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...
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 lowercase (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 $|x| = \sqrt{x^*x}$, where $x^*$ is the conjugate of $x$. The transpose, conjugate transpose, and inverse of a non-singular matrix $X$ are denoted by $X^T$, $X^H$, and $X^{-1}$, respectively. The inner product between two vectors $x, y \in \mathbb{C}^{N \times 1}$ is defined accordingly as $\langle x, y \rangle = \sum_{n=1}^{N} y_n^* x_n$. The $W$-weighted Euclidean norm of a vector $x$ is denoted by $\|x\|_W = \sqrt{x^H W x}$, and the conventional Euclidean norm is written as $\|x\|$. The 2-norm of a matrix $A$ is denoted by $\|A\|$ and the Frobenius norm is denoted by $\|A\|_F$. Given a matrix $A = [a_1, \cdots, a_N]$ where $a_n$ is a column vector, the vectorization operator is defined as $\text{vec}(A) = [a_1^T, \cdots, a_N^T]$. Finally, $I_N$ is an $N \times N$ identity matrix and $1_N$ is an $N$-dimensional all 1 vector.

II. Problem Statement

Let $x \in \mathbb{R}^N$ be an unknown parameter vector associated with a specific network, belonging to a compact convex set $\mathcal{X}$. The network objective is described by a vector-valued continuously differentiable function $g(x) = [g_1(x), \cdots, g_M(x)]^T$ 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

$$\hat{x} = \arg \min_{x \in \mathcal{X}} \|g(x)\|^2. \quad (1)$$

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

**Assumption 1.**

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

$$\|g(x)\| \leq \epsilon_{\text{max}}. \quad (2)$$

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

$$\sigma_{\text{min}} = \min_{x \in \mathcal{X}} \sqrt{\lambda_{\text{min}}(G^T(x)G(x))},$$

$$\sigma_{\text{max}} = \max_{x \in \mathcal{X}} \sqrt{\lambda_{\text{max}}(G^T(x)G(x))},$$

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

3) The Jacobian $G(x)$ satisfies the Lipschitz condition

$$\|G(x) - G(x')\| \leq \omega \|x - x'\|, \quad x, x' \in \mathcal{X},$$

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 $x^0$ and solves the NLLS problem iteratively [3]

$$x^{k+1} = P_\mathcal{X} [x^k - \alpha_k d^k], \quad k = 1, 2, \cdots, \quad (3)$$

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

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

where $G(x)$ is the $M \times N$ Jacobian matrix of $g(x)$. 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

$$G^T(x^*)g(x^*) = 0, \quad x^* \in \mathcal{X}. \quad (5)$$

Note that if $\alpha_k$ 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 $\alpha_k$ 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 $\hat{x}$ is one of the many stationary points $x^*$ and the focus in this paper is to study the local convergence property of the algorithm to a particular fixed point $\check{x} \in \mathcal{X}$.

B. Distributed Formulation

Although the convergence of centralized Gauss-Newton algorithms is well studied [3] under Assumption 1, it is not
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 $g_i : \mathbb{R}^N \rightarrow \mathbb{R}^{M_i}$ from (1), i.e.,

$$g(x) = [g_1^T(x), \ldots, g_I^T(x)]^T$$

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

$$\hat{x} = \arg \min_{x \in \mathbb{R}} \sum_{i=1}^I \|g_i(x)\|^2,$$  

(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 $G_i(x) = \partial g_i(x)/\partial x^T$ is a sub-matrix of the full Jacobian $G(x)$, then we have (cf. [51, Corollary 3.1.3])

$$\|G_i(x) - G_i(x')\| \leq \omega \|x - x'\|, \quad x, x' \in \mathbb{R},$$

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

$$\|G_i^T(x)g_i(x) - G_i^T(x')g_i(x')\| \leq \nu_s \|x - x'\|,$$

$$\|G_i^T(x)G_i(x) - G_i^T(x')G_i(x')\| \leq \nu_S \|x - x'\|,$$

where $\nu_s \geq \omega (\epsilon_{\max} + \sigma_{\max})$ and $\nu_S \geq 2 \sigma_{\max} \omega$ 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.

$$x_{k+1}^i = P_x [x_k^i - \alpha d_k^i],$$

(8)

where the exact decentralized descent is given by

$$d_k^i = [G^T(x_k^i)G(x_k^i)]^{-1} G^T(x_k^i)g(x_k^i).$$

(9)

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

$$G^T(x_k^i)G(x_k^i) = \sum_{j=1}^I G_j^T(x_k^i)G_j(x_k^i)$$

(10)

and

$$G^T(x_k^i)g(x_k^i) = \sum_{j=1}^I G_j^T(x_k^i)g_j(x_k^i),$$

(11)

while the $i$-th agent has only partial information available to compute $G_i^T(x_k^i)g_i(x_k^i)$ and $G_i^T(x_k^i)g_j(x_k^i)$. 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 “$\ell$”) for the network gossiping. All the network agents are assumed to have a synchronous clock that determines the time instants $\tau_k$ 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 $\ell = 1, \ldots, \ell_k$.

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 $\ell = 1, \ldots, \ell_k$ that takes place between every two updates.

#### A. Local Update Model

Let $x_k^i$ be the local iterate at the $i$-th agent after the $k$-th update. For convenience, let

$$q(x_k^i) = \frac{1}{I} \sum_{j=1}^I G_j^T(x_k^i)g_j(x_k^i),$$

(12)

$$Q(x_k^i) = \frac{1}{I} \sum_{j=1}^I G_j^T(x_k^i)G_j(x_k^i).$$

(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

$$d_k^i = Q^{-1}(x_k^i)q(x_k^i),$$

(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 $x_k^i$ at other agents $i \neq j$ as well as that each node only knows its own mapping $g_i$ and $G_j$. In fact, the available information at the $i$-th agent after the $k$-th Gauss-Newton update is $G_i^T(x_k^i)g_i(x_k^i)$ and $G_i^T(x_k^i)G_i(x_k^i)$. Therefore, we propose to use an average surrogate for $q(x_k^i)$ and $Q(x_k^i)$

$$\bar{q}_k = \frac{1}{I} \sum_{i=1}^I G_i^T(x_k^i)g_i(x_k^i),$$

(15)

$$\bar{H}_k = \frac{1}{I} \sum_{i=1}^I G_i^T(x_k^i)G_i(x_k^i),$$

(16)
which can be obtained via network gossiping.

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

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

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

$$\mathbf{h}_{k,i}(0) \triangleq \mathbf{G}^T_i \left( \mathbf{x}_i^k \right) \mathbf{g}_i(\mathbf{x}_i^k)$$

$$\mathbf{H}_{k,i}(0) \triangleq \mathbf{G}^T_i \left( \mathbf{x}_i^k \right) \mathbf{G}_i(\mathbf{x}_i^k).$$

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

After $\ell_k$ exchanges, the “approximate descent” for the $(k+1)$-th update at the $i$-th agent is

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

and the local estimate is updated as

$$\mathbf{x}_i^{k+1} = P_\beta \left[ \mathbf{x}_i^k - \alpha d^k_{i}(\ell_k) \right].$$

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 $[\tau_k, \tau_{k+1})$ for every GN update $k$ and gossip exchange $\ell$. The node set $\mathcal{I} = \{1, \ldots, I\}$ 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 $A_{ij}(\ell) = [A_{ij}^{(k,\ell)}]_{I \times I}$

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

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

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

There exists an integer $L \geq 1$ such that for any $\ell$

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

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 $[\tau_k, \tau_{k+1})$, 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 $\{i, j\} \in \mathcal{M}_{k,\ell}$.

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

1) $W_{ij}^{(k,\ell)} \geq \eta$ 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 $\mathcal{H}_{k}(\ell) \triangleq [\mathcal{H}_{1,k}(\ell), \ldots, \mathcal{H}_{I,k}(\ell)]^T$, the exchange model can be written compactly as

$$\mathcal{H}_{k}(\ell) = [\mathbf{W}_{k}(\ell) \otimes \mathbf{I}_{N\tau_k}] \mathcal{H}_{k}(\ell-1), \quad 1 \leq \ell \leq \ell_k,$$

where $\ell_k$ 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 $\mathcal{H}_{k} = \sum_{\ell=1}^{L} \mathcal{H}_{k,\ell}(0)/I$. 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 $[\tau_k, \tau_{k+1})$ for $\ell = 1, \ldots, \ell_k$. In particular, if the network is fully connected such that $\mathbf{A} = \mathbf{I}_I - 1/I_J^T$, 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 $\mathbf{L} = \text{diag}(\mathbf{A}_1) - \mathbf{A}$ as $\mathbf{W} = \mathbf{I}_I - w\mathbf{L}$ with $w = \beta/\max(\mathbf{A}_I)$ 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 $j$ to random communicate. Suppose agent $i_{k,\ell}$ wakes up at $[\tau_k, \tau_{k+1})$ 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

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

where $e_i$ 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 $\eta$ in Assumption 3, the entries of the matrix product $[\prod_{\ell=0}^{\ell_k} \mathbf{W}_{k}(\ell)]$ converge with a geometric rate uniformly for all $i,j \in \mathcal{I}$ and $k$

$$\left[ \prod_{\ell=0}^{L} \mathbf{W}_{k}(\ell) \right]_{ij} \leq 2 \left( \frac{1 + \eta^{-L_0}}{1 - \eta^{-L_0}} \right) \lambda^L,$$

3This 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.
Algorithm 1 Gossip-based Gauss-Newton (GGN) Algorithm
1: given initial variables $x_i^0$ at all agents $i \in \mathcal{I}$.
2: set $k = 0$.
3: repeat
4: set $k = k + 1$.
5: initialization: For $i \in \mathcal{I}$, each agent $i$ evaluates (18) and constructs $H_{k,i}(0)$ as (17);
6: network gossiping: Each agent $i$ exchanges information with neighbors via network gossiping as (24);
7: local update: For $i \in \mathcal{I}$, each agent $i$ updates its local variables as (21) and (20).
8: until $\|x_i^{k+1} - x_i^k\| \leq \epsilon$ or $k = K$.
9: set the local estimate as $\hat{x}_i = x_i^k$.

with $L_0 = (I - 1)L$ and
\[ \lambda_0 = (1 - \eta L_0)^{1/L_0} \in (0, 1). \] (27)

It is clear from Lemma 1 that the limit of the weight matrix product exists $\lim_{\ell \to \infty} \prod_{\ell' = 0}^{\ell} W_k(\ell') = 1_N 1_N^T / L_0$ and thus
\[ \lim_{\ell \to \infty} H_{k,i}(\ell) = \frac{1}{I} \sum_{i = 1}^{I} H_{k,i}(0), \quad k = 1, 2, \ldots \] (28)

which asymptotically leads to $\lim_{\ell \to \infty} d_k(\ell) = \bar{H}_k^{-1} \bar{H}_k$.

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 $\ell_k$ and the mismatch error by using the surrogates $H_k$ and $\bar{H}_k$ 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 $x_i^{k+1}$ and a fixed point in (5) satisfies the following recursion.

\begin{equation}
\|x_i^{k+1} - \bar{x}\| \leq T_1 \|x_i^k - \bar{x}\| + T_2 \|x_i^k - \bar{x}\| + \alpha \|d^k(\ell_k) - d^k_i\|,
\end{equation}

where
\begin{align}
T_1 & \triangleq \frac{\alpha \omega}{2\sigma_{min}}, \quad \epsilon_{min} \triangleq \|g(\bar{x})\| \tag{30} \\
T_2 & \triangleq (1 - \alpha) \frac{\sigma_{max}}{\sigma_{min}} + \frac{\sqrt{2\alpha \omega}}{\sigma_{min}} \tag{31}
\end{align}

Proof: See Appendix A.

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

Theorem 1. (Convergence with Bounded Perturbation) Let Assumption 1 hold and $\mathcal{X}$ be a compact convex set. Given a step-size chosen as
\[ \max \left\{ 1 - \frac{3\sigma_{min}}{\sigma_{max}}, 0 \right\} < \alpha \leq 1 \] (32)

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

we define the following
\[ \rho_{min} = \frac{(1 - T_2)^2 - \sqrt{(1 - T_2)^2 - 4\alpha T_1 \kappa}}{2T_1} \tag{34} \]
\[ \rho_{max} = \frac{(1 - T_2)^2 + \sqrt{(1 - T_2)^2 - 4\alpha T_1 \kappa}}{2T_1} \tag{35} \]
where $\kappa$ is a bounded perturbation with
\[ 0 < \kappa < (1 - T_2)^2 \frac{1}{4\alpha T_1} \tag{36} \]

If the $\|d^k(\ell_k) - d^k_i\| \leq \kappa$ is bounded for all $k$ and $i \in \mathcal{I}$, then given any $x_i^0$ that falls within the $\rho_{max}$-neighborhood of $\bar{x} \in \mathcal{X}$
\[ \|x_i^0 - \bar{x}\| < \rho_{max}, \tag{37} \]

the asymptotic error of the local iterate $x_i^{k+1}$ at each agent with respect to $\bar{x}$ can be bounded as
\[ \limsup_{k \to \infty} \|x_i^{k+1} - \bar{x}\| \leq \rho_{min}. \tag{38} \]

Proof: See Appendix B.

An intuition that can be drawn from the sufficient condition is that the smaller is the Lipschitz constant $\omega$, the larger is the region of convergence around the global solution $\bar{x}$ 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 $\bar{x}$ 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 $\kappa$

If the perturbation $\kappa$ 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 $\ell$-th exchange
\[ h_k(\ell) \triangleq [h_{k,1}(\ell), \ldots, h_{k,T}(\ell)]^T, \]
\[ H_k(\ell) \triangleq [H_{k,1}^T(\ell), \ldots, H_{k,T}^T(\ell)]^T \]
Lemma 3. Let Assumption 2 and 3 hold. The gossip error $e_k(ℓ)$ and $E_k(ℓ)$ after the $k$-th update can be bounded as

$$
\|e_k(ℓ)\| < C\lambda_k, \quad \|E_k(ℓ)\| < C\lambda_k,
$$

where

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

Proof: See Appendix C.

2) Mismatch of surrogates: Define the errors between the surrogate $h_k$, $H_k$ and exact quantities $q(\bar{x}_k^i)$ and $Q(\bar{x}_k^i)$ as

$$
s_{k,i} = \bar{h}_k - q(\bar{x}_k^i) = \frac{1}{T} \sum_{j=1}^{T} [h_{k,i}(ℓ) - h_{k,j}(ℓ)], \quad (40)
$$

$$
S_{k,i} = \bar{H}_k - Q(\bar{x}_k^i) = \frac{1}{T} \sum_{j=1}^{T} [H_{k,i}(ℓ) - H_{k,j}(ℓ)], \quad (42)
$$

which thus leads to

$$
h_{k,i}(ℓ) = q(\bar{x}_k^i) + s_{k,i} + e_{k,i}(ℓ), \quad (41)
$$

$$
H_{k,i}(ℓ) = Q(\bar{x}_k^i) + S_{k,i} + E_{k,i}(ℓ). \quad (42)
$$

By (40) and Corollary 1, we have

$$
\|s_{k,i}\| \leq \frac{\nu_s}{T} \sum_{j=1}^{T} \|x_k^i - x_j^k\| \quad (43)
$$

$$
\|S_{k,i}\| \leq \frac{\nu_S}{T} \sum_{j=1}^{T} \|x_k^i - x_j^k\|. \quad (44)
$$

Clearly, this discrepancy depends on the disagreement $\|x_k^i - x_j^k\|$ for each pair of $i$-th and $j$-th agents, characterized by the mismatch $S_{k,i}$ and $s_{k,i}$ which originates from the gossip errors $E_{k,i}(ℓ)$ and $e_{k,i}(ℓ)$. Having analyzed the gossip error dynamics in Lemma 3, in the following we bound the disagreement $\|x_k^i - x_j^k\|$.

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

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

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

$$
\ell_{\min} = \left\lfloor \frac{\log \left( \frac{\xi}{4D} \right)}{\log \lambda_\infty} \right\rfloor, \quad (45)
$$

$$
D \triangleq CC_2(\nu\lambda_\infty C_1C_2 + 1). \quad (46)
$$

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

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

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

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 $x_0^i = x_0^j$, the deviation $\|x^K - x^K\|$ for any $i$ and $j$ at the $K$-th update satisfies

$$
\|x^K_i - x^K_j\| \leq \xi \left( \frac{CC_1C_2}{D} \right) \sum_{k=0}^{K-1} \lambda_k^{(\ell_k - \ell_{\min})}, \quad (48)
$$

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

$$
\|x^K_i - x^K_j\| \leq 4CC_1C_2 \sum_{k=0}^{K-1} \lambda_k^{(\ell_k + 1)}. \quad (49)
$$

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 $k$

$$
\|d^K_i(\ell_k) - d^K_i\| < \kappa \quad (49)
$$

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 $\ell_k$’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 $\mathcal{N} \triangleq \{1, \ldots, N\}$. 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 $Y = [-Y_{nm}]_{N \times N}$, where $Y_{nm} = G_{nm} + jB_{nm}$, $\{n,m\} \in \mathcal{E}$ is the line admittance, and the shunt admittance $Y_{nm} = G_{nm} + jB_{nm}$ associated with the II-model of each transmission line $\{n,m\} \in \mathcal{E}$. Note that $Y_{nm} = -\sum_{l \neq n} (Y_{nm} + Y_{nm})$ 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 $\mathbf{x} = [\Theta^T, \mathbf{V}^T]^T$, where $\Theta \triangleq [\theta_1, \ldots, \theta_N]^T$ is the phase vector with $\theta_1$ being the slack bus phase, and $\mathbf{V} \triangleq [V_1, \ldots, V_N]^T$ contains the magnitude.

3The II-model is a circuit equivalent of a transmission line by abstracting two electric buses as a two-port network in the shape of a II connection [54].
where \( f_i(x) \triangleq T_{i}f(x) \), and \( T_{i} \triangleq \text{diag}[T_{i,k},T_{i,l}] \) is a block diagonal binary matrix selecting the corresponding measurements at the \( i \)-th site. Specifically, \( T_{i,k} \in \{0,1\}^{M_i, n \times 2N} \) and \( T_{i,l} \in \{0,1\}^{M_i, n \times 4L} \) are selection matrices with each row having only one “1” entry located at the index of the corresponding element in \( f(x) \) measured by field devices. The number of measurements recorded by each agent is \( M_i = M_{i,k} + M_{i,l} = \text{Tr}(T_{i,k}^T T_{i,l}) \).

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 \( \mathbb{E}\{e^T e\} = \sigma^2 I \), the state is estimated as

\[
\hat{x} = \min_{x \in \mathbb{R}} \sum_{i=1}^{I} \|z_i - f_i(x)\|^2 \tag{54}
\]

where \( X \triangleq \{\theta_n \in [\theta_{\text{max}}, \theta_{\text{min}}], \forall \theta_n \in [0, V_{\text{max}}], \forall n \in N\} \) with \( \theta_{\text{max}} \) and \( V_{\text{max}} \) being the phase angle and voltage limit. By letting \( g_k(x) \triangleq z_i - f_i(x) \) and \( G_{k}(x) \triangleq \partial f_i(x)/\partial x^T \), 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

\[
\hat{x}[t] = \min_{x \in \mathbb{R}} \sum_{i=1}^{I} \|z_i[t] - f_i(x)\|^2 \tag{55}
\]

where \( z_i[t] \in \mathbb{R}^{M_i} \) 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 \( x^0[t] \) with the previous local estimate \( \hat{x}_i[t] \). 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 \( \{\hat{V}_{i,n}, \hat{\theta}_{i,n}\} \) at each update in each area from \( v_i^k = [\ldots, \hat{\theta}_{i,n}, \hat{V}_{i,n}, \ldots]^T \), the local MSE relative to \( V_{n}'s \) and \( \hat{\theta}_{i,n}'s \) is

\[
\text{MSE}_{V_i}^{(k)} = \mathbb{E}\left\{ \sum_{n=1}^{N} (\hat{V}_{i,n} - V_{n})^2 \right\} / N \tag{56}
\]

\[
\text{MSE}_{\theta_{i,n}}^{(k)} = \mathbb{E}\left\{ \sum_{n=1}^{N} (\theta_{i,n} - \hat{\theta}_{i,n})^2 \right\} / N. \tag{57}
\]

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 \( \bar{x} \) and \( f(\bar{x}) \) in per unit (p.u.) values.
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 \( \{z_i\}_{i=1}^I \) are generated adding independent Gaussian errors \( \varepsilon_{i,m} \sim N(0, \sigma^2) \) 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 \( A = I_3 I_3^T - I \). The weight matrix is constructed according to the Laplacian \( L = \text{diag}(A1_I) - A \) as \( W = I_f - wL \) with \( w = \beta / \max(A1_I) \) and \( \beta = 0.3 \). The step-size is \( \alpha_{\text{GNN}} = 0.5 \) 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 \( \ell \), while the GGN algorithm runs \( \ell_k = \ell_0 = \ell_{\text{min}} = 3 \) exchanges for each update. Therefore, the equivalent exchange index \( \ell \) in the figure for the diffusion algorithm is \( \ell = \sum_{k=1}^K \ell_k \) after 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

\[
\text{Val}_k = \sum_{i=1}^I \| z_i - f_i(x_k^\ell) \|^2
\]

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

\[
\text{Grad}_k = \sum_{i=1}^I \| G_i^T(x_k^\ell)(z_i - f_i(x_k^\ell)) \|
\]

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., \( K_{\text{min}} = 30 \) exchanges). It is observed in Fig. 3(a) to 3(b) that although the gossip exchange per update \( \ell_{\text{min}} = 3 \) is small and does not satisfy Assumption 4, both \( \text{Val}_k \) and \( \text{Grad}_k \) 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 \( \ell^{-1} \), 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

\[
\text{MSE}_{\ell}^{(k)} = \frac{1}{I} \sum_{i=1}^I \text{MSE}_{\ell,i}^{(k)} \quad \text{MSE}_{\ell_i}^{(k)} = \frac{1}{I} \sum_{i=1}^I \text{MSE}_{\ell_i}^{(k)}
\]

We generate 3 snapshots of measurements \( \{z_i(t)\}_{i=1}^I \) for \( t = 1, \cdots, 3 \) based on the same state \( x(t) = x \) by adding independent Gaussian noise with variance \( \sigma^2 = 10^{-6} \), similar to the adaptive setting considered in [24]. More specifically, we use \( \ell_{\text{min}} = 3 \) 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.
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 $O(N)$, 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 $\text{Val}_{x}(k) = ||x_i - f(x_i^*)||^2$ as well as the individual MSE$_{x_i}(k)$ and MSE$_{\text{avg}}(k)$ 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 consistency 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 $d_i^{k+1}$ in (14)

$$x_i^{k+1} = P_X [x_i^k - \alpha d_i^k + \alpha (d_i^k - d_i^k (f_k))] .$$

By subtracting the fixed point $\bar{x}$ and using the non-expansivity of the operator $P_X (\cdot)$ on the closed convex set $X$, we have the following recursion

$$\|x_i^{k+1} - \bar{x}\| \leq \|x_i^k - \bar{x} - \alpha d_i^k\| + \alpha \|d_i^k (f_k) - d_i^k\| .$$

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

$$x_i^k - \bar{x} - \alpha d_i^k = x_i^k - \bar{x} - \alpha G^i (x_i^k) g (x_i^k) + \alpha G^i (\bar{x}) g (\bar{x}) .$$

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

$$x_i^k - \bar{x} = G^i (x_i^k) G (x_i^k) (x_i^k - \bar{x}) .$$

Then by substituting (63) into (62), and meanwhile adding and subtracting simultaneously a term $\alpha G^i (x_i^k) g (x_i^k)$, we have the following expression

$$x_i^k - \bar{x} - \alpha d_i^k = G^i (x_i^k) [G (x_i^k) (x_i^k - \bar{x}) - \alpha g (x_i^k) + \alpha g (\bar{x})]$$

$$+ \alpha [G^i (\bar{x}) - G^i (x_i^k)] g (\bar{x}) .$$

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

$$\alpha g (\bar{x}) - \alpha g (x) - G (x) (\bar{x} - x)$$

$$= \alpha \int_0^1 G (x + t (\bar{x} - x)) (\bar{x} - x) dt - G (x) (\bar{x} - x)$$

$$= \alpha \left( \int_0^1 [G (x + t (\bar{x} - x)) - G (x)] (\bar{x} - x) dt \right)$$

$$- (1 - \alpha) G (x) (\bar{x} - x) ,$$

whose norm can be bounded by using Assumption 1 as

$$\|\alpha g (\bar{x}) - \alpha g (x) - G (x) (\bar{x} - x)\|$$

$$\leq \alpha \left( \int_0^1 \|G (x + t (\bar{x} - x)) - G (x)\| dt \right) \|x - \bar{x}\|$$

$$+ (1 - \alpha) \sigma_{\text{max}} \|x - \bar{x}\| .$$

From the Lipschitz condition in Assumption 1, we have

$$\int_0^1 \|G (x + t (\bar{x} - x)) - G (x)\| dt \leq \omega \|x - \bar{x}\| \int_0^1 t dt .$$

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

$$\|\alpha g (\bar{x}) - \alpha g (x) - G (x) (\bar{x} - x)\|$$

$$\leq \frac{\alpha \omega}{2} \|x - \bar{x}\|^2 + (1 - \alpha) \sigma_{\text{max}} \|x - \bar{x}\| ,$$

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

$$\|G (x - G (\bar{x})\| \leq \sqrt{2} \|G (x)\| \|G (\bar{x})\| \|G (x) - G (\bar{x})\|$$

$$\leq \sqrt{2} \omega \|x - \bar{x}\| .$$

By definition we have $\|G^i (x)\|^2 = \|G (x)\|^{-1}$. Also Assumption 1 implies $\|G^i (x)\| \leq 1/\sigma_{\text{min}}$. For convenience, we let $\epsilon_{\text{min}} \triangleq \|g (\bar{x})\|$ be the goodness of fit at $\bar{x}$ and define the following constants

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

Then, substituting $\|G (x)\| \leq 1/\epsilon_{\text{min}}$ (69) and (70) back into (64) and using (30), we have

$$\|x_i^k - \bar{x} - \alpha d_i^k\| \leq T_1 \|x_i^k - \bar{x}\|^2 + T_2 \|x_i^k - \bar{x}\| ,$$

Therefore, we have the error recursion (29).

APPENDIX B

PROOF OF THEOREM 1

If the discrepancy error is upper bounded by a constant $\kappa \geq 0$ such that $\|d_i^k (f_k) - d_i^k\| \leq \kappa$, then from Lemma 2, the recursion can be simplified as

$$\|x_i^{k+1} - x^*\| \leq T_1 \|x_i^k - x^*\|^2 + T_2 \|x_i^k - x^*\| + \alpha \kappa .$$

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

$$\zeta_{i,k+1} \leq T_1 \zeta_i^2 + T_2 \zeta_i + \alpha \kappa , \quad \zeta_{i,k} > 0 .$$
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
\[
\psi(\rho_k) = T_1 \rho_k^2 + T_2 \rho_k + \alpha \kappa, \quad \rho_k > 0,
\]
whose equilibrium points are obtained by solving
\[
\rho = T_1 \rho^2 + T_2 \rho + \alpha \kappa.
\] (75)
When \( \kappa \) satisfies
\[
(1 - T_2)^2 - 4\alpha T_1 \kappa \geq 0,
\] (76)
the equilibrium points of (75) exist and are obtained as (34).

Now let \( \hat{\psi}(\rho) = d\psi(\rho)/d\rho \) be the first order derivative of the dynamics. According to \( [57, \text{cf. Proposition 1.9}] \), an equilibrium point is a stable sink if \( |\hat{\psi}(\cdot)| < 1 \) and unstable otherwise. Thus, the equilibrium point is unstable since the following is always true
\[
\hat{\psi}(\rho_{\text{max}}) = |2T_1 \rho_{\text{max}}^2 + T_2| = |1 + \sqrt{(1 - T_2)^2 - 4\alpha T_1 \kappa}| > 1,
\] (77)
while the point \( \rho_{\text{min}} \) is a sink if
\[
\hat{\psi}(\rho_{\text{min}}) = |1 - \sqrt{(1 - T_2)^2 - 4\alpha T_1 \kappa}| < 1.
\] (78)
To guarantee \( |\hat{\psi}(\rho_{\text{min}})| < 1 \), it requires
\[
0 < (1 - T_2)^2 - 4\alpha T_1 \kappa < 4,
\] (80)
which together with (76) leads to the following condition on the bounded perturbation \( \kappa \)
\[
\frac{T_2^2 - 2T_2 - 3}{4\alpha T_1} < \kappa < \frac{T_2^2 - 2T_2 + 1}{4\alpha T_1}.
\] (81)
Clearly, given an arbitrary \( \alpha \in (0, 1) \), the lower bound on \( \kappa \) 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
\[
T_2^2 - 2T_2 - 3 < 0 \implies -1 < T_2 < 3.
\] (82)
Since \( T_2 \geq 0 \) by definition (30), the condition becomes
\[
0 \leq (1 - \alpha) \frac{\sigma_{\text{max}}^2}{\sigma_{\text{min}}^2} + \frac{\sqrt{2} \alpha \omega_{\min}}{\sigma_{\text{min}}^2} < 3.
\] (83)
By re-arranging the terms, this condition is equivalent to
\[
\left. \begin{array}{c}
\frac{\sqrt{2} \alpha \omega_{\min}}{\sigma_{\text{min}}^2} < 3 - (1 - \alpha) \frac{\sigma_{\text{max}}^2}{\sigma_{\text{min}}^2} \\
3 - (1 - \alpha) \frac{\sigma_{\text{max}}^2}{\sigma_{\text{min}}^2} > 0
\end{array} \right\} \quad \omega_{\min} \leq \frac{\sigma_{\text{min}}^2}{\sqrt{2} \alpha}, \quad \frac{3 \sigma_{\text{min}}^2}{\sigma_{\text{max}}^2} > 0.
\] (84)
which can be simplified as
\[
\left. \begin{array}{c}
\omega_{\min} \leq \frac{\sigma_{\text{min}}^2}{\sqrt{2} \alpha} \\
\max \left\{ 1 - \frac{3 \sigma_{\text{min}}^2}{\sigma_{\text{max}}^2}, 0 \right\} < \alpha \leq 1
\end{array} \right\}.
\] (85)
Thus, if the initial error \( \zeta_{i,0} > \rho_{\text{max}} \), the error keeps growing. On the other hand, if the errors are bounded by \( 0 < \zeta_{i,k} < \rho_{\text{max}} \) for all \( i \)’s and \( k \)’s, the algorithm reaches the equilibrium error floor \( \rho_{\text{min}} \). Thus, as long as the initialization error \( \zeta_{i,0} \) satisfies \( 0 < \zeta_{i,0} < \rho_{\text{max}} \) for \( i = 1, \ldots, I \), the algorithm progresses with contracting error until reaching the error floor \( \rho_{\text{min}} \) due to the constant bounded perturbation \( \kappa \).

As a result, as long as the initial condition \( x_0^i \) satisfies \( \|x_0^i - x^*\| < \rho_{\text{max}} \) with respect to a certain fixed point \( x^* \), the error norm is upper bounded by
\[
\lim_{k \to \infty} \|x^i_k - x^*\| \leq \rho_{\text{min}}.
\]
and hence the norm of each component is bounded by the total norm $\|e_k(\ell)\| < C\lambda_\eta^\ell$ and $\|E_k(\ell)\|_F < C\lambda_\eta^\ell$.

**APPENDIX D**

**PROOF OF LEMMA 4**

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

$$(Z + \delta Z)^{-1} = \sum_{q=0}^\infty (-1)^q (Z^{-1}\delta Z)^q Z^{-1}$$  \hspace{1cm} (89)

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

**A. Initial Case: $k = 1$**

Given $x_0^i = x^0$ for all $i$, then for any $i \neq j$ we have

$$\|x^1_i - x^1_j\| \leq \|d^0_j(\ell_0) - d^0_j(\ell_0)\|,$$  \hspace{1cm} (90)

where the discrepancy is expressed explicitly as

$$d^0_j(\ell_0) - d^0_j(\ell_0) = [H_0 + E_{0,j}(\ell_0)]^{-1} [\tilde{H}_0 + e_{0,j}(\ell_0)] - [H_0 + E_{0,j}(\ell_0)]^{-1} [\tilde{H}_0 + e_{0,j}(\ell_0)].$$

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

1) **Matrix series expansion**:

Since $x_0^i = x^0$ for all $i$ such that $H_0 = Q(x_0^0)$ and $I_0 = q(x_0^0)$, they can be bounded based on Assumption 1 as follows

$$\|\tilde{H}_0\| = \|Q(x_0^0)\| = \frac{1}{J} \frac{1}{J} \frac{1}{J} \|G^T(x_0^0)G(x_0^0)\| \leq \sigma_{\text{max}} \xi_{\text{max}} I,$$

$$\|H_0^{-1}\| = \|Q^{-1}(x_0^0)\| = I \left\|G^T(x_0^0)G(x_0^0)^{-1}\right\| \leq \frac{I}{\sigma_{\text{min}}}.$$  \hspace{1cm} (91)

Note that from the norm inequality of sub-matrices

$$\|H_0^{-1}E_{0,i}(\ell_0)\| \leq \|H_0^{-1}\| \|E_{0,i}(\ell_0)\| \leq \|H_0^{-1}\| \|E_0(\ell_0)\|_F,$$

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

$$\|H_0^{-1}\| \|E_0(\ell_0)\|_F \leq \frac{1}{\sigma_{\text{min}}} C^\ell_0 \lambda_\eta(\ell_0 - \ell_{\text{min}}) \frac{IC}{\sigma_{\text{min}}} \lambda_\eta^\ell_{\text{min}}.$$  \hspace{1cm} (92)

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

$$\ell_{\text{min}} > \log \left( \frac{\xi}{4D} \right) / \log \lambda_\eta \Rightarrow \frac{IC}{\sigma_{\text{min}}} \lambda_\eta^\ell_{\text{min}} < \frac{IC}{4\sigma_{\text{min}}^2 D} \xi.$$  \hspace{1cm} (93)

For notation convenience, we define

$$\hat{\xi} = \frac{IC}{\sigma_{\text{min}}^2 D} \xi,$$

and clearly, we have $0 < \hat{\xi} < \xi < 1/2$ according to the definition of $D$ in (45) by Assumption 4. Therefore, letting $\delta Z = E_0(\ell_0)$ and $Z = H_0$, it follows from Lemma 3 that

$$\|Z^{-1}\delta Z\| = \|H_0^{-1}\| \|E_0(\ell_0)\|_F \leq \frac{1}{\sigma_{\text{min}}} \lambda_\eta(\ell_0 - \ell_{\text{min}}) \hat{\xi} < \frac{1}{8}.$$  \hspace{1cm} (94)

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

$$d^0_j(\ell_0) - d^0_j(\ell_0) = \left[ H_0^{-1} - \sum_{q=0}^\infty (-1)^q (H_0^{-1}E_{0,i}(\ell_0))^{q+1} H_0^{-1} \right] [H_0 + e_{0,i}(\ell_0)]$$

$$\left[ H_0^{-1} - \sum_{q=1}^\infty (-1)^q (H_0^{-1}E_{0,j}(\ell_0))^{q+1} H_0^{-1} \right] [H_0 + e_{0,j}(\ell_0)].$$

To simplify the above expression, we write it in three terms $D_1(\ell_0), D_2(\ell_0)$ and $D_3(\ell_0)$ as follows

$$d^0_j(\ell_0) - d^0_j(\ell_0) = D_1(\ell_0) + D_2(\ell_0) + D_3(\ell_0),$$

where $D_1(\ell_0) \equiv H_0^{-1} [e_{0,i}(\ell_0) - e_{0,j}(\ell_0)]$ and

$$D_2(\ell_0) \equiv \sum_{q=1}^\infty (-1)^q (H_0^{-1}E_{0,i}(\ell_0))^{q+1} H_0^{-1} [\tilde{H}_0 + e_{0,j}(\ell_0)] \left[ H_0^{-1} - \sum_{q=1}^\infty (-1)^q (H_0^{-1}E_{0,j}(\ell_0))^{q+1} H_0^{-1} \right] [H_0 + e_{0,j}(\ell_0)] \right].$$

2) **Proof of success when $k = 1$**:

According to the triangular inequality for norms, we can bound

$$\|e_{0,i}(\ell_0) - e_{0,j}(\ell_0)\| \leq 2 \|e_{0,i}(\ell_0)\|$$

$$\|e_{0,i}(\ell_0) - e_{0,j}(\ell_0)\| \leq 2 \|e_{0,i}(\ell_0)\|_F.$$

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

$$\|D_1(\ell_0)\| \leq 2 \|\|H_0^{-1}\|\|e_{0,i}(\ell_0)\| \leq \frac{1}{2} \hat{\xi} \lambda_\eta^\ell(\ell_0 - \ell_{\text{min}}).$$  \hspace{1cm} (96)

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

$$\sum_{q=1}^\infty (-1)^q \left[ (H_0^{-1}E_{0,i}(\ell_0))^{q+1} - (H_0^{-1}E_{0,j}(\ell_0))^{q+1} \right]$$

$$\leq 2 \sum_{q=1}^\infty \left\| (H_0^{-1}E_{0,i}(\ell_0))^{q+1} \right\|^q$$

$$\leq 2 \sum_{q=1}^\infty \left\{ \frac{1}{2} \hat{\xi} \lambda_\eta^\ell(\ell_0 - \ell_{\text{min}}) \right\}^q \leq \frac{1}{2} \frac{\hat{\xi} \lambda_\eta^\ell(\ell_0 - \ell_{\text{min}})}{1 - \frac{1}{2} \hat{\xi} \lambda_\eta^\ell(\ell_0 - \ell_{\text{min}})},$$  \hspace{1cm} (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 $0 < \xi, \xi < 1/2$ and $\lambda_\eta(\ell_0 - \ell_{\text{min}}) \leq 1$, then

$$\frac{\hat{\xi} \lambda_\eta^\ell(\ell_0 - \ell_{\text{min}})}{1 - \frac{1}{2} \hat{\xi} \lambda_\eta^\ell(\ell_0 - \ell_{\text{min}})} < 2\hat{\xi} \lambda_\eta^\ell(\ell_0 - \ell_{\text{min}}),$$  \hspace{1cm} (98)

and thus the norm of the second term is bounded as

$$\|D_2(\ell_0)\| \leq \hat{\xi} \lambda_\eta^\ell(\ell_0 - \ell_{\text{min}}) \|H_0^{-1}\| \|e_{0,i}(\ell_0)\| \leq \frac{\sigma_{\text{max}} \xi_{\text{max}}}{\sigma_{\text{min}}^2} \hat{\xi} \lambda_\eta^\ell(\ell_0 - \ell_{\text{min}}),$$
where the last inequality comes from (91).

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

$$
\|D_3(\ell_0)\| \leq 2 \sum_{q=1}^{\infty} \left( \|H_0^{-1}\| \|E_0(\ell_0)\|_F \right)^q \|H_0^{-1}e_0(\ell_0)\|
$$

which leads to

$$
\|D_3(\ell_0)\| \leq 2 \sum_{q=1}^{\infty} \left( \frac{1}{2} \xi \lambda_q(\ell_0 - \ell_{\text{min}}) \right)^q \left( 1 - \frac{1}{2} \xi \lambda_q(\ell_0 - \ell_{\text{min}}) \right) \cdot \frac{1}{8} \xi \lambda_q(\ell_0 - \ell_{\text{min}})
$$

(101)

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

$$
\|d^K_0(\ell_0) - d^K_0(\ell_0)\| \leq \left(1 + \frac{\sigma_{\text{max}} \xi}{\sigma_{\text{min}}^2 D} \right) \frac{IC}{\sigma_{\text{min}}^2 D} \xi \lambda_q(\ell_0 - \ell_{\text{min}})
$$

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

$$
\|x_i^K - x_j^K\| \leq \xi \left( \frac{CC_1 C_2}{D} \right) \lambda_q(\ell_0 - \ell_{\text{min}}),
$$

(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 $i \neq j$

$$
\|x_i^K - x_j^K\| \leq \xi \left( \frac{CC_1 C_2}{D} \right) \sum_{k=0}^{K} \lambda_q(\ell_k - \ell_{\text{min}}).
$$

(103)

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

$$
\|x_i^{K+1} - x_j^{K+1}\| \leq \|x_i^K - x_j^K\| + \|d^K(\ell_K) - d^K(\ell_K)\|,
$$

where

$$
d^K(\ell_K) - d^K(\ell_K) = [H_K + E_{K,i}(\ell_K)]^{-1} [H_K + e_{K,i}(\ell_K)]
- [H_K + E_{K,j}(\ell_K)]^{-1} [H_K + e_{K,j}(\ell_K)].
$$

(104)

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

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

$$
\|H_K^{-1}\| = \left\| \left[ (Q(x^K) + S_{K,i})^{-1} \right] \right\|,
$$

(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

$$
\|S_{K,i}\| < \nu_{\xi} \xi \left( \frac{CC_1 C_2}{D} \right) \sum_{k=0}^{K} \lambda_q(\ell_k - \ell_{\text{min}})
$$

(106)

$$
< \xi \left( \frac{\nu_{\xi} CC_1 C_2}{D} \right) \lambda_{\infty},
$$

(107)

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

$$
\|Q^{-1}(x^K)S_{K,i}\| \leq \|Q^{-1}(x^K)\| \|S_{K,i}\| \leq \|Q^{-1}(x^K)\| \|S_{K,i}\| \leq \frac{I}{\sigma_{\text{min}}^2} \left( \frac{\nu_{\xi} CC_1 C_2}{D} \right) \lambda_{\infty} \xi < \frac{1}{2},
$$

(108)

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

$$
\|H_K^{-1}\| \leq \|Q^{-1}(x^K)\|
$$

(109)

$$
+ \sum_{q=1}^{\infty} \left( \|Q^{-1}(x^K)\| \|S_{K,i}\| \right)^q \|Q^{-1}(x^K)\|
$$

$$
\leq \frac{I}{\sigma_{\text{min}}^2} + \frac{I}{\sigma_{\text{min}}^2} \sum_{q=1}^{\infty} \xi^q
$$

$$
= \frac{I}{\sigma_{\text{min}}^2} \left(1 + \frac{\xi}{1 - \xi}\right) = \frac{I}{\sigma_{\text{min}}^2} \frac{1}{1 - \xi} < \frac{2I}{\sigma_{\text{min}}^2}.
$$

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

$$
\|H_K^{-1}E_{K,i}(\ell_K)\| \leq \|H_K^{-1}\| \|E_{K,i}(\ell_K)\| \leq \|H_K^{-1}\| \|E_{K}(\ell_K)\|_F
$$

$$
\|H_K^{-1}E_{K,j}(\ell_K)\| \leq \|H_K^{-1}\| \|E_{K,j}(\ell_K)\| \leq \|H_K^{-1}\| \|E_{K}(\ell_K)\|_F.
$$

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

$$
\|H_K^{-1}\| \|E_{K}(\ell_K)\|_F \leq \frac{2I}{\sigma_{\text{min}}^2} C \lambda_{\ell_{\text{min}}} = \frac{2I}{\sigma_{\text{min}}^2} \frac{IC}{\sigma_{\text{min}}^2 D} \xi.
$$

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

$$
\|Z^{-1}\delta Z\| = \|H_K^{-1}\| \|E_{K}(\ell_K)\|_F
$$

(110)

$$
\leq \frac{2I}{\sigma_{\text{min}}^2} C \lambda_{\ell_{\text{min}}} \tilde{\xi} < \frac{1}{4}.
$$

(111)

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

$$
d^K(\ell_K) - d^K(\ell_K) = \left[ H_K^{-1} - \sum_{q=1}^{\infty} (-1)^q (H_K^{-1}E_{K,i}(\ell_K))^q H_K^{-1} \right] [H_K + e_{K,i}(\ell_K)]
- \left[ H_K^{-1} - \sum_{q=1}^{\infty} (-1)^q (H_K^{-1}E_{K,j}(\ell_K))^q H_K^{-1} \right] [H_K + e_{K,j}(\ell_K)].
$$

(104)
To simplify the above expression, we write it in three terms $D_1(\ell K), D_2(\ell K)$ and $D_3(\ell K)$ as follows
\[
d_k^K(\ell K) - d_j^K(\ell K) = D_1(\ell K) + D_2(\ell K) + D_3(\ell K),
\]
where $D_1(\ell K) \triangleq \bar{H}_K^{-1} e_{\ell K,i}(\ell K) - e_{\ell K,j}(\ell K)$ and
\[
D_2(\ell K) \triangleq \sum_{q=1}^{\infty} (-1)^q \left( \bar{H}_K^{-1} E_{K,j}(\ell K) \right)^q \bar{H}_K^{-1} \bar{h}_K
- \sum_{q=1}^{\infty} (-1)^q \left( \bar{H}_K^{-1} E_{K,i}(\ell K) \right)^q \bar{H}_K^{-1} \bar{h}_K
\]
\[
D_3(\ell K) \triangleq \sum_{q=1}^{\infty} (-1)^q \left( \bar{H}_K^{-1} E_{K,i}(\ell K) \right)^q \bar{H}_K^{-1} e_{\ell K,j}(\ell K)
- \sum_{q=1}^{\infty} (-1)^q \left( \bar{H}_K^{-1} E_{K,j}(\ell K) \right)^q \bar{H}_K^{-1} e_{\ell K,i}(\ell K).
\]

2) Proof of success when $k = K + 1$: According to the triangular inequality for norms, we can bound
\[
\| e_{\ell K,i}(\ell K) - e_{\ell K,j}(\ell K) \| \leq 2 \| e_{\ell K}(\ell K) \|
\]
\[
\| E_{K,i}(\ell K) - E_{K,j}(\ell K) \| \leq 2 \| E_{\ell K}(\ell K) \|_F.
\]
Using (110), we can bound the norm of the first term as
\[
\| D_1(\ell K) \| \leq 2 \| H_K^{-1} \| \| e_{\ell K}(\ell K) \| \leq \xi \lambda_{\eta}^{\ell K - \ell_{\min}}.
\]
Similarly, the infinite sum in the second term is bounded as
\[
\sum_{q=1}^{\infty} (-1)^q \left[ \left( H_K^{-1} E_{K,j}(\ell K) \right)^q - \left( H_K^{-1} E_{K,i}(\ell K) \right)^q \right] \leq 2 \sum_{q=1}^{\infty} \left( \| H_K^{-1} \| \| E_{\ell K}(\ell K) \|_F \right)^q
\]
\[
\leq 2 \sum_{q=1}^{\infty} \left( \frac{\xi \lambda_{\eta}^{\ell K - \ell_{\min}}}{1 - \frac{1}{2} \xi \lambda_{\eta}^{\ell K - \ell_{\min}}} \right)^q = \frac{\xi \lambda_{\eta}^{\ell K - \ell_{\min}}}{\left( 1 - \frac{1}{2} \xi \lambda_{\eta}^{\ell K - \ell_{\min}} \right)^q},
\]
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 $0 < \xi < \xi < 1/2$ and $\lambda_{\eta}^{\ell K - \ell_{\min}} < 1$, then
\[
\frac{\xi \lambda_{\eta}^{\ell K - \ell_{\min}}}{\left( 1 - \frac{1}{2} \xi \lambda_{\eta}^{\ell K - \ell_{\min}} \right)^q} < 2 \xi \lambda_{\eta}^{\ell K - \ell_{\min}}\).
\]
and thus the norm of the second term is bounded as
\[
\| D_2(\ell K) \| \leq 2 \xi \lambda_{\eta}^{\ell K - \ell_{\min}} \| H_K^{-1} \| \| \bar{h}_K \|
\]
\[
\leq 2 \sigma_{\text{max}} \lambda_{\eta}^{\ell K - \ell_{\min}},
\]
where the last inequality comes from (91). Following the same rationale, the norm of the third term can be bounded as
\[
\| D_3(\ell K) \| \leq 2 \sum_{q=1}^{\infty} \left( \| H_K^{-1} \| \| E_{\ell K}(\ell K) \|_F \right)^q \| H_K^{-1} \|
\]
\[
\leq 2 \sum_{q=1}^{\infty} \left( \| H_K^{-1} \| \| E_{\ell K}(\ell K) \|_F \right)^q + 1
\]
\[
\leq 2 \sum_{q=1}^{\infty} \left( \frac{1}{\xi \lambda_{\eta}^{\ell K - \ell_{\min}}} \right)^q + 1
\]
\[
= \frac{\xi \lambda_{\eta}^{\ell K - \ell_{\min}}}{1 - \frac{1}{2} \xi \lambda_{\eta}^{\ell K - \ell_{\min}}} < \xi \lambda_{\eta}^{\ell K - \ell_{\min}}.
\]
where the last inequality has used the results in (115). Note that this is again a very loose bound.

Substituting $\xi = IC \xi / (\sigma_{\text{min}}^2 D)$ in (93) back to (113), (114) and (118) and using the constants $C_1$ and $C_2$, we have
\[
\| d_j^K(\ell K) - d_j^K(\ell K) \| \leq \xi \left( \frac{CC_1 C_2}{D} \right) \lambda_{\eta}^{\ell K - \ell_{\min}}
\]
Similarly, based on (47) and (90), we have
\[
\| x_{j}^{K+1} - x_{j}^{K+1} \|
\leq \| x_{j}^{K} - x_{j}^{K} \| + \| d_j^K(\ell K) - d_j^K(\ell K) \|
\leq \xi \left( \frac{CC_1 C_2}{D} \right) \sum_{k=0}^{K-1} \lambda_{\eta}^{\ell K - \ell_{\min}} + \xi \left( \frac{CC_1 C_2}{D} \right) \lambda_{\eta}^{\ell K - \ell_{\min}}
\leq \xi \left( \frac{CC_1 C_2}{D} \right) \sum_{k=0}^{K} \lambda_{\eta}^{\ell K - \ell_{\min}},
\]
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 $\xi \leq 4D\lambda_{\eta}^{\ell_{\min} + 1}$, and
\[
\| x_{j}^{K+1} - x_{j}^{K+1} \| \leq 4CC_1 C_2 \sum_{k=0}^{K} \lambda_{\eta}^{k+1}.
\]

**APPENDIX E**

**Proof of Theorem 2**

By the decomposition in (41), we have
\[
d_j^k(\ell_k) - d_j^k
\]
\[
= \left[ Q(x_k^j) + S_{k,i} + E_{k,i}(\ell_k) \right] - \left[ Q(x_k^j) - Q(x_k^j)^{-1} q(x_k^j) \right]
- Q(x_k^j)^{-1} q(x_k^j).
\]
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 $\| S_{k,i}(\ell_k) \| \leq \nu_s C_1 \lambda_{\infty} \xi / D$.
The expansion depends on the quantity
\[
\| Q^{-1}(x_k^j) (S_{k,i} + E_{k,i}(\ell_k)) \|
\]
\[
\leq \| Q^{-1}(x_k^j) \| \| S_{k,i} \| + \| Q^{-1}(x_k^j) \| \| E_{k,i}(\ell_k) \|.
\]
Using the derivation in (108) and $C_1, C_2$ in (47), we have
\[
\|Q^{-1}(x^k_i)\left(S_{k,i} + E_{k,i}(\ell_k)\right)\| \\
< C_2 \left(\frac{\nu SC_1 C_2}{D}\right) \lambda_\infty \xi + \left(\frac{C_2}{4D}\right) \xi \lambda_{\eta}(\ell_k - \ell_0) \\
= CC_2 \left(\frac{\nu SC_1 C_2}{D}\right) \lambda_\infty + \left(\frac{1}{4}\right) \xi \lambda_{\eta}(\ell_k - \ell_0) \\
< CC_2 \left(\frac{\nu SC_1 C_2}{D}\right) \lambda_\infty + \left(\frac{2}{4}\right) \xi \lambda_{\eta}(\ell_k - \ell_0) \\
=1, \text{ from (45)}
\]

where the last inequality is by the definition of $D$ in (45). Then (119) can be re-written as
\[
d^k_i(\ell_k) - d^k_i \\
= \left[Q^{-1}(x^k_i) - \sum_{q=1}^{\infty} \left(\sum_{q=1}^{\infty} \left(Q^{-1}(x^k_i)(S_{k,i} + E_{k,i}(\ell_k))\right)^q Q^{-1}(x^k_i)\right)\right] q(x^k_i) \\
= Q^{-1}(x^k_i) \left[s_{k,i} + e_{k,i}(\ell_k)\right] - Q^{-1}(x^k_i) q(x^k_i) \\
= \left[Q^{-1}(x^k_i) \left[s_{k,i} + e_{k,i}(\ell_k)\right]\right] - Q^{-1}(x^k_i) q(x^k_i) \\
= \sum_{q=1}^{\infty} (Q^{-1}(x^k_i)(S_{k,i} + E_{k,i}(\ell_k)))^q Q^{-1}(x^k_i)q(x^k_i) \\
= \sum_{q=1}^{\infty} (Q^{-1}(x^k_i)(S_{k,i} + E_{k,i}(\ell_k)))^q Q^{-1}(x^k_i)s_{k,i} + e_{k,i}(\ell_k). \quad (120)
\]

According to Lemma 4 and Assumption 4, we have $\|s_{k,i}(\ell_k)\| \leq \nu \frac{CC_1 C_2 \lambda_\infty \xi}{D}$, and the norm of the first term above can be bounded similarly as (120)
\[
\|Q^{-1}(x^k_i) \left[s_{k,i} + e_{k,i}(\ell_k)\right]\| < \xi. \quad (121)
\]

Likewise, the norm of the second term is bounded as
\[
\sum_{q=1}^{\infty} \left(\sum_{q=1}^{\infty} \left(Q^{-1}(x^k_i)(S_{k,i} + E_{k,i}(\ell_k))\right)^q Q^{-1}(x^k_i)q(x^k_i)\right) \\
= \sum_{q=1}^{\infty} \left(\|Q^{-1}(x^k_i)(S_{k,i} + E_{k,i}(\ell_k))\|\right)^q \|Q^{-1}(x^k_i)q(x^k_i)\| \\
< \frac{\sigma_{\max} \xi \max}{\sigma_{\min}^2} \sum_{q=1}^{\infty} \xi^q = \frac{\sigma_{\max} \xi \max}{\sigma_{\min}^2} \frac{\xi}{1 - \xi} < 2 \frac{\sigma_{\max} \xi \max}{\sigma_{\min}^2} \xi, \quad (122)
\]

and similarly for the third term
\[
\sum_{q=1}^{\infty} \left(\|Q^{-1}(x^k_i)(S_{k,i} + E_{k,i}(\ell_k))\|\right)^{q+1} < \sum_{q=1}^{\infty} \xi^{q+1} = \frac{\xi^2}{1 - \xi}. \quad (123)
\]

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
\[
\|d^k_i(\ell_k) - d^k_i\| \leq 2 \left(1 + \frac{\sigma_{\max} \xi \max}{\sigma_{\min}^2}\right) \xi = C_1 \xi \quad (122)
\]
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 $\xi$ specified by the system. Given (45), we have
\[
\xi < 4D\lambda_{\eta}(\ell_{\min} + 1), \quad (123)
\]
and therefore the perturbation bound $\kappa$ on the error recursion in Lemma 2 can be obtained as
\[
\kappa \leq 4C_1 D\lambda_{\eta}(\ell_{\min} + 1). \quad (124)
\]


