^{-1}(\mathbf{x}_i^k)(\mathbf{S}_{k,i}+\mathbf{E}_{k,i}(\ell_k))\right)^q \mathbf{Q}^{-1}(\mathbf{x}_i^k)\mathbf{q}(\mathbf{x}_i^k)\right\| \\
&\le \sum_{q=1}^{\infty}\left(\left\|\mathbf{Q}^{-1}(\mathbf{x}_i^k)(\mathbf{S}_{k,i}+\mathbf{E}_{k,i}(\ell_k))\right\|\right)^q \left\|\mathbf{Q}^{-1}(\mathbf{x}_i^k)\mathbf{q}(\mathbf{x}_i^k)\right\| \\
&< \frac{\sigma_{\max}\epsilon_{\max}}{\sigma_{\min}^2}\sum_{q=1}^{\infty}\xi^q = \frac{\sigma_{\max}\epsilon_{\max}}{\sigma_{\min}^2}\frac{\xi}{1-\xi} < 2\frac{\sigma_{\max}\epsilon_{\max}}{\sigma_{\min}^2}\xi,
\end{aligned}$$

and similarly for the third term

$$\begin{aligned}
&\left\|\sum_{q=1}^{\infty}\left(\mathbf{Q}^{-1}(\mathbf{x}_i^k)(\mathbf{S}_{k,i}+\mathbf{E}_{k,i}(\ell_k))\right)^q \mathbf{Q}^{-1}(\mathbf{x}_i^k)\left[\mathbf{s}_{k,i}+\mathbf{e}_{k,i}(\ell_k)\right]\right\| \\
&< \sum_{q=1}^{\infty}\left\|\mathbf{Q}^{-1}(\mathbf{x}_i^k)(\mathbf{S}_{k,i}+\mathbf{E}_{k,i}(\ell_k))\right\|^{q+1} < \sum_{q=1}^{\infty}\xi^{q+1} = \frac{\xi^2}{1-\xi}.
\end{aligned}$$

Furthermore, since $\xi \in (0, 1/2)$, the above expression can be simplified as $\xi^2/(1-\xi) < 2\xi$. Finally, summing them up and using the constant C_1 we have

$$\left\|\mathbf{d}_i^k(\ell_k) - \mathbf{d}_i^k\right\| \le 2\left(1+\frac{\sigma_{\max}\epsilon_{\max}}{\sigma_{\min}^2}\right)\xi = C_1\xi \tag{122}$$

for all i and k. Now we have established that the discrepancy between the decentralized descent and the exact descent can be bounded by an arbitrarily small error ξ specified by the system. Given (45), we have

$$\xi < 4D\lambda_\eta^{(\ell_{\min}+1)}, \tag{123}$$

and therefore, the perturbation bound κ on the error recursion in Lemma 2 can be obtained as

$$\kappa \triangleq 4C_1 D\lambda_\eta^{(\ell_{\min}+1)}. \tag{124}$$

References

[1] J. Nocedal and S. Wright, *Numerical Optimization*. Springer verlag, 1999.

[2] J. Dennis and R. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Society for Industrial Mathematics, 1996, vol. 16.

[3] Å. Björck, *Numerical Methods for Least Squares Problems*. Society for Industrial Mathematics, 1996, no. 51.

[4] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge Univ Pr, 2004.

[5] A. Monticelli, "Electric Power System State Estimation," *Proceedings of the IEEE*, vol. 88, no. 2, pp. 262–282, 2000.

[6] C. Mensing and S. Plass, "Positioning Algorithms for Cellular Networks using TDOA," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 4. IEEE, 2006, pp. IV–IV.

[7] P. Stoica, R. Moses, B. Friedlander, and T. Soderstrom, "Maximum Likelihood Estimation of the Parameters of Multiple Sinusoids from Noisy Measurements," *Acoustics, Speech and IEEE Trans. Signal Process.*, vol. 37, no. 3, pp. 378–392, 1989.

[8] B. Bell and F. Cathey, "The Iterated Kalman Filter Update as a Gauss-Newton Method," *Automatic Control, IEEE Transactions on*, vol. 38, no. 2, pp. 294–297, 1993.

[9] M. Schweiger, S. Arridge, and I. Nissilä, "Gauss-Newton method for Image Reconstruction in Diffuse Optical Tomography," *Physics in medicine and biology*, vol. 50, p. 2365, 2005.

[10] J. Tsitsiklis, "Problems in Decentralized Decision Making and Computation." DTIC Document, Tech. Rep., 1984.

[11] R. Karp, C. Schindelhauer, S. Shenker, and B. Vocking, "Randomized Rumor Spreading," in *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*. IEEE, 2000, pp. 565–574.

[12] R. Olfati-Saber and R. Murray, "Consensus Problems in Networks of Agents with Switching Topology and Time-Delays," *Automatic Control, IEEE Transactions on*, vol. 49, no. 9, pp. 1520–1533, 2004.

[13] A. Dimakis, S. Kar, J. Moura, M. Rabbat, and A. Scaglione, "Gossip Algorithms for Distributed Signal Processing," *Proc. IEEE*, vol. 98, no. 11, pp. 1847–1864, 2010.

[14] D. Kempe, A. Dobra, and J. Gehrke, "Gossip-based Computation of Aggregate Information," in *Foundations of Computer Science, 2003. Proceedings. 44th Annual IEEE Symposium on*. IEEE, 2003, pp. 482–491.

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

[16] R. Olfati-Saber, J. Fax, and R. Murray, "Consensus and Cooperation in Networked Multi-agent Systems," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 215–233, 2007.

[17] B. Johansson, T. Keviczky, M. Johansson, and K. Johansson, "Subgradient Methods and Consensus Algorithms for Solving Convex Optimization Problems," in *Decision and Control, 2008. CDC 2008. 47th IEEE Conference on*. IEEE, 2008, pp. 4185–4190.

[18] D. Bertsekas, M. I. of Technology. Laboratory for Information, and D. Systems, "A New Class of Incremental Gradient Methods for Least Squares Problems," *SIAM Journal on Optimization*, vol. 7, no. 4, pp. 913–926, 1997.

[19] A. Nedic and D. Bertsekas, "Incremental Subgradient Methods for Nondifferentiable Optimization," *SIAM Journal of Optimization*, vol. 12, no. 1, pp. 109–138, 2001.

[20] A. Nedic and A. Ozdaglar, "Distributed Subgradient Methods for Multi-agent Optimization," *Automatic Control, IEEE Transactions on*, vol. 54, no. 1, pp. 48–61, 2009.

[21] S. Ram, A. Nedic, and V. Veeravalli, "Distributed Stochastic Subgradient Projection Algorithms for Convex Optimization," *Journal of optimization theory and applications*, vol. 147, no. 3, pp. 516–545, 2010.

[22] A. Nedic, "Asynchronous Broadcast-based Convex Optimization over a Network," *Automatic Control, IEEE Transactions on*, no. 99, pp. 1–1, 2010.

[23] K. Srivastava and A. Nedic, "Distributed Asynchronous Constrained Stochastic Optimization," *Selected Topics in Signal Processing, IEEE Journal of*, no. 99, pp. 1–1, 2011.

[24] S. Kar, J. Moura, and K. Ramanan, "Distributed Parameter Estimation in Sensor Networks: Nonlinear Observation Models and Imperfect Communication," *Information Theory, IEEE Transactions on*, vol. 58, no. 6, pp. 3575 –3605, june 2012.

[25] I. Matei and J. Baras, "Performance Evaluation of the Consensus-based Distributed Subgradient Method under Random Communication Topologies," *Selected Topics in Signal Processing, IEEE Journal of*, no. 99, pp. 1–1, 2011.

[26] J. Chen and A. Sayed, "Diffusion Adaptation Strategies for Distributed Optimization and Learning over Networks," *Signal Processing, IEEE Transactions on*, vol. 60, no. 8, pp. 4289–4305, 2012.

[27] C. Lopes and A. Sayed, "Diffusion Least-Mean Squares over Adaptive Networks: Formulation and Performance Analysis," *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 3122–3136, 2008.

[28] F. Cattivelli and A. Sayed, "Diffusion LMS Strategies for Distributed Estimation," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1035–1048, 2010.

[29] F. Cattivelli, C. Lopes, and A. Sayed, "Diffusion Recursive Least-Squares for Distributed Estimation over Adaptive Networks," *IEEE Trans. Signal Process.*, vol. 56, no. 5, pp. 1865–1877, 2008.

[30] E. Wei, A. Ozdaglar, and A. Jadbabaie, "A Distributed Newton Method for Network Utility Maximization," in *Decision and Control (CDC), 2010 49th IEEE Conference on*. IEEE, 2010, pp. 1816–1821.

[31] M. Ilic and A. Hsu, "Toward Distributed Contingency Screening using Line Flow Calculators and Dynamic Line Rating Units (DLRs)," in *2012 45th Hawaii International Conference on System Sciences*. IEEE, 2012, pp. 2027–2035.

[32] B. Bejar, P. Belanovic, and S. Zazo, "Distributed Gauss-Newton Method for Localization in Ad-hoc Networks," in *Signals, Systems and Computers (ASILOMAR), 2010 Conference Record of the Forty Fourth Asilomar Conference on*. IEEE, 2010, pp. 1452–1454.

[33] B. Cheng, R. Hudson, F. Lorenzelli, L. Vandenberghe, and K. Yao, "Distributed Gauss-Newton Method for Node Localization in Wireless Sensor Networks," in *Signal Processing Advances in Wireless Communications, 2005 IEEE 6th Workshop on*. IEEE, 2005, pp. 915–919.

[34] G. Calafiore, L. Carlone, and M. Wei, "A Distributed Gauss-Newton Approach for Range-based Localization of Multi-agent Formations," in *Computer-Aided Control System Design (CACSD), 2010 IEEE International Symposium on*. IEEE, 2010, pp. 1152–1157.

[35] T. Zhao and A. Nehorai, "Information-Driven Distributed Maximum Likelihood Estimation based on Gauss-Newton Method in Wireless Sensor Networks," *IEEE Trans. Signal Process.*, vol. 55, no. 9, pp. 4669–4682, 2007.

[36] F. Schweppe and E. Handschin, "Static State Estimation in Electric Power Systems," *Proceedings of the IEEE*, vol. 62, no. 7, pp. 972–982, 1974.

[37] R. Larson, W. Tinney, and J. Peschon, "State Estimation in Power Systems Part I: Theory and Feasibility," *IEEE Trans. Power App. Syst.*, no. 3Part-I, pp. 345–352, 1970.

[38] C. Brice and R. Cavin, "Multiprocessor Static State Estimation," *IEEE Trans. Power App. Syst.*, no. 2, pp. 302–308, 1982.

[39] M. Kurzyn, "Real-Time State Estimation for Large-Scale Power Systems," *IEEE Trans. Power App. Syst.*, no. 7, pp. 2055–2063, 1983.

[40] T. Yang, H. Sun, and A. Bose, "Transition to a Two-Level Linear State Estimator : Part i & ii," *IEEE Trans. Power Syst.*, no. 99, pp. 1–1, 2011.

[41] A. Gómez-Expósito, A. Abur, A. de la Villa Jaén, and C. Gómez-Quiles, "A Multilevel State Estimation Paradigm for Smart Grids," *Proceedings of the IEEE*, no. 99, pp. 1–25, 2011.

[42] D. Falcao, F. Wu, and L. Murphy, "Parallel and Distributed State Estimation," *IEEE Trans. Power Syst.*, vol. 10, no. 2, pp. 724–730, 1995.

[43] S. Lin, "A Distributed State Estimator for Electric Power Systems," *IEEE Trans. Power Syst.*, vol. 7, no. 2, pp. 551–557, 1992.

[44] R. Ebrahimian and R. Baldick, "State Estimation Distributed Processing," *IEEE Trans. Power Syst.*, vol. 15, no. 4, pp. 1240–1246, 2000.

[45] T. Van Cutsem, J. Horward, and M. Ribbens-Pavella, "A Two-Level Static State Estimator for Electric Power Systems," *IEEE Trans. Power App. Syst.*, no. 8, pp. 3722–3732, 1981.

[46] L. Zhao and A. Abur, "Multi-area State Estimation using Synchronized Phasor Measurements," *IEEE Trans. Power Syst.*, vol. 20, no. 2, pp. 611–617, 2005.

[47] W. Jiang, V. Vittal, and G. Heydt, "A Distributed State Estimator Utilizing Synchronized Phasor Measurements," *IEEE Trans. Power Syst.*, vol. 22, no. 2, pp. 563–571, 2007.

[48] L. Xie, D. Choi, S. Kar, and H. Poor, "Fully Distributed State Estimation for Wide-Area Monitoring Systems," *Smart Grid, IEEE Transactions on*, vol. 3, no. 3, pp. 1154–1169, 2012.

[49] V. Kekatos and G. Giannakis, "Distributed Robust Power System State Estimation," *Arxiv preprint arXiv:1204.0991*, 2012.

[50] X. Li, A. Scaglione, and T.-H. Chang, "Optimal Sensor Placement for Hybrid State Estimation in Smart Grid," *Acoustics, Speech and Signal Processing, 2013. ICASSP 2013 Proceedings. 2013 IEEE International Conference on*.

[51] R. Horn and C. Johnson, "Topics in Matrix Analysis, 1991."

[52] K. Eriksson, D. Estep, and C. Johnson, *Applied Mathematics, Body and Soul: Derivates and Geometry in* $\mathbb{R}^3$. Springer Verlag, 2004, vol. 3.

[53] V. Blondel, J. Hendrickx, A. Olshevsky, and J. Tsitsiklis, "Convergence in Multiagent Coordination, Consensus, and Flocking," in *Decision and Control, 2005 and 2005 European Control Conference. CDC-ECC'05. 44th IEEE Conference on*. IEEE, 2005, pp. 2996–3000.

[54] A. Monticelli, *State Estimation in Electric Power Systems: a Generalized Approach*. Springer, 1999, vol. 507.

[55] "U. K. National Grid-Real Time Operational Data," 2009, [Online; accessed 22-July-2004]. [Online]. Available: http://www.nationalgrid.com/uk/Electricity/Data/

[56] S. Salzo and S. Villa, "Convergence Analysis of a Proximal Gauss-Newton Method," *Arxiv preprint arXiv:1103.0414*, 2011.

[57] O. Galor, *Discrete Dynamical Systems*. Springer, 2007.

Xiao Li has been pursuing the Ph.D. degree at the University of California, Davis, since 2009. He received the B.Eng. degree (2007) from Sun Yat-Sen (Zhongshan) University, China, and the M.Phil. degree (2009) from the University of Hong Kong. His research interests are in the theoretical and algorithmic studies in signal processing and optimizations, statistical learning and inferences for high dimensional data, distributed optimizations and adaptive algorithms, as well as their applications in communications, networked systems, and smart grid.

Anna Scaglione (IEEE F10') received the Laurea (M.Sc. degree) in 1995 and the Ph.D. degree in 1999 from the University of Rome, "La Sapienza." She is currently Professor in Electrical and Computer Engineering at University of California at Davis, where she joined in 2008. She was previously at Cornell University, Ithaca, NY, from 2001 where became Associate Professor in 2006; prior to joining Cornell she was Assistant Professor in the year 2000-2001, at the University of New Mexico.

She served as Associate Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS from 2002 to 2005, and serves since 2008 the Editorial Board of the IEEE TRANSACTIONS ON SIGNAL PROCESSING from 2008 to 2011, where she is Area Editor. Dr. Scaglione received the 2000 IEEE Signal Processing Transactions Best Paper Award the NSF Career Award in 2002 and she is co-recipient of the Ellersick Best Paper Award (MILCOM 2005) and the 2013 IEEE Donald G. Fink Prize Paper Award. Her expertise is in the broad area of signal processing for communication systems and networks. Her current research focuses on communication and wireless networks, sensors' systems for monitoring, control and energy management and network science.