Communication-Efficient Distributed Learning Over Networks—Part I: Sufficient Conditions for Accuracy

Zhenyu Liu, *Member, IEEE*, Andrea Conti, *Fellow, IEEE*, Sanjoy K. Mitter, *Life Fellow, IEEE*, and Moe Z. Win, *Fellow, IEEE*

***Abstract*— Distributed learning is an important task in emerging applications such as localization and navigation, Internet-of-Things, and autonomous vehicles. This paper establishes a theoretical framework for learning states that evolve in real time over networks. Specifically, each agent node in the network aims to infer a time-varying state in a decentralized manner by using the node's local observations and the messages received from other nodes within its communication range. As a result, the inference accuracy of a node is significantly affected by the quality of its received messages. This calls for carefully designed strategies for generating messages that are able to provide sufficient information for the receiver and are robust to channel impairments. This paper presents communication-efficient encoding strategies for generating transmitted messages and derives a sufficient condition for the boundedness of the distributed inference error of all the agent nodes over time. The findings of this paper provide guidelines for the design of communication-efficient distributed learning in complex networked systems.**

***Index Terms*— Distributed learning, decentralized network inference, noisy inference, anytime capacity, multi-agent networks.**

I. Introduction

DISTRIBUTED learning is critical for complex networked systems and enables various applications such as location based services [1], [2], [3], [4] and Internet-of-Things [5], [6], [7]. In distributed learning, the sensing, communication, and computing capabilities of different nodes in a network are exploited for learning unknown states with no or minimum involvement of a central processor [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19]. In many applications, the states to be learned are time-varying and central processors may not be available. For these applications, a node needs to learn the unknown states in real time by efficiently fusing the information contained in its sensor observations and that in the messages received from other nodes. Such a learning task is referred to as distributed inference in this paper. Example applications of distributed inference include localization and navigation [20], [21], [22], [23], [24], environmental monitoring [25], [26], [27], [28], as well as target detection and tracking [29], [30], [31].

Distributed inference is challenging due to the following reasons. First, the nodes in the network typically have constraints on their communication resources (e.g., power and bandwidth). Therefore, communication-efficient algorithms and protocols are required in order to reduce the communication overhead and to ensure desirable inference accuracy simultaneously. Second, inter-node communication is typically affected by channel impairments, leading to corruption and failures of communication. Consequently, channel coding techniques are required to protect the transmitted messages against channel impairments. Finally, distributed inference is a time-sensitive task as the states of interest vary with time. This calls for low-latency techniques for communication and computing in order to improve the nodes' capabilities of inferring their current states in real time.

Distributed learning and related optimization techniques have been studied extensively [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42]. An emerging technique that has attracted significant research interest is federated learning [13], [14], [15], [16], where nodes in the network aim to learn a statistical model by performing local computing based on data they generated and by communicating with a central processor iteratively. Consensus and diffusion techniques have been proposed for distributed inference over networks [35], [36], [37], [38], [39], where each node iteratively exchanges messages containing information of the unknown states with other nodes and updates its own estimator based on the received messages. In addition, many papers study real-time methods for generating encoded messages exchanged among nodes, where the encoder only uses currently available information without waiting to collect more data [43], [44], [45], [46], [47], [48], [49], [50], [51], [52], [53], [54]. An important notion for real-time encoding is the *anytime capacity* proposed in [55], which is used for establishing necessary and sufficient conditions under which a dynamic system can be stabilized over noisy channels.

Common limitations of known results on distributed learning and distributed inference are that they do not consider

Manuscript received 16 April 2022; revised 12 September 2022; accepted 30 November 2022. Date of publication 13 February 2023; date of current version 17 March 2023. This work was supported in part by the Office of Naval Research under Grant N00014-16-1-2141 and Grant N62909-22-1-2009 and in part by the Army Research Office through the Massachusetts Institute of Technology (MIT) Institute for Soldier Nanotechnologies under Contract W911NF-13-D-0001. *(Corresponding author: Moe Z. Win.)*

Zhenyu Liu is with the Wireless Information and Network Sciences Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139 USA.

Andrea Conti is with the Department of Engineering and CNIT, University of Ferrara, 44122 Ferrara, Italy.

Sanjoy K. Mitter and Moe Z. Win are with the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: moewin@mit.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JSAC.2023.3242731.

Digital Object Identifier 10.1109/JSAC.2023.3242731

constraints on the data rates of transmitted messages or do not account for channel impairments. In addition, some works rely on iterative mechanisms for exchanging messages and integrating information across the network, which can incur significant latency and communication overhead. As a result, communication-efficient distributed inference in complex networked systems remains a challenging problem.

Fundamental questions related to distributed inference include what are the requirements on sensing and communication capabilities of the network for achieving desirable accuracy and how to perform communication-efficient message generation and message exchange in the network? Answers to these questions will provide guidelines for the design of distributed learning algorithms in networked systems. The goal of this paper is to establish a theoretical foundation for distributed inference of time-varying states in complex networked systems. The key contributions of this paper are as follows:

- we present a general model for distributed inference of time-varying states in complex networked systems without any central processors;
- we design real-time methods for generating encoded messages exchanged among nodes and for inferring the unknown states in a distributed manner;
- we establish a sufficient condition on the network's sensing and communication capabilities under which the distributed inference error is bounded; and
- we evaluate the performance of the designed distributed encoding and inference methods for the application of network localization and navigation (NLN).

This paper focuses on sufficient conditions for the distributed inference error to be bounded. A companion paper [56] derives a necessary condition on the network's sensing and communication capabilities of the network under which the distributed inference error is bounded. The remaining sections are organized as follows: Section II presents the problem formulation. Section III describes preliminaries used in the paper. Section IV presents a few important notions for presenting the results of the paper. Section V presents the main result of the paper: a sufficient condition under which the distributed inference error is bounded over time. Section VI presents case studies of the established sufficient condition, together with results for the application of NLN. Finally, Section VII gives our conclusions.

Notations: Random variables are displayed in sans serif, upright fonts; their realizations in serif, italic fonts. Vectors and matrices are denoted by bold lowercase and uppercase letters, respectively. For example, a random variable and its realization are denoted by x and x; a random vector and its realization are denoted by $\mathbf{x}$ and $\boldsymbol{x}$, respectively. The expectation of $\mathbf{x}$ is denoted by $\mathbb{E}\{\mathbf{x}\}$, whereas the conditional expectation of $\mathbf{x}$ given a random vector $\mathbf{y}$ is denoted by $\mathbb{E}\{\mathbf{x}\,|\,\mathbf{y}\}$. Given a discrete-time stochastic process $\{\mathbf{x}_t\}_{t\geqslant 0}$, the notation $\mathbf{x}_{s:t}$ represents the vertical concatenation of $\mathbf{x}_s, \mathbf{x}_{s+1}, \ldots, \mathbf{x}_t$ for integers $0 \leqslant s \leqslant t$. The sets of real numbers and complex numbers are denoted by $\mathbb{R}$ and $\mathbb{C}$, respectively. Logarithms of a positive number x with base 2 is denoted by $\log x$. The cardinality of a set $\mathcal{X}$ is denoted by $|\mathcal{X}|$. The dimensionality of a linear subspace $\mathcal{S}$ is denoted by $\dim(\mathcal{S})$. The sum and direct sum of two subspaces $\mathcal{S}_1$ and $\mathcal{S}_2$ are denoted by $\mathcal{S}_1 + \mathcal{S}_2$ and $\mathcal{S}_1 \oplus \mathcal{S}_2$, respectively. The precedence of sum and direct sum is lower than that of the operator $\cap$ in all expressions. For example, $\mathcal{S}_1 + \mathcal{S}_2 \cap \mathcal{S}_3 = \mathcal{S}_1 + (\mathcal{S}_2 \cap \mathcal{S}_3)$ for subspaces $\mathcal{S}_1$, $\mathcal{S}_2$, and $\mathcal{S}_3$. The Euclidean norm and the ith entry of a vector $\boldsymbol{x}$ are denoted by $\|\boldsymbol{x}\|$ and $[\boldsymbol{x}]_i$, respectively. The transpose, column space, and spectral norm (i.e., the largest singular value) of matrix $\boldsymbol{A}$ are denoted by $\boldsymbol{A}^{\mathrm{T}}$, $\mathcal{C}(\boldsymbol{A})$, and $\|\boldsymbol{A}\|$, respectively. Notation $\mathrm{diag}\{\cdot\}$ represents a block diagonal matrix with the arguments being its diagonal blocks. For example, $\mathrm{diag}\{\boldsymbol{A}, \boldsymbol{B}\} := \left[\begin{smallmatrix} \boldsymbol{A} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B} \end{smallmatrix}\right]$. The horizontal concatenation of matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ (resp. row vectors $\boldsymbol{a}^{\mathrm{T}}$ and $\boldsymbol{b}^{\mathrm{T}}$) with the same number of rows is denoted by $[\boldsymbol{A} \quad \boldsymbol{B}]$ (resp. $[\boldsymbol{a}^{\mathrm{T}} \quad \boldsymbol{b}^{\mathrm{T}}]$). The Kronecker product of matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ is denoted by $\boldsymbol{A} \otimes \boldsymbol{B}$. The m-by-n matrix of zeros (resp. ones) is denoted by $\boldsymbol{0}_{m\times n}$ (resp. $\boldsymbol{1}_{m\times n}$); when $n = 1$, the m-dimensional vector of zeros (resp. ones) is simply denoted by $\boldsymbol{0}_m$ (resp. $\boldsymbol{1}_m$); the m-by-m identity matrix is denoted by $\boldsymbol{I}_m$: the subscript is omitted when the size of the matrix is clear from the context. Notations and definitions for important quantities used in the paper are summarized in Table I.

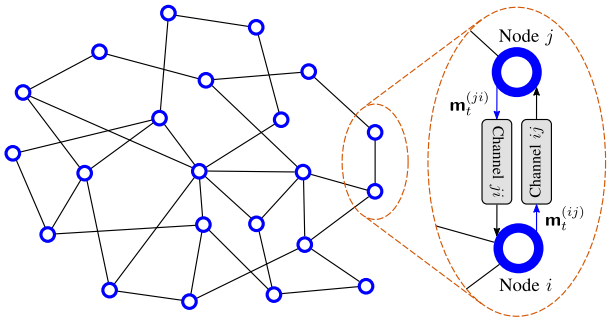


Fig. 1. Distributed inference over networks via sensing and communication.

II. Problem Formulation

Consider a set $\mathcal{V}$ of nodes where each node is associated with a time-varying unknown state. Each node in the network can perform observations and exchange messages with other nodes within its communication range. Based on whether a pair of nodes are within the communication range of each other, an undirected graph $\mathscr{G}_{\mathrm{u}} = \{\mathcal{V}, \mathcal{E}_{\mathrm{u}}\}$ with vertex set $\mathcal{V}$ and edge set $\mathcal{E}_{\mathrm{u}}$ is constructed, as shown in Fig. 1. Specifically, the vertex set $\mathcal{V}$ consists of all the nodes, and an edge $(i, j) \in \mathcal{E}_{\mathrm{u}}$ exists if and only if nodes i and j are within the communication range of each other. In this case, node j is a neighbor of node i and vice versa. The neighbor set of node j is denoted by $\mathcal{N}_{\mathrm{u}}^{(j)}$. This paper considers connected graph $\mathscr{G}_{\mathrm{u}}$, i.e., there is a path between any two different vertices in the graph.

Each node in the network performs noisy observations of its own state and the states of its neighbors. Moreover, a node generates encoded messages and transmits them to its neighbors. A block diagram containing the unknown state, observations, and encoded messages associated with a node j

TABLE I
NOTATIONS AND DEFINITIONS OF IMPORTANT QUANTITIES

Notation	Definition	Notation	Definition
$\mathscr{G}_{\mathrm{u}}$	undirected graph associated with the network	$\mathcal{V}$	set of all the nodes in the network
v	the number of nodes in the network	$\mathbf{x}_t^{(j)}$	state of node j at time t
$\boldsymbol{A}^{(j)}$	dynamic matrix of node j	$\mathbf{z}_t^{(jj)}$	intra-node observation obtained by node j at time t
$\mathbf{z}_t^{(ji)}$	inter-node observation obtained by node j with node i at time t	$\mathbf{m}_t^{(ji)}$	message transmitted from node j to node i at time t
$\mathbf{r}_t^{(ji)}$	message received by node i from node j at time t	$\mathcal{V}_{\mathrm{a}}$	subset containing all the agents in the network
$\hat{\mathbf{x}}_{\mathrm{opt},t}^{(j)}$	distributed minimum-mean-square-error estimator of $\mathbf{x}_t^{(j)}$	$\varepsilon_t^{(j)}$	individual distributed inference mean-square error of agent j at time t
F_t	learning objective function at time t, namely the total distributed inference mean-square error of all agents	$\mathcal{I}_{\boldsymbol{F}}(\mathcal{Y})$	minimum $\boldsymbol{F}$-invariant subspace over $\mathcal{Y}$
$\Lambda(\boldsymbol{F})$	set of all the eigenvalues of a real square matrix $\boldsymbol{F}$ that are either real or have positive imaginary parts	$\mathcal{M}_\lambda(\boldsymbol{F})$	real generalized eigenspace of a real square matrix $\boldsymbol{F}$ associated with eigenvalue λ
$\mathbf{x}_t$	concatenation of the unknown states of all the nodes at time t	$\boldsymbol{A}$	block-diagonal matrix with $\boldsymbol{A}^{(j)}$ being its jth diagonal block
$\boldsymbol{e}_{k,m}$	unit vector of m entries with its kth entry being 1 and other entries being 0	$\mathcal{X}^{(j)}$	node j's state subspace
$\mathcal{T}$	ordered tree based on graph $\mathscr{G}_{\mathrm{u}}$	$\mathcal{E}$	set of edges in ordered tree $\mathcal{T}$
$\mathcal{N}^{(j)}$	set of neighbors of node j in $\mathcal{T}$	$\mathcal{N}_{\mathrm{s}}^{(j)}$	subset of $\mathcal{N}^{(j)}$
$\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)$	subset consisting of node j and nodes still connected to j if (j,l) are removed from $\mathcal{T}$ for all $l \in \mathcal{N}_{\mathrm{s}}^{(j)}$	$\mathring{\mathbf{z}}_t^{(j)}$	concatenation of node j's intra-node observation as well as the inter-node observations obtained by node j with all its neighbors in $\mathcal{N}^{(j)}$ at time t
$\mathring{\mathbf{z}}_{0:t}^{(j)}$	vertical concatenation of node j's observations $\mathring{\mathbf{z}}_0^{(j)}, \mathring{\mathbf{z}}_1^{(j)}, \ldots, \mathring{\mathbf{z}}_t^{(j)}$ from time 0 to time t	$\mathring{\mathbf{r}}_t^{(j)}$	messages received by node j from its neighbors in $\mathcal{N}^{(j)}$ at time t
$\mathring{\mathbf{r}}_{0:t}^{(j)}$	vertical concatenation of received messages $\mathring{\mathbf{r}}_0^{(j)}, \mathring{\mathbf{r}}_1^{(j)}, \ldots, \mathring{\mathbf{r}}_t^{(j)}$ from time 0 to time t	$\mathring{\boldsymbol{\Gamma}}^{(j)}$	sensor gain matrix corresponding to $\mathring{\mathbf{z}}_t^{(j)}$
$\boldsymbol{O}\big(\mathring{\boldsymbol{\Gamma}}^{(j)}, \boldsymbol{A}\big)$	observability matrix corresponding to observations obtained by node j	$\mathcal{S}(\mathcal{V}_0)$	observable subspace corresponding to observations obtained by nodes in a subset $\mathcal{V}_0$ of $\mathcal{V}$
$\mathcal{H}^{(ij)}$	invariant encoding subspace for generating encoded messages from node i to node j	$\boldsymbol{H}^{(ij)}$	matrix whose columns form an orthonormal basis of $\mathcal{H}^{(ij)}$
$\breve{C}^{(ij)}\big(\alpha^{(j)}\big)$	$\alpha^{(j)}$-anytime capacity for the channel from node i to node j	$\gamma_{\mathcal{T}}^{(ij)}$	threshold for $\breve{C}^{(ij)}\big(\alpha^{(j)}\big)$ in the communication subcondition of the sufficient condition
$\mathbf{y}_t^{(ij)}$	estimator of $\big(\boldsymbol{H}^{(ij)}\big)^{\mathrm{T}}\mathbf{x}_t$ computed by node i at time t using $\mathring{\mathbf{z}}_{0:t}^{(i)}$ and $\mathring{\mathbf{r}}_{0:t-1}^{(i)}$	$\hat{\mathbf{y}}_t^{(ij)}$	estimator of $\mathbf{y}_t^{(ij)}$ computed by node j at time t using $\mathring{\mathbf{r}}_{0:t}^{(ij)}$
$\hat{\boldsymbol{\xi}}_t^{(j)}$	local estimator of $\boldsymbol{O}\big(\mathring{\boldsymbol{\Gamma}}^{(j)}, \boldsymbol{A}\big)\mathbf{x}_t$ computed by node j at time t using $\mathring{\mathbf{z}}_{0:t}^{(j)}$	$\hat{\mathbf{x}}_t^{(j)}$	designed distributed estimator of $\mathbf{x}_t^{(j)}$ computed by node j at time t

is shown in Fig. 2. A discrete-time model is adopted for these quantities with details presented below.

- Unknown state: The d-dimensional state of node j at time t is represented by $\mathbf{x}_t^{(j)} \in \mathbb{R}^d$. In particular, $\mathbf{x}_t^{(j)}$ satisfies
$$\mathbf{x}_t^{(j)} = \boldsymbol{A}^{(j)}\mathbf{x}_{t-1}^{(j)} + \boldsymbol{\zeta}_t^{(j)} \quad \text{for } t = 1, 2, \ldots \tag{1}$$
where $\boldsymbol{A}^{(j)}$ is a known matrix called dynamic matrix of node j, and $\boldsymbol{\zeta}_t^{(j)}$ is a zero-mean random vector representing the disturbance to the state. In this paper, the dimensions of the states for different nodes are the same for notational simplicity. The extension to scenarios where these dimensions are different is straightforward.
- Observations: Node j is equipped with sensors j_{a} and j_{b} for obtaining intra-node observations and inter-node observations, respectively. Specifically, node j obtains an intra-node observation $\mathbf{z}_t^{(jj)}$ and obtains an inter-node observation $\mathbf{z}_t^{(ji)}$ with each neighbor $i \in \mathcal{N}_{\mathrm{u}}^{(j)}$ at time t. Therefore, the observations obtained by node j at time t consist of $\mathbf{z}_t^{(jj)}$ and $\mathbf{z}_t^{(ji)}$ for all $i \in \mathcal{N}_{\mathrm{u}}^{(j)}$. In particular, $\mathbf{z}_t^{(jj)}$ depends on node j's state and can be written as
$$\mathbf{z}_t^{(jj)} = \boldsymbol{\Gamma}^{(jj)}\mathbf{x}_t^{(j)} + \mathbf{n}_t^{(jj)} \quad \text{for } t = 0, 1, \ldots \tag{2}$$
where $\boldsymbol{\Gamma}^{(jj)}$ is a known sensor gain matrix, and $\mathbf{n}_t^{(jj)}$ is a zero-mean random vector representing observation noise. The observation $\mathbf{z}_t^{(ji)}$ depends on the states of node j as well as node i and can be written as
$$\mathbf{z}_t^{(ji)} = \boldsymbol{\Gamma}_1^{(ji)}\mathbf{x}_t^{(j)} + \boldsymbol{\Gamma}_2^{(ji)}\mathbf{x}_t^{(i)} + \mathbf{n}_t^{(ji)} \quad \text{for } t = 0, 1, \ldots \tag{3}$$
where $\boldsymbol{\Gamma}_1^{(ji)}$ and $\boldsymbol{\Gamma}_2^{(ji)}$ are known sensor gain matrices, and $\mathbf{n}_t^{(ji)}$ is a zero-mean random vector representing observation noise.

The sensor gain matrices $\boldsymbol{\Gamma}^{(jj)}$, $\boldsymbol{\Gamma}_1^{(ji)}$, and $\boldsymbol{\Gamma}_2^{(ji)}$ determine the sensing capability of node j. For example, if $\boldsymbol{\Gamma}^{(jj)} = \boldsymbol{I}$, then the intra-node observation $\mathbf{z}_t^{(jj)}$ is a noisy version of node j's state $\mathbf{x}_t^{(j)}$ and is informative for inferring $\mathbf{x}_t^{(j)}$. By contrast, if $\boldsymbol{\Gamma}^{(jj)} = \mathbf{0}$, then $\mathbf{z}_t^{(jj)}$

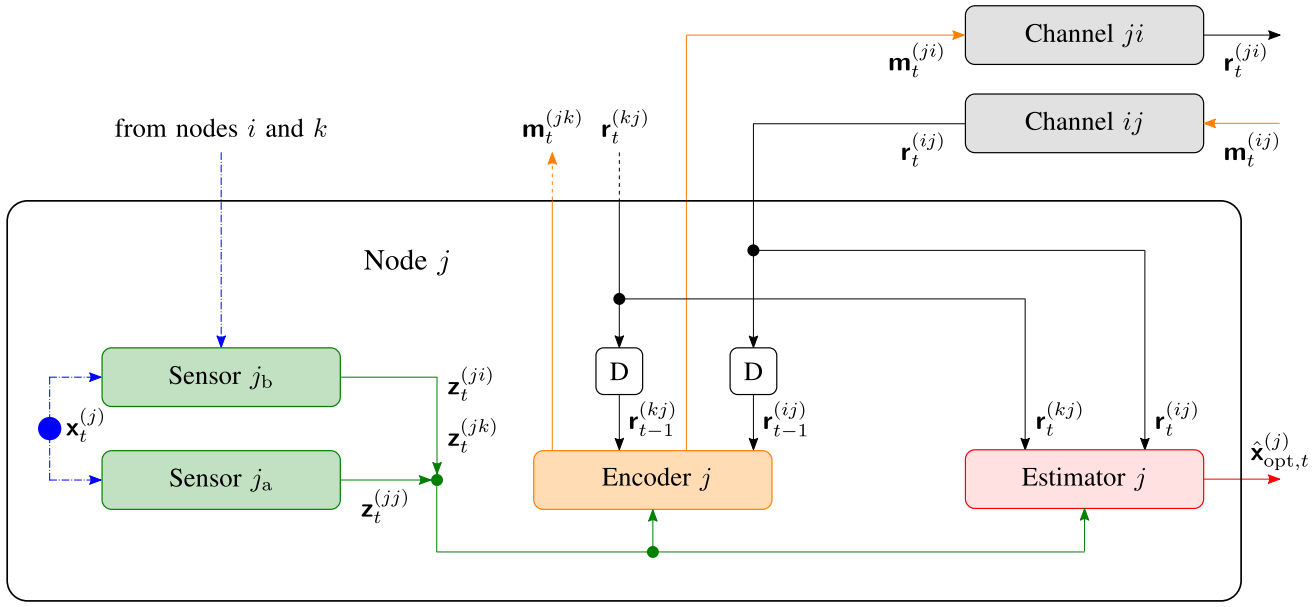


Fig. 2. Block diagram of node j: sensor j_a observes $\mathbf{z}_t^{(jj)}$ whereas sensor j_b observes $\mathbf{z}_t^{(ji)}$ and $\mathbf{z}_t^{(jk)}$ at time t. The observations and received messages are used for generating encoded messages $\mathbf{m}_t^{(ji)}$ and $\mathbf{m}_t^{(jk)}$ and for computing estimator $\hat{\mathbf{x}}_{\mathrm{opt},t}^{(j)}$.

contains only noise $\mathbf{n}_t^{(jj)}$ and is thus not useful for inferring $\mathbf{x}_t^{(j)}$.

- Encoded message: Node j performs real-time encoding and transmits encoded message $\mathbf{m}_t^{(ji)}$ to its neighbor $i \in \mathcal{N}_\mathrm{u}^{(j)}$ via channel ji. The transmitted message $\mathbf{m}_t^{(ji)}$ is a deterministic function of the local observations made by node j up to time t as well as the messages received by node j up to time $t-1$. Specifically, let $\mathbf{r}_t^{(kj)}$ represent the data received by node j from node $k \in \mathcal{N}_\mathrm{u}^{(j)}$ at a general time step t. Then $\mathbf{m}_t^{(ji)}$ can be written as

$$\mathbf{m}_t^{(ji)} = \boldsymbol{\mu}_t^{(ji)}\left(\mathbf{z}_{0:t}^{(jj)}, \left\{\mathbf{z}_{0:t}^{(jk)}, \mathbf{r}_{0:t-1}^{(kj)} : k \in \mathcal{N}_\mathrm{u}^{(j)}\right\}\right)$$

where $\boldsymbol{\mu}_t^{(ji)}$ is referred to as the encoding function of node j for node i at time t. Moreover, the sequence of encoding functions $\boldsymbol{\mu}_0^{(ji)}, \boldsymbol{\mu}_1^{(ji)}, \ldots$ is referred to as the encoding strategy of node j for node i.

The following assumptions are made on the initial state $\mathbf{x}_0^{(j)}$, state disturbance $\boldsymbol{\zeta}_t^{(j)}$, as well as observation noise $\mathbf{n}_t^{(jj)}$ and $\mathbf{n}_t^{(ji)}$.

A1. Vectors $\mathbf{x}_0^{(j)}$, $\boldsymbol{\zeta}_t^{(j)}$, $\mathbf{n}_t^{(jj)}$, and $\mathbf{n}_t^{(ji)}$ have probability densities for all $j \in \mathcal{V}$, $i \in \mathcal{N}_\mathrm{u}^{(j)}$, and $t \geqslant 0$.

A2. Vectors $\boldsymbol{\zeta}_t^{(j)}$ are independent over time t. Similarly, $\mathbf{n}_t^{(jj)}$ and $\mathbf{n}_t^{(ji)}$ are independent over t for all $j \in \mathcal{V}$ and $i \in \mathcal{N}_\mathrm{u}^{(j)}$. Moreover, $\mathbf{x}_0^{(j)}$, $\{\boldsymbol{\zeta}_t^{(j)}\}_{t\geqslant 0}$, $\{\mathbf{n}_t^{(jj)}\}_{t\geqslant 0}$, and $\{\mathbf{n}_t^{(ji)}\}_{t\geqslant 0}$ are independent over all $j \in \mathcal{V}$ and $i \in \mathcal{N}_\mathrm{u}^{(j)}$.

A3. For any real number $a > 0$, sequences $\{\mathbb{E}\{\|\boldsymbol{\zeta}_t^{(j)}\|^a\}\}_{t\geqslant 0}$, $\{\mathbb{E}\{\|\mathbf{n}_t^{(jj)}\|^a\}\}_{t\geqslant 0}$, and $\{\mathbb{E}\{\|\mathbf{n}_t^{(ji)}\|^a\}\}_{t\geqslant 0}$ are bounded from above for all $j \in \mathcal{V}$ and $i \in \mathcal{N}_\mathrm{u}^{(j)}$.

Assumption A3 holds if the tail of the distribution for each entry of $\boldsymbol{\zeta}_t^{(j)}$, $\mathbf{n}_t^{(jj)}$, and $\mathbf{n}_t^{(ji)}$ is not heavy. As an example, if $\boldsymbol{\zeta}_t^{(j)}$ has an identical distribution for all time steps t, and each entry of $\boldsymbol{\zeta}_t^{(j)}$ is a sub-exponential random variable,[1] then $\{\mathbb{E}\{\|\boldsymbol{\zeta}_t^{(j)}\|^a\}\}_{t\geqslant 0}$ is bounded from above for all $a > 0$. This example shows that Assumption A3 applies to various types of distributions for $\boldsymbol{\zeta}_t^{(j)}$, $\mathbf{n}_t^{(jj)}$, and $\mathbf{n}_t^{(ji)}$. Therefore, unlike some existing works with strong assumptions on the distributions of the disturbance and noise (e.g., they have Gaussian distributions or bounded supports), the model considered in this paper is general.

For simplicity of the presentation, the magnitudes of all the eigenvalues of $\boldsymbol{A}^{(j)}$ are considered to be no smaller than one for all $j \in \mathcal{V}$. Results in this paper can be extended to scenarios where the magnitudes of certain eigenvalues of $\boldsymbol{A}^{(j)}$ are smaller than one.

The following assumptions are made for the channels in the network.

A4. Given the transmitted message $\mathbf{m}_t^{(ji)}$, the received message $\mathbf{r}_t^{(ji)}$ is conditionally independent of $\mathbf{x}_0^{(j)}$, $\{\boldsymbol{\zeta}_t^{(j)}\}_{t\geqslant 0}$, $\{\mathbf{n}_t^{(jj)}\}_{t\geqslant 0}$, and $\{\mathbf{n}_t^{(ji)}\}_{t\geqslant 0}$.

A5. The channel between each pair of nodes is memoryless. Moreover, the channel state information is known to both the transmitter and the receiver.

Moreover, the paper considers scenarios where a multiple access scheme such as time-division multiple access or frequency-division multiple access is employed so that the communications among different channels do not interfere with each other. In fact, the sufficient condition established in Section V can be used to facilitate the design of a multiple access scheme.

The network contains a subset $\mathcal{V}_\mathrm{a} \subseteq \mathcal{V}$ of nodes referred to as agents. Note that $\mathcal{V}_\mathrm{a}$ can be any non-empty subset of $\mathcal{V}$, from a singleton $\{j\}$ to the entire set $\mathcal{V}$. An agent node j aims

[1] A random variable x is sub-exponential if there exists a constant $c > 0$ such that $\mathbb{P}\{|\mathsf{x}| > x\} \leqslant 2\exp\{-cx\}$ for any $x \geqslant 0$ [57, Chapter 2].

to learn its state $\mathbf{x}_t^{(j)}$ in a distributed manner using its local observations and messages received from its neighbors (see Fig. 2). In particular, a distributed estimator $\hat{\mathbf{x}}_t^{(j)}$ of $\mathbf{x}_t^{(j)}$ is a function of node j's observations $\{\mathbf{z}_{0:t}^{(jj)}\} \cup \{\mathbf{z}_{0:t}^{(ji)} : i \in \mathcal{N}_{\mathrm{u}}^{(j)}\}$ and received messages $\{\mathbf{r}_{0:t}^{(ij)} : i \in \mathcal{N}_{\mathrm{u}}^{(j)}\}$ up to time t. A widely used metric for the inference error associated with $\hat{\mathbf{x}}_t^{(j)}$ is its mean-square error (MSE). To minimize the MSE, the optimal distributed estimator for node j at time t is the distributed minimum-mean-square-error (MMSE) estimator $\hat{\mathbf{x}}_{\mathrm{opt},t}^{(j)}$ given by the conditional expectation

$$\hat{\mathbf{x}}_{\mathrm{opt},t}^{(j)} := \mathbb{E}\Big\{\mathbf{x}_t^{(j)} \,\Big|\, \mathbf{z}_{0:t}^{(jj)}, \{\mathbf{z}_{0:t}^{(ji)}, \mathbf{r}_{0:t}^{(ij)} : i \in \mathcal{N}_{\mathrm{u}}^{(j)}\}\Big\}.$$

The MSE of this estimator, which is referred to as the individual distributed inference MSE of agent j at time t, is given by

$$\varepsilon_t^{(j)} := \mathbb{E}\Big\{\big\|\hat{\mathbf{x}}_{\mathrm{opt},t}^{(j)} - \mathbf{x}_t^{(j)}\big\|^2\Big\}. \tag{4}$$

The objective function F_t at time t for the learning problem is the total distributed inference MSE of all the agents when distributed MMSE estimators are employed, i.e.,

$$F_t := \sum_{j \in \mathcal{V}_{\mathrm{a}}} \varepsilon_t^{(j)}.$$

The objective function F_t is affected by the messages exchanged across the network, which are determined by the encoding strategies employed by the nodes. The fundamental limit of distributed inference is the value of F_t achieved by optimal encoding strategies. Such a value is determined by the sensing and communication capabilities of the network. The paper aims to study the fundamental limit of distributed inference by establishing a sufficient condition under which encoding strategies can be designed so that the sequence $\{F_t\}_{t\geqslant 0}$ is bounded from above, i.e.,

$$\sup_{t\geqslant 0} F_t < \infty. \tag{5}$$

This is equivalent to

$$\sup_{t\geqslant 0} \varepsilon_t^{(j)} < \infty \quad \forall j \in \mathcal{V}_{\mathrm{a}}. \tag{6}$$

In other words, the boundedness of the total distributed inference MSE over time is equivalent to the boundedness of the individual distributed inference MSE of each agent. Boundedness of error is an important property of estimators and it has been studied in [25], [58], and [59]. Moreover, the boundedness of $\{F_t\}_{t\geqslant 0}$ can be viewed as the stability of the inference error $\hat{\mathbf{x}}_{\mathrm{opt},t}^{(j)} - \mathbf{x}_t^{(j)}$ for all $j \in \mathcal{V}_{\mathrm{a}}$. Stability is so critical for dynamic systems that it has piqued significant research interest [46], [51], [60].

The distributed learning problem described above is nonlinear even though the state evolution and observation models are linear, as the encoding function $\boldsymbol{\mu}_t^{(ji)}$ is not limited to be linear. This increases the difficulty of the problem as conventional linear estimation techniques such as Kalman filtering cannot be directly applied. The problem studied in this paper is different from those in many existing works on the distributed inference of a global unknown state. These works typically employ a consensus mechanism to ensure that the estimators of the global unknown state at different nodes in the network are consistent. By contrast, each agent in our problem aims to infer in real time its own state, and thus a consensus mechanism is not employed.

III. Preliminaries

This section presents preliminaries on invariant subspaces and real generalized eigenspaces as well as notions of anytime reliability and anytime capacity.

A. Invariant Subspaces and Real Generalized Eigenspaces

First, the notion of invariant subspace, which is used for the design of real-time encoder in Section V, is introduced. Consider a subspace $\mathcal{Y} \subseteq \mathbb{R}^n$ and a linear mapping $\boldsymbol{f} : \mathbb{R}^n \mapsto \mathbb{R}^n$ defined as $\boldsymbol{f}(\boldsymbol{u}) = \boldsymbol{F}\boldsymbol{u}$ for all $\boldsymbol{u} \in \mathbb{R}^n$, where $\boldsymbol{F}$ is an n-by-n real matrix. A subspace $\mathcal{Y}$ is said to be $\boldsymbol{F}$-invariant if and only if $\boldsymbol{F}\boldsymbol{u} \in \mathcal{Y}$ for all $\boldsymbol{u} \in \mathcal{Y}$. The sum and intersection of a finite number of $\boldsymbol{F}$-invariant subspaces are also $\boldsymbol{F}$-invariant. Define subspace $\mathcal{I}_{\boldsymbol{F}}(\mathcal{Y})$ as

$$\mathcal{I}_{\boldsymbol{F}}(\mathcal{Y}) := \mathcal{C}\big(\begin{bmatrix}\boldsymbol{Y} & \boldsymbol{F}\boldsymbol{Y} & \cdots & \boldsymbol{F}^{n-1}\boldsymbol{Y}\end{bmatrix}\big) \tag{7}$$

where $\boldsymbol{Y}$ is a matrix whose columns form a basis of $\mathcal{Y}$. Subspace $\mathcal{I}_{\boldsymbol{F}}(\mathcal{Y})$ contains $\mathcal{Y}$ and is $\boldsymbol{F}$-invariant. Subspace $\mathcal{I}_{\boldsymbol{F}}(\mathcal{Y})$ is referred to as the minimum $\boldsymbol{F}$-invariant subspace over $\mathcal{Y}$ [61, Chapter 2], since there is no $\boldsymbol{F}$-invariant subspace that both contains $\mathcal{Y}$ and is strictly contained by $\mathcal{I}_{\boldsymbol{F}}(\mathcal{Y})$. Moreover,

$$\mathcal{I}_{\boldsymbol{F}}(\mathcal{Y}_1 + \mathcal{Y}_2) = \mathcal{I}_{\boldsymbol{F}}(\mathcal{Y}_1) + \mathcal{I}_{\boldsymbol{F}}(\mathcal{Y}_2) \tag{8}$$

for any two subspaces $\mathcal{Y}_1, \mathcal{Y}_2 \subseteq \mathbb{R}^n$.

Next, the notion of real generalized eigenspace is introduced. Let $\Lambda(\boldsymbol{F})$ represent the set of all the eigenvalues of $\boldsymbol{F} \in \mathbb{R}^{n\times n}$ that are either real or have positive imaginary parts. Mathematically,

$$\Lambda(\boldsymbol{F}) := \big\{\lambda : \mathrm{Im}(\lambda) \geqslant 0;\, \exists \boldsymbol{u} \in \mathbb{C}^n, \boldsymbol{u} \neq \boldsymbol{0} \text{ s.t. } \boldsymbol{F}\boldsymbol{u} = \lambda\boldsymbol{u}\big\} \tag{9}$$

where $\mathrm{Im}(\lambda)$ represents the imaginary part of λ. A real generalized eigenspace $\mathcal{M}_\lambda(\boldsymbol{F})$ is defined for each $\lambda \in \Lambda(\boldsymbol{F})$ as

$$\mathcal{M}_\lambda(\boldsymbol{F}) := \begin{cases} \mathcal{M}_\lambda^{\mathrm{C}}(\boldsymbol{F}) \cap \mathbb{R}^n & \text{if } \lambda \in \mathbb{R} \\ \big(\mathcal{M}_\lambda^{\mathrm{C}}(\boldsymbol{F}) + \mathcal{M}_{\lambda^*}^{\mathrm{C}}(\boldsymbol{F})\big) \cap \mathbb{R}^n & \text{otherwise} \end{cases} \tag{10}$$

where

$$\mathcal{M}_\lambda^{\mathrm{C}}(\boldsymbol{F}) := \big\{\boldsymbol{u} \in \mathbb{C}^n : \big(\boldsymbol{F} - \lambda\boldsymbol{I}\big)^n \boldsymbol{u} = \boldsymbol{0}\big\} \tag{11}$$

represents the subspace spanned by generalized eigenvectors [62] of $\boldsymbol{F}$ associated with λ over the field $\mathbb{C}$ of complex numbers. Definition (10) indicates that $\mathcal{M}_\lambda(\boldsymbol{F})$ is a set of real vectors. If $\lambda \in \mathbb{R}$, then each element of $\mathcal{M}_\lambda(\boldsymbol{F})$ is a generalized eigenvector of $\boldsymbol{F}$ associated with λ. Otherwise, each element of $\mathcal{M}_\lambda(\boldsymbol{F})$ can be written as the sum of two generalized eigenvectors of $\boldsymbol{F}$ associated with λ and λ^*, respectively. The set $\mathcal{M}_\lambda(\boldsymbol{F})$ is a subspace of $\mathbb{R}^n$ and we refer to such a subspace as a real generalized eigenspace of $\boldsymbol{F}$ associated with λ. Real generalized eigenspaces can be

used for representing a real square matrix as its real Jordan canonical form via similarity transformation. This technique has been used for the investigation of dynamic systems in the literature. It is also used in this paper for analyzing the accuracy of the designed distributed estimator.

Real generalized eigenspace $\mathcal{M}_\lambda(\boldsymbol{F})$ can be shown to be $\boldsymbol{F}$-invariant. Moreover, the next proposition shows a decomposition of invariant subspaces using real generalized eigenspaces.

Proposition 1: For any real square matrix $\boldsymbol{F}$, an arbitrary $\boldsymbol{F}$-invariant subspace $\mathcal{Y}$ can be decomposed as

$$\mathcal{Y} = \bigoplus_{\lambda \in \Lambda(\boldsymbol{F})} \left(\mathcal{Y} \cap \mathcal{M}_\lambda(\boldsymbol{F})\right).$$

Notation $\oplus$ in this proposition represents the direct sum of subspaces and its definition can be found in [62, Chapter 4]. Proposition 1 is adapted from the theorem stating that a complex invariant subspace can be decomposed using complex generalized eigenspaces [62, Chapter 4]. Different from that theorem, Proposition 1 focuses on the decomposition of a real subspace using real generalized eigenspaces. As a special case, consider $\boldsymbol{F} \in \mathbb{R}^{n\times n}$ and $\mathcal{Y} = \mathbb{R}^n$. Proposition 1 shows that $\mathbb{R}^n = \oplus_{\lambda\in\Lambda(\boldsymbol{F})}\mathcal{M}_\lambda(\boldsymbol{F})$.

B. Anytime Reliability and Anytime Capacity

The notions of anytime reliability and anytime capacity were introduced in [55]. Consider a communication system consisting of a transmitter node and a receiver node. The transmitter aims to send a sequence of data symbols to the receiver via a channel. Unlike classical block-coding where all the data symbols are available to the transmitter beforehand, the transmitter in this communication system obtains data symbols sequentially: at each time t, a new data symbol $\mathbf{s}_t$ from an alphabet with 2^r elements is available to the transmitter. Based on available data symbols $\mathbf{s}_{0:t}$, the transmitter generates an encoded message $\mathbf{m}_t$ and sends it at each time t. In other words, the transmitter performs real-time encoding without waiting for the entire data symbol sequence to be available. Such a transmitter is referred to as an anytime encoder.

Let $\mathbf{r}_t$ represent the message obtained by the receiver at time t, which can be different from $\mathbf{m}_t$ due to impairments in the communication channel. At each time, the receiver estimates all the symbols that have been sent by the transmitter. Specifically, at time t, the receiver generates an estimator $\hat{\mathbf{s}}_{t'}(\mathbf{r}_{0:t})$ of $\mathbf{s}_{t'}$ based on $\mathbf{r}_{0:t}$ for all $t' \in \{0, 1, \ldots, t\}$. In other words, the receiver decodes the transmitted symbols without waiting for the transmitter to complete sending the entire data symbol sequence. Such a receiver is referred to as an anytime decoder. The estimators $\hat{\mathbf{s}}_{t'}(\mathbf{r}_{0:t_1})$ and $\hat{\mathbf{s}}_{t'}(\mathbf{r}_{0:t_2})$ for $\mathbf{s}_{t'}$ computed at two different times $t' \leqslant t_1 < t_2$ are not necessarily equal. In fact, the accuracy of the estimator $\hat{\mathbf{s}}_{t'}(\mathbf{r}_{0:t_2})$ is expected to be higher compared with that of $\hat{\mathbf{s}}_{t'}(\mathbf{r}_{0:t_1})$ as more messages have been received at t_2 than at t_1.

The system described above is called rate r sequential communication system. This system achieves *α-anytime reliability* if there exists a constant K such that

$$\mathbb{P}\left\{\hat{\mathbf{s}}_{0:t'}(\mathbf{r}_{0:t}) \neq \mathbf{s}_{0:t'}\right\} \leqslant K 2^{-\alpha(t-t')} \quad \forall\, t \geqslant 0,\ 0 \leqslant t' \leqslant t \tag{12}$$

where $\hat{\mathbf{s}}_{0:t'}(\mathbf{r}_{0:t})$ represents the concatenation of $\hat{\mathbf{s}}_0(\mathbf{r}_{0:t})$, $\hat{\mathbf{s}}_1(\mathbf{r}_{0:t}), \ldots, \hat{\mathbf{s}}_{t'}(\mathbf{r}_{0:t})$. Anytime reliability is interpreted as follows. The left-hand side of (12) represents the probability of decoding error for the transmitted symbols $\mathbf{s}_{0:t'}$ at time t, i.e., the probability of error if a decoding delay of $t - t'$ time steps is allowed for estimating $\mathbf{s}_{0:t'}$. Anytime reliability requires that such probability should decrease at least exponentially fast with respect to the delay $t - t'$ at rate α. Note that t is an arbitrary integer that is no smaller than t', which indicates that $\mathbf{s}_{0:t'}$ is required to be estimated for any delay. This is the reason for the qualifier "anytime".

The *α-anytime capacity* $\breve{C}(\alpha)$ of a channel is defined as the smallest upper bound of r such that a rate r sequential communication system that achieves α-anytime reliability exists. The anytime capacity $\breve{C}(\alpha)$ is a monotonically non-increasing function of the parameter α. For a memoryless channel, it holds that $\breve{C}(\alpha) \leqslant C$ for any $\alpha \geqslant 0$, where C represents the Shannon capacity of the channel. Similar to Shannon capacity, the anytime capacity of a channel is affected by the communication resource available to the transmitter (e.g., power and bandwidth) and the signal degradation introduced by the channel (e.g., fading and noise). More properties of anytime capacity can be found in [55], [63], [64].

IV. Notions for Establishing Sufficient Condition

This section presents important notions for establishing the sufficient condition on the boundedness of the total distributed inference MSE over time. First, this section introduces a concatenated state and describes ordered trees based on the graph corresponding to the network. Then, it introduces the observable subspace. Finally, this section presents invariant encoding subspaces (IESs), which are used for the design of encoders and distributed estimators detailed in Section V.

A. Concatenated State

Define the concatenated state $\mathbf{x}_t$, noise $\boldsymbol{\zeta}_t$, and matrix $\boldsymbol{A}$ as

$$\mathbf{x}_t := \left[\left(\mathbf{x}_t^{(1)}\right)^{\mathrm{T}} \quad \left(\mathbf{x}_t^{(2)}\right)^{\mathrm{T}} \quad \cdots \quad \left(\mathbf{x}_t^{(v)}\right)^{\mathrm{T}}\right]^{\mathrm{T}} \tag{13a}$$

$$\boldsymbol{\zeta}_t := \left[\left(\boldsymbol{\zeta}_t^{(1)}\right)^{\mathrm{T}} \quad \left(\boldsymbol{\zeta}_t^{(2)}\right)^{\mathrm{T}} \quad \cdots \quad \left(\boldsymbol{\zeta}_t^{(v)}\right)^{\mathrm{T}}\right]^{\mathrm{T}} \tag{13b}$$

$$\boldsymbol{A} := \operatorname{diag}\left\{\boldsymbol{A}^{(1)}, \boldsymbol{A}^{(2)}, \ldots, \boldsymbol{A}^{(v)}\right\}. \tag{13c}$$

where $v := |\mathcal{V}|$ is the number of nodes in the network. According to (1) and (13),

$$\mathbf{x}_t = \boldsymbol{A}\mathbf{x}_{t-1} + \boldsymbol{\zeta}_t \tag{14}$$

$$\mathbf{x}_t^{(j)} = (\boldsymbol{e}_{j,v} \otimes \boldsymbol{I}_d)^{\mathrm{T}}\mathbf{x}_t \tag{15}$$

where for a given positive integer m, $\boldsymbol{e}_{k,m}$ represents a unit vector of m entries with its kth entry being one and other entries being zero for $1 \leqslant k \leqslant m$. An agent j aims to infer

$(\boldsymbol{e}_{j,v} \otimes \boldsymbol{I}_d)^{\mathrm{T}} \mathbf{x}_t$ according to (15). Denote the column space of $\boldsymbol{e}_{j,v} \otimes \boldsymbol{I}_d$ by $\mathcal{X}^{(j)}$, i.e.,

$$\mathcal{X}^{(j)} := \mathcal{C}(\boldsymbol{e}_{j,v} \otimes \boldsymbol{I}_d) \subseteq \mathbb{R}^{dv}. \tag{16}$$

Subspace $\mathcal{X}^{(j)}$ can be viewed as a subspace that corresponds to $\mathbf{x}_t^{(j)}$ and is thus referred to as node j's state subspace.

Observations $\mathbf{z}_t^{(jj)}$ and $\mathbf{z}_t^{(ji)}$ can be expressed in terms of the concatenated state $\mathbf{x}_t$. To this end, define $\mathring{\boldsymbol{\Gamma}}^{(jj)}$ and $\mathring{\boldsymbol{\Gamma}}^{(ji)}$ as

$$\mathring{\boldsymbol{\Gamma}}^{(jj)} := \boldsymbol{e}_{j,v}^{\mathrm{T}} \otimes \boldsymbol{\Gamma}^{(jj)} \tag{17a}$$

$$\mathring{\boldsymbol{\Gamma}}^{(ji)} := \boldsymbol{e}_{j,v}^{\mathrm{T}} \otimes \boldsymbol{\Gamma}_1^{(ji)} + \boldsymbol{e}_{i,v}^{\mathrm{T}} \otimes \boldsymbol{\Gamma}_2^{(ji)}. \tag{17b}$$

Then $\mathbf{z}_t^{(jj)}$ and $\mathbf{z}_t^{(ji)}$ can be written as

$$\mathbf{z}_t^{(jj)} = \mathring{\boldsymbol{\Gamma}}^{(jj)} \mathbf{x}_t + \mathbf{n}_t^{(jj)} \tag{18a}$$

$$\mathbf{z}_t^{(ji)} = \mathring{\boldsymbol{\Gamma}}^{(ji)} \mathbf{x}_t + \mathbf{n}_t^{(ji)}. \tag{18b}$$

The introduction of $\mathbf{x}_t$ as well as the view of $\mathbf{z}_t^{(jj)}$ and $\mathbf{z}_t^{(ji)}$ as observations of $\mathbf{x}_t$ only serve to facilitate the development and presentation of the results. They do not change either the system model or the decentralized nature of the inference problem being studied. In particular, (14) is equivalent to (1), whereas (18) is equivalent to (2) and (3). Moreover, each node performs inference based only on its local observations and messages received from its neighbors.

B. Ordered Tree Based on the Graph

Recall that $\mathcal{G}_{\mathrm{u}}$ is an undirected graph constructed based on whether nodes in the network are within the communication range of each other, as described in Section II. Here, we introduce a tree referred to as an ordered tree based on $\mathcal{G}_{\mathrm{u}}$. First, construct an arbitrary spanning tree of $\mathcal{G}_{\mathrm{u}}$ by removing edges from it. Then, assign a node in $\mathcal{V}$ as the root node of the tree, based on which a neighbor of each node is classified as either the parent or a child of that node. Finally, an order of all the children of each node is specified. The tree $\mathcal{T}$ constructed in this manner is referred to as an ordered tree based on $\mathcal{G}_{\mathrm{u}}$. Define $\mathcal{E}$ as the set of edges in $\mathcal{T}$, and let $\mathcal{N}^{(j)}$ represent the set of neighbors of node j in $\mathcal{T}$, i.e.,

$$\mathcal{N}^{(j)} := \{i : (i,j) \in \mathcal{E}\}.$$

Note that $\mathcal{E}$ and $\mathcal{N}^{(j)}$ are subsets of $\mathcal{E}_{\mathrm{u}}$ and $\mathcal{N}_{\mathrm{u}}^{(j)}$ presented in Section II, respectively. In addition, define $N^{(j)} := |\mathcal{N}^{(j)}|$ as the number of neighbors for node j in $\mathcal{T}$. If node j is the root of $\mathcal{T}$, then all its neighbors are its children with the nth child denoted by $\check{c}_n^{(j)}$ for $n \in \{1, 2, \ldots, N^{(j)}\}$. Otherwise node j has $N^{(j)} - 1$ children and one parent. In this case, the nth child is denoted by $\check{c}_n^{(j)}$ for $n \in \{1, 2, \ldots, N^{(j)} - 1\}$ and the parent is denoted by $\check{c}_{N^{(j)}}^{(j)}$.

Given an ordered tree $\mathcal{T}$, define $\mathring{\mathbf{z}}_t^{(j)}$ as the concatenation of node j's intra-node observation as well as the inter-node observations obtained by node j with all its neighbors in $\mathcal{N}^{(j)}$ at time t. Mathematically,

$$\mathring{\mathbf{z}}_t^{(j)} := \left[\left(\mathbf{z}_t^{(jj)}\right)^{\mathrm{T}} \quad \left(\mathbf{z}_t^{(j\,i_1)}\right)^{\mathrm{T}} \quad \left(\mathbf{z}_t^{(j\,i_2)}\right)^{\mathrm{T}} \quad \cdots \quad \left(\mathbf{z}_t^{(j\,i_n)}\right)^{\mathrm{T}} \right]^{\mathrm{T}}. \tag{19}$$

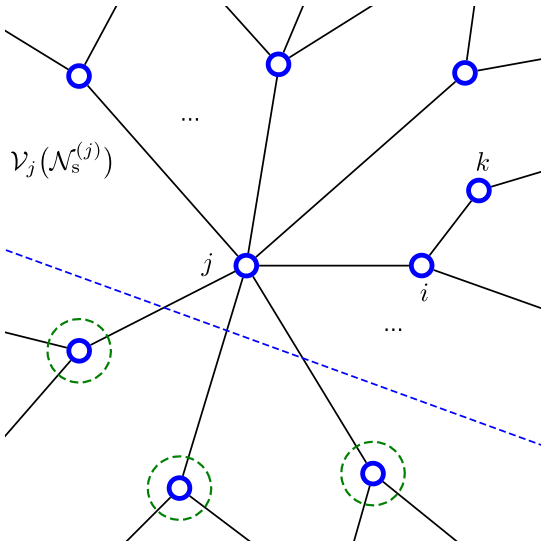


Fig. 3. Set $\mathcal{V}_j(\mathcal{N}_{\mathrm{s}}^{(j)})$ in an arbitrary ordered tree based on the graph: $\mathcal{N}_{\mathrm{s}}^{(j)}$ consists of the three nodes inside dashed green circles, whereas $\mathcal{V}_j(\mathcal{N}_{\mathrm{s}}^{(j)})$ consists of nodes to the upper-right of the dashed blue line.

where indices $i_1, i_2, \ldots, i_n$ are all the elements of $\mathcal{N}^{(j)}$. Moreover, define $\mathring{\mathbf{r}}_t^{(j)}$ as the messages received by node j from its neighbors in $\mathcal{N}^{(j)}$ at time t, i.e.,

$$\mathring{\mathbf{r}}_t^{(j)} := \left[\left(\mathbf{r}_t^{(i_1\,j)}\right)^{\mathrm{T}} \quad \left(\mathbf{r}_t^{(i_2\,j)}\right)^{\mathrm{T}} \quad \cdots \quad \left(\mathbf{r}_t^{(i_n\,j)}\right)^{\mathrm{T}} \right]^{\mathrm{T}}. \tag{20}$$

Finally, a useful subset of nodes is defined given an arbitrary ordered tree $\mathcal{T}$ based on $\mathcal{G}_{\mathrm{u}}$. For any $j \in \mathcal{V}$ and subset $\mathcal{N}_{\mathrm{s}}^{(j)}$ of its neighbor set $\mathcal{N}^{(j)}$, define $\mathcal{V}_j(\mathcal{N}_{\mathrm{s}}^{(j)}) \subseteq \mathcal{V}$ as a subset consisting of node j and nodes that are connected to j if edges (j,l) are removed from $\mathcal{T}$ for all $l \in \mathcal{N}_{\mathrm{s}}^{(j)}$. Mathematically,

$$\mathcal{V}_j(\mathcal{N}_{\mathrm{s}}^{(j)}) := \{j\} \cup \Big\{ k \in \mathcal{V} : k \leftrightarrow j \text{ in graph } \big\{\mathcal{V}, \mathcal{E} \setminus \{(j,l) : l \in \mathcal{N}_{\mathrm{s}}^{(j)}\}\big\} \Big\} \tag{21}$$

where $k \leftrightarrow j$ represents that there is a path between nodes k and j. This definition is illustrated in Fig. 3.

C. Observable Subspace

The observations $\mathring{\mathbf{z}}_t^{(j)}$ can be expressed as a function of the concatenated state $\mathbf{x}_t$. To this end, define the sensor gain matrix $\mathring{\boldsymbol{\Gamma}}^{(j)}$ and concatenated observation noise $\mathring{\mathbf{n}}_t^{(j)}$ as

$$\mathring{\boldsymbol{\Gamma}}^{(j)} := \left[\left(\mathring{\boldsymbol{\Gamma}}^{(jj)}\right)^{\mathrm{T}} \quad \left(\mathring{\boldsymbol{\Gamma}}^{(j\,i_1)}\right)^{\mathrm{T}} \quad \left(\mathring{\boldsymbol{\Gamma}}^{(j\,i_2)}\right)^{\mathrm{T}} \cdots \left(\mathring{\boldsymbol{\Gamma}}^{(j\,i_n)}\right)^{\mathrm{T}} \right]^{\mathrm{T}}$$

$$\mathring{\mathbf{n}}_t^{(j)} := \left[\left(\mathbf{n}_t^{(jj)}\right)^{\mathrm{T}} \quad \left(\mathbf{n}_t^{(j\,i_1)}\right)^{\mathrm{T}} \quad \left(\mathbf{n}_t^{(j\,i_2)}\right)^{\mathrm{T}} \quad \cdots \left(\mathbf{n}_t^{(j\,i_n)}\right)^{\mathrm{T}} \right]^{\mathrm{T}}$$

where indices $i_1, i_2, \ldots, i_n$ are all the elements of $\mathcal{N}^{(j)}$. Combining these definitions with (18) and (19), $\mathring{\mathbf{z}}_t^{(j)}$ can be written as

$$\mathring{\mathbf{z}}_t^{(j)} = \mathring{\boldsymbol{\Gamma}}^{(j)} \mathbf{x}_t + \mathring{\mathbf{n}}_t^{(j)}. \tag{22}$$

Viewing $\mathring{\mathbf{z}}_t^{(j)}$ as observations of $\mathbf{x}_t$, the observability matrix corresponding to observations obtained by node j is $\boldsymbol{O}(\mathring{\boldsymbol{\Gamma}}^{(j)}, \boldsymbol{A})$, where $\boldsymbol{O}(\boldsymbol{G}, \boldsymbol{F})$ is defined as [65], [66], [67]

$$\boldsymbol{O}(\boldsymbol{G}, \boldsymbol{F}) := \left[\boldsymbol{G}^{\mathrm{T}} \quad \boldsymbol{F}^{\mathrm{T}}\boldsymbol{G}^{\mathrm{T}} \quad \cdots \quad (\boldsymbol{F}^{k-1})^{\mathrm{T}}\boldsymbol{G}^{\mathrm{T}} \right]^{\mathrm{T}} \tag{23}$$

for general matrices $\boldsymbol{F} \in \mathbb{R}^{k\times k}$ and $\boldsymbol{G}$ with k columns. Define $\mathcal{S}(\{j\})$ as the column space of $\boldsymbol{O}\big(\mathring{\boldsymbol{\Gamma}}^{(j)}, \boldsymbol{A}\big)^{\mathrm{T}}$, i.e.,

$$\mathcal{S}(\{j\}) := \mathcal{C}\Big(\boldsymbol{O}\big(\mathring{\boldsymbol{\Gamma}}^{(j)}, \boldsymbol{A}\big)^{\mathrm{T}}\Big) \quad \forall j \in \mathcal{V} \tag{24}$$

and we call it the observable subspace corresponding to observations obtained by node j.

Next, we present a lemma on constructing a local estimator of a linear transformation of the concatenated state. This lemma provides insight on the observable subspace $\mathcal{S}(\{j\})$ and is used for designing distributed encoders and estimators.

Lemma 1: Any node j in the network can construct a local estimator $\hat{\boldsymbol{\xi}}_t^{(j)}$ of $\boldsymbol{O}\big(\mathring{\boldsymbol{\Gamma}}^{(j)}, \boldsymbol{A}\big)\mathbf{x}_t$ at each time $t \geqslant 0$ using its local observations $\mathring{\mathbf{z}}_{0:t}^{(j)}$ such that the following inequality holds for any $a \geqslant 2$

$$\sup_{t\geqslant 0} \mathbb{E}\Big\{\big\|\hat{\boldsymbol{\xi}}_t^{(j)} - \boldsymbol{O}\big(\mathring{\boldsymbol{\Gamma}}^{(j)}, \boldsymbol{A}\big)\mathbf{x}_t\big\|^a\Big\} < \infty\,. \tag{25}$$

Proof: See Appendix A. □

Lemma 1 indicates that node j can construct an estimator of $\boldsymbol{O}\big(\mathring{\boldsymbol{\Gamma}}^{(j)}, \boldsymbol{A}\big)\mathbf{x}_t$ with bounded error using only its observations. This is why the column space $\mathcal{S}(\{j\})$ of $\boldsymbol{O}\big(\mathring{\boldsymbol{\Gamma}}^{(j)}, \boldsymbol{A}\big)^{\mathrm{T}}$ is referred to as an observable subspace in this paper.

Remark 1: The relationship between $\mathcal{X}^{(j)}$ and $\mathcal{S}(\{j\})$ determines whether node j is able to compute an estimator of $\mathbf{x}_t^{(j)}$ with bounded MSE using only its local observations $\mathring{\mathbf{z}}_{0:t}^{(j)}$. Specifically, if $\mathcal{X}^{(j)} \subseteq \mathcal{S}(\{j\})$, then $\boldsymbol{e}_{j,v}\otimes\boldsymbol{I}_d$ can be written as $\boldsymbol{e}_{j,v}\otimes\boldsymbol{I}_d = \boldsymbol{O}\big(\mathring{\boldsymbol{\Gamma}}^{(j)}, \boldsymbol{A}\big)^{\mathrm{T}}\boldsymbol{\Phi}^{(jj)}$ for some deterministic matrix $\boldsymbol{\Phi}^{(jj)}$. Then, node j can employ $\big(\boldsymbol{\Phi}^{(jj)}\big)^{\mathrm{T}}\hat{\boldsymbol{\xi}}_t^{(j)}$ as an estimator of $\mathbf{x}_t^{(j)} = (\boldsymbol{e}_{j,v}\otimes\boldsymbol{I}_d)^{\mathrm{T}}\mathbf{x}_t$, and it can be shown using (25) that $\sup_{t\geqslant 0}\mathbb{E}\big\{\big\|\big(\boldsymbol{\Phi}^{(jj)}\big)^{\mathrm{T}}\hat{\boldsymbol{\xi}}_t^{(j)} - \mathbf{x}_t^{(j)}\big\|^2\big\} < \infty$. In other words, node j can construct an estimator of $\mathbf{x}_t^{(j)}$ whose MSE is bounded over time using only its local observations and not its received messages. Of course, the relationship $\mathcal{X}^{(j)} \subseteq \mathcal{S}(\{j\})$ does not hold in general. Consequently, node j requires not only its local observations but also its received messages in order to construct a distributed estimator of $\mathbf{x}_t^{(j)}$ whose MSE is bounded over time. This will be explained in Section IV-D.

The definition of $\mathcal{S}(\{j\})$ can be generalized to cases where the argument of $\mathcal{S}(\cdot)$ is an arbitrary non-empty subset $\mathcal{V}_0 \subseteq \mathcal{V}$. Specifically, define $\mathcal{S}(\mathcal{V}_0)$ as

$$\mathcal{S}(\mathcal{V}_0) := \mathcal{C}\Big(\boldsymbol{O}\big(\big[\mathring{\boldsymbol{\Gamma}}^{(j)}\big]_{j\in\mathcal{V}_0}, \boldsymbol{A}\big)^{\mathrm{T}}\Big) \tag{26}$$

where $\big[\mathring{\boldsymbol{\Gamma}}^{(j)}\big]_{j\in\mathcal{V}_0}$ represents the vertical concatenation of $\mathring{\boldsymbol{\Gamma}}^{(j)}$ for all $j \in \mathcal{V}_0$. Subspace $\mathcal{S}(\mathcal{V}_0)$ is referred to as the observable subspace corresponding to observations obtained by nodes in $\mathcal{V}_0$. Definition (24) is a special case of (26) with $\mathcal{V}_0 = \{j\}$. Subspace $\mathcal{S}(\mathcal{V}_0)$ can be interpreted in a similar manner as $\mathcal{S}(\{j\})$. In particular, if $\mathcal{X}^{(j)} \subseteq \mathcal{S}(\mathcal{V}_0)$, then an estimator of $\mathbf{x}_t^{(j)}$ whose MSE is bounded over time can be constructed based on observations $\big\{\mathring{\mathbf{z}}_{0:t}^{(j)} : j \in \mathcal{V}_0\big\}$. Subspace $\mathcal{S}(\mathcal{V}_0)$ has the following properties: for any subsets $\mathcal{V}_0 \subseteq \mathcal{V}$ and $\mathcal{V}_0' \subseteq \mathcal{V}$, it holds that

$$\begin{aligned}\mathcal{S}(\mathcal{V}_0) &= \mathcal{I}_{\boldsymbol{A}^{\mathrm{T}}}\big(\mathcal{S}(\mathcal{V}_0)\big)\\ \mathcal{S}(\mathcal{V}_0\cup\mathcal{V}_0') &= \mathcal{S}(\mathcal{V}_0) + \mathcal{S}(\mathcal{V}_0')\,.\end{aligned} \tag{27}$$

In particular, the first equality shows that $\mathcal{S}(\mathcal{V}_0)$ is $\boldsymbol{A}^{\mathrm{T}}$-invariant.

D. Invariant Encoding Subspace (IES)

This subsection presents the notion of IES, which is used for constructing encoders and estimators in distributed learning. Given an ordered tree $\mathcal{T}$ based on $\mathscr{G}_{\mathrm{u}}$, a subspace $\mathcal{H}^{(ij)}$ named IES can be constructed for each $j \in \mathcal{V}$ and $i \in \mathcal{N}^{(j)}$. Specifically, an IES is a subspace that satisfies the properties described in the following proposition, where we recall that $\mathcal{V}_{\mathrm{a}}$ is the set consisting of all the agents in the network.

Proposition 2: Consider an arbitrary ordered tree $\mathcal{T}$ based on $\mathscr{G}_{\mathrm{u}}$. If the state subspace of every agent is contained in the observable subspace corresponding to observations obtained by all the nodes in the network, namely

$$\mathcal{X}^{(j)} \subseteq \mathcal{S}(\mathcal{V}) \tag{28}$$

for all $j \in \mathcal{V}_{\mathrm{a}}$, then there exists an $\boldsymbol{A}^{\mathrm{T}}$-invariant subspace $\mathcal{H}^{(ij)} \subseteq \mathbb{R}^{dv}$ for every $j \in \mathcal{V}$ and $i \in \mathcal{N}^{(j)}$ with the following properties

$$\mathcal{X}^{(j)} \subseteq \mathcal{S}(\{j\}) + \sum_{i\in\mathcal{N}^{(j)}} \mathcal{H}^{(ij)} \quad \forall j \in \mathcal{V}_{\mathrm{a}} \tag{29a}$$

$$\mathcal{H}^{(ij)} \subseteq \mathcal{S}(\{i\}) + \sum_{k\in\mathcal{N}^{(i)}\setminus\{j\}} \mathcal{H}^{(ki)} \quad \forall j \in \mathcal{V},\ i \in \mathcal{N}^{(j)} \tag{29b}$$

$$\mathcal{H}^{(ij)} \subseteq \mathcal{X}^{(i)} + \mathcal{X}^{(j)} \quad \forall j \in \mathcal{V},\ i \in \mathcal{N}^{(j)}. \tag{29c}$$

Furthermore, there exists a subspace $\mathcal{G}_\lambda^{(ij)} \subseteq \mathcal{M}_\lambda(\boldsymbol{A}^{\mathrm{T}})$ for each $j \in \mathcal{V}$, $i \in \mathcal{N}^{(j)}$, and $\lambda \in \Lambda\big(\boldsymbol{A}^{(j)}\big)$ such that

$$\mathcal{H}^{(ij)} = \sum_{\lambda\in\Lambda(\boldsymbol{A}^{(j)})} \mathcal{I}_{\boldsymbol{A}^{\mathrm{T}}}\big(\mathcal{G}_\lambda^{(ij)}\big) \tag{30}$$

and the dimension of $\mathcal{G}_\lambda^{(ij)}$ is given by (59) in Appendix B.

Proof: See Appendix B. In particular, $\mathcal{G}_\lambda^{(ij)}$ is given in (63) if node i is a child of node j and is given in (64) if node i is the parent of node j. □

Define $\boldsymbol{H}^{(ij)}$ as a matrix whose columns form an orthonormal basis of $\mathcal{H}^{(ij)}$, i.e.,

$$\mathcal{C}\big(\boldsymbol{H}^{(ij)}\big) = \mathcal{H}^{(ij)}, \quad \big(\boldsymbol{H}^{(ij)}\big)^{\mathrm{T}}\boldsymbol{H}^{(ij)} = \boldsymbol{I} \quad \forall j \in \mathcal{V},\ i \in \mathcal{N}^{(j)}. \tag{31}$$

The IES $\mathcal{H}^{(ij)}$ is employed by the designed encoder for generating the transmitted messages from node i to node j. In particular, the properties of IES given in (29) are important for the design of distributed encoder and estimator as well as for proving the sufficient condition. Here, the design of distributed estimator at agent j using (29a) is explained as an example. As described in Section IV-C, if $\mathcal{X}^{(j)} \not\subseteq \mathcal{S}(\{j\})$, then node j needs to use both its local observations and its received messages for computing a distributed estimator of $\mathbf{x}_t^{(j)} = (\boldsymbol{e}_{j,v}\otimes\boldsymbol{I}_d)^{\mathrm{T}}\mathbf{x}_t$ with bounded MSE. Specifically, if (28) holds, and thus (29a) holds according to Proposition 2, then $\boldsymbol{e}_{j,v}\otimes\boldsymbol{I}_d$ can be written as a linear combination of $\boldsymbol{O}\big(\mathring{\boldsymbol{\Gamma}}^{(j)}, \boldsymbol{A}\big)^{\mathrm{T}}$ and $\big\{\boldsymbol{H}^{(ij)} : i \in \mathcal{N}^{(j)}\big\}$, i.e.,

$$\boldsymbol{e}_{j,v}\otimes\boldsymbol{I}_d = \boldsymbol{O}\big(\mathring{\boldsymbol{\Gamma}}^{(j)}, \boldsymbol{A}\big)^{\mathrm{T}}\boldsymbol{\Phi}^{(jj)} + \sum_{i\in\mathcal{N}^{(j)}} \boldsymbol{H}^{(ij)}\boldsymbol{\Phi}^{(ij)} \tag{32}$$

for some matrices $\boldsymbol{\Phi}^{(jj)}$ and $\{\boldsymbol{\Phi}^{(ij)} : i \in \mathcal{N}^{(j)}\}$. Here, we recall that the column spaces of $\boldsymbol{e}_{j,v} \otimes \boldsymbol{I}_d$ and $\boldsymbol{O}\big(\mathring{\boldsymbol{\Gamma}}^{(j)}, \boldsymbol{A}\big)^{\mathrm{T}}$ are $\mathcal{X}^{(j)}$ and $\mathcal{S}(\{j\})$ according to (16) and (24), respectively. Taking transpose of (32) and right-multiplying it with $\mathbf{x}_t$ gives

$$\mathbf{x}_t^{(j)} = \big(\boldsymbol{\Phi}^{(jj)}\big)^{\mathrm{T}} \boldsymbol{O}\big(\mathring{\boldsymbol{\Gamma}}^{(j)}, \boldsymbol{A}\big)\mathbf{x}_t + \sum_{i \in \mathcal{N}^{(j)}} \big(\boldsymbol{\Phi}^{(ij)}\big)^{\mathrm{T}} \big(\boldsymbol{H}^{(ij)}\big)^{\mathrm{T}} \mathbf{x}_t \,. \tag{33}$$

At time t, agent j computes an estimator of $\boldsymbol{O}\big(\mathring{\boldsymbol{\Gamma}}^{(j)}, \boldsymbol{A}\big)\mathbf{x}_t$ using its local observations $\mathring{\mathbf{z}}_{0:t}^{(j)}$ and computes an estimator of $\big(\boldsymbol{H}^{(ij)}\big)^{\mathrm{T}}\mathbf{x}_t$ using messages $\mathbf{r}_{0:t}^{(ij)}$ received from node i for each $i \in \mathcal{N}^{(j)}$. Then, agent j linearly combines these estimators using coefficient matrices $\boldsymbol{\Phi}^{(jj)}$ and $\big\{\boldsymbol{\Phi}^{(ij)} : i \in \mathcal{N}^{(j)}\big\}$ to obtain a distributed estimator of $\mathbf{x}_t^{(j)}$. Details of such a distributed estimator is presented in Section V-C.

Similar to (29a), relationship (29b) is used for designing the encoder as detailed in Section V-B. Furthermore, (29c) shows that $\dim(\mathcal{H}^{(ij)}) \leqslant \dim(\mathcal{X}^{(i)}) + \dim(\mathcal{X}^{(j)}) = 2d$. In other words, the dimension of $\mathcal{H}^{(ij)}$ is at most twice the number of entries in $\mathbf{x}_t^{(j)}$, which is not affected by the size of the network. Finally, (30) shows how $\mathcal{H}^{(ij)}$ is constructed.

V. Sufficient Condition for the Boundedness of Total Distributed Inference MSE

This section first states the established sufficient condition for the boundedness of the total distributed inference MSE over time. Then, the designed distributed encoder and estimator are presented, and the accuracy of the designed estimator is analyzed. Finally, the tightness of the established condition as well as the communication efficiency of the designed encoder and estimator are discussed.

A. Statement of the Sufficient Condition

The next theorem presents a sufficient condition for the existence of encoding strategies that ensure the learning objective function, namely the total distributed inference MSE, is bounded over time.

Theorem 1: Consider the distributed learning problem presented in Section II. For this problem, a sufficient condition for achieving $\sup_{t \geqslant 0} F_t < \infty$ is given as follows: there exists an ordered tree $\mathcal{T}$ based on $\mathscr{G}_{\mathrm{u}}$ such that both of the following two subconditions hold:

(i) Relationship (28) holds for every $j \in \mathcal{V}_{\mathrm{a}}$.

(ii) The $\alpha^{(j)}$-anytime capacity $\breve{C}^{(ij)}\big(\alpha^{(j)}\big)$ of the channel from node i to node j satisfies

$$\breve{C}^{(ij)}\big(\alpha^{(j)}\big) > \sum_{\lambda \in \Lambda(\boldsymbol{A}^{(j)})} \dim\big(\mathcal{I}_{\boldsymbol{A}^{\mathrm{T}}}\big(\mathcal{G}_{\lambda}^{(ij)}\big)\big) \log|\lambda| =: \gamma_{\mathcal{T}}^{(ij)} \qquad \forall j \in \mathcal{V},\, i \in \mathcal{N}^{(j)} \tag{34}$$

for $\alpha^{(j)}$ given by

$$\alpha^{(j)} := 2^{D+3} \max_{\lambda \in \Lambda(\boldsymbol{A}^{(j)})} \log|\lambda| \tag{35}$$

where D is the diameter of $\mathcal{T}$.[2] Here, $\mathcal{G}_{\lambda}^{(ij)}$ is a subspace specified in Proposition 2, and operator $\mathcal{I}$ is defined in (7).

Proof: The theorem is proved by designing a real-time encoder for every node to each of its neighbors and by designing a distributed estimator at each agent. In particular, the MSE of the distributed estimator at every agent is shown to be bounded over time if the sufficient condition in Theorem 1 holds. This indicates that (6) holds, as the MSE of the distributed MMSE estimator is no greater than that of the designed distributed estimator. Since (6) is equivalent to (5), the desired result is proved. Details of the proof are presented in the following sections: Sections V-B and V-C present the designed encoder and estimator, respectively. Section V-D analyzes the MSE of the designed distributed estimator. Some details of the proof are presented in Appendices C and D. □

We clarify the statement of Theorem 1 with respect to the ordered tree $\mathcal{T}$. Given the graph $\mathscr{G}_{\mathrm{u}}$, an ordered tree $\mathcal{T}$ based on this graph can be constructed in different manners (e.g., different assignments of roots and specifications of the order for each node's children). The construction of $\mathcal{T}$ affects the definition of subspace $\mathcal{G}_{\lambda}^{(ij)}$ and thus affects the threshold $\gamma_{\mathcal{T}}^{(ij)}$ for the anytime capacity. Theorem 1 states that if Subcondition (i) holds and if Subcondition (ii) holds for any ordered tree, then encoding strategies can be designed to ensure that the total distributed inference MSE is bounded over time. A case study of distributed learning in a network with three nodes is presented in Section VI-A to explain Theorem 1.

Theorem 1 is interpreted as follows. Subcondition (i) of Theorem 1 describes a sensing capability of the network for ensuring that the total distributed inference MSE is bounded over time. In particular, $\mathcal{X}^{(j)}$ in (28) is the subspace corresponding to the state $\mathbf{x}_t^{(j)}$ of node j, whereas $\mathcal{S}(\mathcal{V})$ represents the observable subspace corresponding to observations obtained by all the nodes in the network. Subcondition (i) indicates that the network has sufficient sensing capability such that the state of every agent is observable given observations obtained by the entire network.

To better understand Subcondition (i), consider the scenario where there is no communication constraint in the network so that every node can transmit infinite amount of data to each of its neighbors without any loss. In this scenario, the anytime capacity of each link becomes infinity, and Subcondition (ii) holds automatically. Consequently, Theorem 1 states that if Subcondition (i) holds, then a distributed estimator can be designed for each agent such that the total distributed inference MSE is bounded over time. In fact, each agent can collect the observations of all the nodes in the network in this scenario since there is no communication constraint. As a result, the observable subspace corresponding to these observations becomes $\mathcal{S}(\mathcal{V})$. As discussed in Section IV-A, if such an observable subspace satisfies (28) for all $j \in \mathcal{V}_{\mathrm{a}}$, then each agent can construct an estimator of its state using

[2]The diameter of a tree is defined as the largest of shortest-path distances between all pairs of nodes in the tree, where the shortest-path distance between a pair of nodes is the smallest number of edges on any path connecting the two nodes [68].

the observations of all the nodes such that the MSE of this estimator is bounded over time.

In addition to sensing capability, the network also needs to have enough communication capability so that each agent can obtain useful information from received messages. In particular, Subcondition (ii) of Theorem 1 describes a communication capability of the network for ensuring that the total distributed inference MSE is bounded. This subcondition states that if the anytime capacity of the channel from node i to node j is above a threshold for all $j \in \mathcal{V}$ and $i \in \mathcal{N}^{(j)}$, then there exists an encoding strategy for each node in the network to each of its neighbors such that the total distributed inference MSE is bounded over time.

Remark 2: The threshold $\gamma_{\mathcal{T}}^{(ij)}$ for the anytime capacity $\breve{C}^{(ij)}\big(\alpha^{(j)}\big)$ in (34) increases with the magnitudes of eigenvalues of dynamic matrix $\boldsymbol{A}^{(j)}$ and the dimensions of subspaces $\mathcal{I}_{\boldsymbol{A}^{\mathrm{T}}}\big(\mathcal{G}_{\lambda}^{(ij)}\big)$. This is explained as follows. First, as the magnitudes of eigenvalues of $\boldsymbol{A}^{(j)}$ become larger, the unknown disturbances to agent j's state in the past will have more significant effect on its current state (see (1)). Consequently, agent j needs to extract more information from its received messages in order to achieve desirable distributed inference accuracy. This demands stronger communication capabilities between agent j and its neighbors and thus $\gamma_{\mathcal{T}}^{(ij)}$ becomes larger. Second, as will be shown in Section V-B, node i transmits to node j the information of an estimator $\mathbf{y}_t^{(ij)}$ for $\big(\boldsymbol{H}^{(ij)}\big)^{\mathrm{T}}\mathbf{x}_t$ at each time t. Here, we recall that the columns of $\boldsymbol{H}^{(ij)}$ form an orthonormal basis of $\mathcal{H}^{(ij)}$. According to (30), if the dimension of $\mathcal{I}_{\boldsymbol{A}^{\mathrm{T}}}\big(\mathcal{G}_{\lambda}^{(ij)}\big)$ increases, then the dimension of $\mathcal{H}^{(ij)}$ also increases. Consequently, $\mathbf{y}_t^{(ij)}$ becomes a vector with a larger dimension, and thus node i needs to transmit more information to agent j at each time step. This requires a better communication channel from node i to node j and thus $\gamma_{\mathcal{T}}^{(ij)}$ increases.

A favorable property of the threshold $\gamma_{\mathcal{T}}^{(ij)}$ is that it is not affected by the number of nodes in the network irrespective of the specification of the ordered tree $\mathcal{T}$. In particular, combining (58c) and the definition (7), and noting that $\boldsymbol{A}^{\mathrm{T}}$ is block-diagonal according to (13c), we can show that $\mathcal{I}_{\boldsymbol{A}^{\mathrm{T}}}\big(\mathcal{G}_{\lambda}^{(ij)}\big) \subseteq \mathcal{X}^{(i)} + \mathcal{X}^{(j)}$, and thus $\dim\big(\mathcal{I}_{\boldsymbol{A}^{\mathrm{T}}}\big(\mathcal{G}_{\lambda}^{(ij)}\big)\big) \leqslant 2d$. Consequently, the threshold on the anytime capacity $\gamma_{\mathcal{T}}^{(ij)}$ does not depend on the number of nodes in the network. This shows desirable scalability of the designed encoder and estimator. In particular, scalability is important for modern networks, which can consist of a massive number of nodes thanks to the proliferation of mobile devices and Internet-of-Things.

B. Design of Encoders for Distributed Learning

If the sufficient condition given in Theorem 1 holds, then there exists an ordered tree $\mathcal{T}$ such that (34) is satisfied. The designed encoder and estimator based on $\mathcal{T}$ are presented in this section and in Section V-C, respectively.

Consider the encoding procedure performed by node i for generating messages to its neighbor j. At time t, node i computes an estimator $\mathbf{y}_t^{(ij)}$ of $\big(\boldsymbol{H}^{(ij)}\big)^{\mathrm{T}}\mathbf{x}_t$ using its observations $\mathring{\mathbf{z}}_{0:t}^{(i)}$ and received messages $\mathring{\mathbf{r}}_{0:t-1}^{(i)}$, where the definitions of $\mathring{\mathbf{z}}_t^{(j)}$ and $\mathring{\mathbf{r}}_t^{(j)}$ for a general $j \in \mathcal{V}$ and $t \geqslant 0$ are given in (19) and (20), respectively. To construct $\mathbf{y}_t^{(ij)}$ at node i, the property (29b) of an IES is employed. Specifically, if (28) holds, and thus (29b) holds according to Proposition 2, then $\boldsymbol{H}^{(ij)}$ can be written as a linear combination of $\boldsymbol{O}\big(\mathring{\boldsymbol{\Gamma}}^{(i)}, \boldsymbol{A}\big)^{\mathrm{T}}$ and $\big\{\boldsymbol{H}^{(ki)} : k \in \mathcal{N}^{(i)} \setminus \{j\}\big\}$, i.e.,

$$\boldsymbol{H}^{(ij)} = \boldsymbol{O}\big(\mathring{\boldsymbol{\Gamma}}^{(i)}, \boldsymbol{A}\big)^{\mathrm{T}}\boldsymbol{\Psi}_i^{(ij)} + \sum_{k\in\mathcal{N}^{(i)}\setminus\{j\}} \boldsymbol{H}^{(ki)}\boldsymbol{\Psi}_k^{(ij)} \tag{36}$$

for some matrices $\boldsymbol{\Psi}_i^{(ij)}$ and $\big\{\boldsymbol{\Psi}_k^{(ij)} : k \in \mathcal{N}^{(i)} \setminus \{j\}\big\}$. Here, we recall that the column space of matrix $\boldsymbol{H}^{(ij)}$, matrix $\boldsymbol{O}\big(\mathring{\boldsymbol{\Gamma}}^{(i)}, \boldsymbol{A}\big)^{\mathrm{T}}$, and matrix $\boldsymbol{H}^{(ki)}$ are $\mathcal{H}^{(ij)}$, $\mathcal{S}(\{i\})$, and $\mathcal{H}^{(ki)}$, respectively, according to (31) and (24). Taking transpose of (36) and right-multiplying it with $\mathbf{x}_t$ gives

$$\big(\boldsymbol{H}^{(ij)}\big)^{\mathrm{T}}\mathbf{x}_t = \big(\boldsymbol{\Psi}_i^{(ij)}\big)^{\mathrm{T}}\boldsymbol{O}\big(\mathring{\boldsymbol{\Gamma}}^{(i)}, \boldsymbol{A}\big)\mathbf{x}_t + \sum_{k\in\mathcal{N}^{(i)}\setminus\{j\}} \big(\boldsymbol{\Psi}_k^{(ij)}\big)^{\mathrm{T}}\big(\boldsymbol{H}^{(ki)}\big)^{\mathrm{T}}\mathbf{x}_t\,. \tag{37}$$

Node i first computes a local estimator $\hat{\boldsymbol{\xi}}_t^{(i)}$ of $\boldsymbol{O}\big(\mathring{\boldsymbol{\Gamma}}^{(i)}, \boldsymbol{A}\big)\mathbf{x}_t$ using its local observations $\mathring{\mathbf{z}}_{0:t}^{(i)}$, as described in Lemma 1. Moreover, node i computes an estimator $\big(\boldsymbol{H}^{(ki)}\big)^{\mathrm{T}}\boldsymbol{A}\boldsymbol{H}^{(ki)}\hat{\mathbf{y}}_{t-1}^{(ki)}$ of $\big(\boldsymbol{H}^{(ki)}\big)^{\mathrm{T}}\mathbf{x}_{t-1}$ using messages $\mathbf{r}_{0:t-1}^{(ki)}$ received from node k for each $k \in \mathcal{N}^{(i)} \setminus \{j\}$. Vector $\hat{\mathbf{y}}_{t-1}^{(ki)}$ is described in the next paragraph. Node i then linearly combines $\hat{\boldsymbol{\xi}}_t^{(i)}$ and $\big\{\big(\boldsymbol{H}^{(ki)}\big)^{\mathrm{T}}\boldsymbol{A}\boldsymbol{H}^{(ki)}\hat{\mathbf{y}}_{t-1}^{(ki)} : k \in \mathcal{N}^{(i)} \setminus \{j\}\big\}$ to generate the estimator $\mathbf{y}_t^{(ij)}$ of $\big(\boldsymbol{H}^{(ij)}\big)^{\mathrm{T}}\mathbf{x}_t$ as follows

$$\mathbf{y}_t^{(ij)} := \big(\boldsymbol{\Psi}_i^{(ij)}\big)^{\mathrm{T}}\hat{\boldsymbol{\xi}}_t^{(i)} + \sum_{k\in\mathcal{N}^{(i)}\setminus\{j\}} \big(\boldsymbol{\Psi}_k^{(ij)}\big)^{\mathrm{T}}\big(\boldsymbol{H}^{(ki)}\big)^{\mathrm{T}}\boldsymbol{A}\boldsymbol{H}^{(ki)}\hat{\mathbf{y}}_{t-1}^{(ki)}\,. \tag{38}$$

Finally, node i generates an encoded message $\mathbf{m}_t^{(ij)}$ based on $\mathbf{y}_{0:t}^{(ij)}$ and transmits this message to agent j. The generation of $\mathbf{m}_t^{(ij)}$ is presented in Section V-D. A block diagram for the encoding procedure performed by node i for generating messages to node j is shown in Fig. 4.

Random vector $\hat{\mathbf{y}}_{t-1}^{(ki)}$ is explained as follows. Analogous to the computation of $\mathbf{y}_t^{(ij)}$ by node i described above, node k computes $\mathbf{y}_{t-1}^{(ki)}$ as an estimator of $\big(\boldsymbol{H}^{(ki)}\big)^{\mathrm{T}}\mathbf{x}_{t-1}$ at time $t-1$. Based on $\mathbf{y}_{0:t-1}^{(ki)}$, node k generates an encoded message $\mathbf{m}_{t-1}^{(ki)}$ and transmits it to node i. According to $\mathbf{r}_{0:t-1}^{(ki)}$, node i computes an estimator $\hat{\mathbf{y}}_{t-1}^{(ki)}$ of $\mathbf{y}_{t-1}^{(ki)}$ at time $t-1$. Recall that $\mathbf{y}_{t-1}^{(ki)}$ is an estimator of $\big(\boldsymbol{H}^{(ki)}\big)^{\mathrm{T}}\mathbf{x}_{t-1}$. Therefore, writing $\hat{\mathbf{y}}_{t-1}^{(ki)} = \mathbf{y}_{t-1}^{(ki)} + \big(\hat{\mathbf{y}}_{t-1}^{(ki)} - \mathbf{y}_{t-1}^{(ki)}\big)$, we can view $\hat{\mathbf{y}}_{t-1}^{(ki)}$ also as an estimator of $\big(\boldsymbol{H}^{(ki)}\big)^{\mathrm{T}}\mathbf{x}_{t-1}$ with additional error $\hat{\mathbf{y}}_{t-1}^{(ki)} - \mathbf{y}_{t-1}^{(ki)}$ compared to $\mathbf{y}_{t-1}^{(ki)}$.

The employment of $\big(\boldsymbol{H}^{(ki)}\big)^{\mathrm{T}}\boldsymbol{A}\boldsymbol{H}^{(ki)}\hat{\mathbf{y}}_{t-1}^{(ki)}$ by node i as an estimator for $\big(\boldsymbol{H}^{(ki)}\big)^{\mathrm{T}}\mathbf{x}_t$ is explained as follows. Left multiplying (14) by $\big(\boldsymbol{H}^{(ki)}\big)^{\mathrm{T}}$, using the equality $\big(\boldsymbol{H}^{(ki)}\big)^{\mathrm{T}}\boldsymbol{H}^{(ki)} = \boldsymbol{I}$ and that $\mathcal{H}^{(ki)}$ is $\boldsymbol{A}^{\mathrm{T}}$-invariant (see

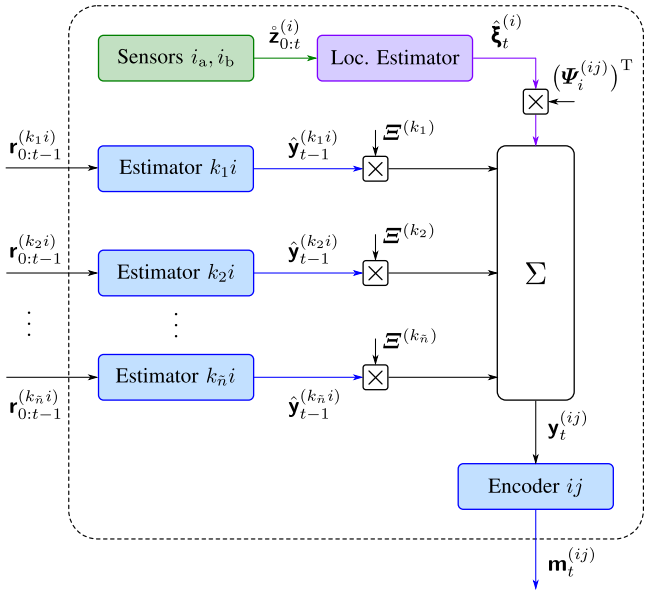


Fig. 4. Block diagram of the encoding procedure performed by node i for generating messages to node j, where the neighbor set of node i in $\mathcal{T}$ is $\mathcal{N}^{(i)} = \{j, k_1, k_2, \ldots, k_{\tilde{n}}\}$. In the figure, $\boldsymbol{\Xi}^{(k_\ell)} := (\boldsymbol{\Psi}_{k_\ell}^{(ij)})^{\mathrm{T}} (\boldsymbol{H}^{(k_\ell i)})^{\mathrm{T}} \boldsymbol{A} \boldsymbol{H}^{(k_\ell i)}$ for $\ell \in \{1, 2, \ldots, \tilde{n}\}$, and a block with a cross inside represents the multiplication of a matrix with a vector. Moreover, Estimator $k_\ell i$ and Encoder ij represent $\mathrm{Est}\big(\mathbf{y}_t^{(k_\ell i)}, (\boldsymbol{H}^{(k_\ell i)})^{\mathrm{T}} \boldsymbol{A} \boldsymbol{H}^{(k_\ell i)}, a_{k_\ell i}, a_{k_\ell i}/2\big)$ and $\mathrm{Enc}\big(\mathbf{y}_t^{(ij)}, (\boldsymbol{H}^{(ij)})^{\mathrm{T}} \boldsymbol{A} \boldsymbol{H}^{(ij)}, a_{ij}, a_{ij}/2\big)$, respectively (see Proposition 3).

Proposition 2), we can show

$$\begin{aligned}&(\boldsymbol{H}^{(ki)})^{\mathrm{T}} \boldsymbol{A} \boldsymbol{H}^{(ki)} (\boldsymbol{H}^{(ki)})^{\mathrm{T}} \mathbf{x}_{t-1} \\ &\qquad = (\boldsymbol{H}^{(ki)})^{\mathrm{T}} \mathbf{x}_t - (\boldsymbol{H}^{(ki)})^{\mathrm{T}} \boldsymbol{\zeta}_t \,. \end{aligned} \tag{39}$$

Therefore, if node i knows $(\boldsymbol{H}^{(ki)})^{\mathrm{T}} \mathbf{x}_{t-1}$, then $(\boldsymbol{H}^{(ki)})^{\mathrm{T}} \boldsymbol{A} \boldsymbol{H}^{(ki)} (\boldsymbol{H}^{(ki)})^{\mathrm{T}} \mathbf{x}_{t-1}$ can be used as an estimator of $(\boldsymbol{H}^{(ki)})^{\mathrm{T}} \mathbf{x}_t$ with error $-(\boldsymbol{H}^{(ki)})^{\mathrm{T}} \boldsymbol{\zeta}_t$. However, $(\boldsymbol{H}^{(ki)})^{\mathrm{T}} \mathbf{x}_{t-1}$ is unknown to node i and thus cannot be used as an estimator. On the other hand, node i can construct $\hat{\mathbf{y}}_{t-1}^{(ki)}$, which is an estimator of $(\boldsymbol{H}^{(ki)})^{\mathrm{T}} \mathbf{x}_{t-1}$ as described in the previous paragraph. Substituting $(\boldsymbol{H}^{(ki)})^{\mathrm{T}} \mathbf{x}_{t-1}$ by $\hat{\mathbf{y}}_{t-1}^{(ki)}$, node i employs $(\boldsymbol{H}^{(ki)})^{\mathrm{T}} \boldsymbol{A} \boldsymbol{H}^{(ki)} \hat{\mathbf{y}}_{t-1}^{(ki)}$ as an estimator of $(\boldsymbol{H}^{(ki)})^{\mathrm{T}} \mathbf{x}_t$ at time t.

C. Design of Estimators for Distributed Learning

Consider the distributed inference performed by agent j. As described in Section IV-D, at time t, agent j computes a local estimator $\hat{\boldsymbol{\xi}}_t^{(j)}$ of $\boldsymbol{O}(\mathring{\boldsymbol{I}}^{(j)}, \boldsymbol{A}) \mathbf{x}_t$ using its local observations $\mathring{\mathbf{z}}_{0:t}^{(j)}$ and computes an estimator $\hat{\mathbf{y}}_t^{(ij)}$ of $(\boldsymbol{H}^{(ij)})^{\mathrm{T}} \mathbf{x}_t$ using received messages $\mathbf{r}_{0:t}^{(ij)}$ for each $i \in \mathcal{N}^{(j)}$. These estimators are linearly combined at node j to generate an estimator $\hat{\mathbf{x}}_t^{(j)}$ of $\mathbf{x}_t^{(j)}$ as follows

$$\hat{\mathbf{x}}_t^{(j)} := (\boldsymbol{\Phi}^{(jj)})^{\mathrm{T}} \hat{\boldsymbol{\xi}}_t^{(j)} + \sum_{i \in \mathcal{N}^{(j)}} (\boldsymbol{\Phi}^{(ij)})^{\mathrm{T}} \hat{\mathbf{y}}_t^{(ij)} \tag{40}$$

where $\boldsymbol{\Phi}^{(jj)}$ and $\{\boldsymbol{\Phi}^{(ij)} : i \in \mathcal{N}^{(j)}\}$ are matrices that satisfy (33). A block diagram for the distributed estimator $\hat{\mathbf{x}}_t^{(j)}$ of agent j is shown in Fig. 5.

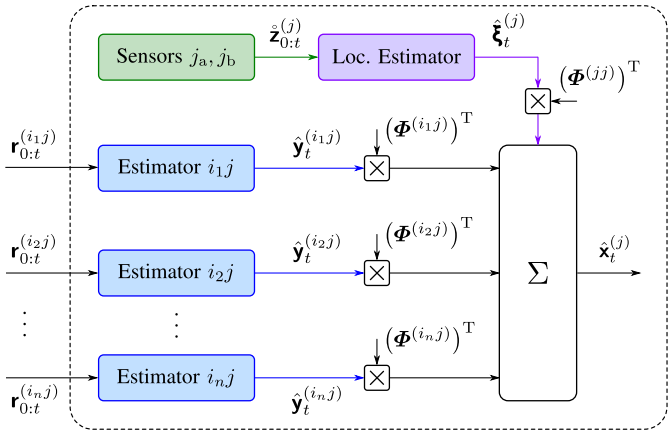


Fig. 5. Block diagram of the distributed estimator at agent j, where the neighbor set of agent j in $\mathcal{T}$ is $\mathcal{N}^{(j)} = \{i_1, i_2, \ldots, i_n\}$. In the figure, a block with a cross inside represents the multiplication of a matrix with a vector, and Estimator $i_\ell j$ represents $\mathrm{Est}\big(\mathbf{y}_t^{(i_\ell j)}, (\boldsymbol{H}^{(i_\ell j)})^{\mathrm{T}} \boldsymbol{A} \boldsymbol{H}^{(i_\ell j)}, a_{i_\ell j}, a_{i_\ell j}/2\big)$ (see Proposition 3), for $\ell \in \{1, 2, \ldots, n\}$.

Analogous to $\hat{\mathbf{y}}_{t-1}^{(ki)}$ described in Section V-B, $\hat{\mathbf{y}}_t^{(ij)}$ is an estimator of $\mathbf{y}_t^{(ij)}$ computed by node j at time t using messages $\mathbf{r}_{0:t}^{(ij)}$ received from node i. Since $\mathbf{y}_t^{(ij)}$ is an estimator of $(\boldsymbol{H}^{(ij)})^{\mathrm{T}} \mathbf{x}_t$, random vector $\hat{\mathbf{y}}_t^{(ij)}$ is also an estimator of $(\boldsymbol{H}^{(ij)})^{\mathrm{T}} \mathbf{x}_t$ with additional error $\hat{\mathbf{y}}_t^{(ij)} - \mathbf{y}_t^{(ij)}$.

D. Analysis of the Accuracy of the Distributed Estimator

In this subsection, the accuracy of the distributed estimator $\hat{\mathbf{x}}_t^{(j)}$ in (40) is analyzed and the desired result (5) is proved when the sufficient condition in Theorem 1 holds. According to (40), the accuracy of $\hat{\mathbf{x}}_t^{(j)}$ is affected by that of the local estimator $\hat{\boldsymbol{\xi}}_t^{(j)}$ and of $\{\hat{\mathbf{y}}_t^{(ij)} : i \in \mathcal{N}^{(j)}\}$. In this subsection, the accuracy of $\hat{\boldsymbol{\xi}}_t^{(j)}$ and $\hat{\mathbf{y}}_t^{(ij)}$ is analyzed to prove (5).

First, $\hat{\boldsymbol{\xi}}_t^{(j)}$ has been shown to satisfy (25). Second, node j can construct an estimator $\hat{\mathbf{y}}_t^{(ij)}$ of $\mathbf{y}_t^{(ij)}$ such that the estimation error $\hat{\mathbf{y}}_t^{(ij)} - \mathbf{y}_t^{(ij)}$ satisfies

$$\sup_{t \geqslant 0} \mathbb{E}\Big\{ \big\| \hat{\mathbf{y}}_t^{(ij)} - \mathbf{y}_t^{(ij)} \big\|^{a_{ij}/2} \Big\} < \infty \tag{41}$$

where $a_{ij} \geqslant 4$ is a scalar whose definition can be found in (67) of Appendix D. To show (41), the following inequality proved in Appendix D via induction is employed:

$$\sup_{t \geqslant 0} \mathbb{E}\Big\{ \big\| \mathbf{y}_t^{(kl)} - (\boldsymbol{H}^{(kl)})^{\mathrm{T}} \mathbf{x}_t \big\|^{a_{kl}} \Big\} < \infty \quad \forall l \in \mathcal{V},\ k \in \mathcal{N}^{(l)}. \tag{42}$$

Inequality (42) shows that the a_{kl}th moment of the Euclidean norm of $\mathbf{y}_t^{(kl)} - (\boldsymbol{H}^{(kl)})^{\mathrm{T}} \mathbf{x}_t$, which is the error of $\mathbf{y}_t^{(kl)}$ as an estimator of $(\boldsymbol{H}^{(kl)})^{\mathrm{T}} \mathbf{x}_t$, is bounded over time. Define

$$\boldsymbol{\omega}_t^{(kl)} := \mathbf{y}_t^{(kl)} - (\boldsymbol{H}^{(kl)})^{\mathrm{T}} \boldsymbol{A} \boldsymbol{H}^{(kl)} \mathbf{y}_{t-1}^{(kl)} \quad \forall l \in \mathcal{V},\ k \in \mathcal{N}^{(l)}. \tag{43}$$

It is shown in Appendix D using (42) and Assumption A3 that

$$\sup_{t \geqslant 0} \mathbb{E}\Big\{ \big\| \boldsymbol{\omega}_t^{(kl)} \big\|^{a_{kl}} \Big\} < \infty \quad \forall l \in \mathcal{V},\ k \in \mathcal{N}^{(l)}. \tag{44}$$

Substituting k and l in (43) by i and j, respectively, gives

$$\mathbf{y}_t^{(ij)} = \left(\boldsymbol{H}^{(ij)}\right)^{\mathrm{T}} \boldsymbol{A} \boldsymbol{H}^{(ij)} \mathbf{y}_{t-1}^{(ij)} + \boldsymbol{\omega}_t^{(ij)}.$$

This shows that the process $\{\mathbf{y}_t^{(ij)}\}_{t\geqslant 0}$ can be viewed as a linear system affected by disturbance $\{\boldsymbol{\omega}_t^{(ij)}\}_{t\geqslant 0}$ that satisfies (44). We aim to design an encoder at node i for generating transmitted messages based on this linear system as well as an estimator at node j for inferring this linear system using received messages such that (41) holds. This problem has been studied in our previous work [69] and Proposition 1 in that work is used for the design of encoder and estimator at nodes i and j, respectively. The proposition in [69] is summarized in the next paragraph.

Consider a linear system $\{\boldsymbol{\theta}_t\}_{t\geqslant 0}$ that satisfies

$$\boldsymbol{\theta}_t = \boldsymbol{F}\boldsymbol{\theta}_{t-1} + \boldsymbol{\eta}_t \tag{45}$$

with the disturbance $\{\boldsymbol{\eta}_t\}_{t\geqslant 0}$ satisfying $\sup_{t\geqslant 0} \mathbb{E}\{\|\boldsymbol{\eta}_t\|^a\} < \infty$ for some $a > 2$. Moreover, suppose the magnitudes of all the eigenvalues of real matrix $\boldsymbol{F}$ is no smaller than 1. Consider a transmitter that computes a message $\mathbf{m}_t$ via real-time encoding based on $\boldsymbol{\theta}_{0:t}$ and transmits this message over a memoryless channel to a receiver at each time t. Let $\mathbf{r}_t$ represent the received message at time t. The receiver computes an estimator $\hat{\boldsymbol{\theta}}_t$ of $\boldsymbol{\theta}_t$ based on messages $\mathbf{r}_{0:t}$ received up to time t. The following proposition shows a sufficient condition under which there exist an encoder and estimator at the transmitter and receiver, respectively, such that the bth moment of the estimation error's Euclidean norm is bounded for any $2 \leqslant b < a$.

Proposition 3 ([69]): Consider the system given by (45). For any $2 \leqslant b < a$, if there is an α such that $\alpha > \frac{ab}{a-b} \max_{\lambda\in\Lambda(\boldsymbol{F})} \log|\lambda|$ and the α-anytime capacity $\breve{C}(\alpha)$ of the channel satisfies $\breve{C}(\alpha) > \log\left(|\det(\boldsymbol{F})|\right)$, then there exist an encoder that generates a transmitted message $\mathbf{m}_t$ based on $\boldsymbol{\theta}_{0:t}$ at each time t and an estimator $\hat{\boldsymbol{\theta}}_t$ of $\boldsymbol{\theta}_t$ based on received messages $\mathbf{r}_{0:t}$ such that

$$\sup_{t\geqslant 0} \mathbb{E}\left\{\left\|\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_t\right\|^b\right\} < \infty. \tag{46}$$

This encoder and estimator are denoted by $\mathrm{Enc}(\boldsymbol{\theta}_t, \boldsymbol{F}, a, b)$ and $\mathrm{Est}(\boldsymbol{\theta}_t, \boldsymbol{F}, a, b)$, respectively.

Proof: The encoder and estimator that achieve (46) employ adaptive quantization [51] as well as anytime encoding and decoding [55] techniques. See [69] for the proof of this proposition as well as the design of $\mathrm{Enc}(\boldsymbol{\theta}_t, \boldsymbol{F}, a, b)$ and $\mathrm{Est}(\boldsymbol{\theta}_t, \boldsymbol{F}, a, b)$. □

Inequality (41) can then be proved by applying Proposition 3. Specifically, set $\boldsymbol{\theta}_t = \mathbf{y}_t^{(ij)}$, $\boldsymbol{F} = \left(\boldsymbol{H}^{(ij)}\right)^{\mathrm{T}} \boldsymbol{A} \boldsymbol{H}^{(ij)}$, $\boldsymbol{\eta}_t = \boldsymbol{\omega}_t^{(ij)}$, $a = a_{ij}$, $b = a_{ij}/2$, $\alpha = \alpha^{(j)}$, and $\breve{C}(\alpha) = \breve{C}^{(ij)}(\alpha^{(j)})$. Moreover, it is shown in Appendix D that (34) and (35) imply the following

$$\alpha^{(j)} > a_{ij} \max_{\lambda\in\Lambda((\boldsymbol{H}^{(ij)})^{\mathrm{T}} \boldsymbol{A} \boldsymbol{H}^{(ij)})} \log|\lambda| \tag{47a}$$

$$\breve{C}^{(ij)}(\alpha^{(j)}) > \log\left(\left|\det\left(\left(\boldsymbol{H}^{(ij)}\right)^{\mathrm{T}} \boldsymbol{A} \boldsymbol{H}^{(ij)}\right)\right|\right) \tag{47b}$$

which is the condition required for applying Proposition 3. According to this proposition, there exist an encoder $\mathrm{Enc}\left(\mathbf{y}_t^{(ij)}, \left(\boldsymbol{H}^{(ij)}\right)^{\mathrm{T}} \boldsymbol{A} \boldsymbol{H}^{(ij)}, a_{ij}, a_{ij}/2\right)$ at node i and an estimator $\mathrm{Est}\left(\mathbf{y}_t^{(ij)}, \left(\boldsymbol{H}^{(ij)}\right)^{\mathrm{T}} \boldsymbol{A} \boldsymbol{H}^{(ij)}, a_{ij}, a_{ij}/2\right)$ at node j, respectively, such that (41) holds. In particular, the encoder generates $\mathbf{m}_t^{(ij)}$ at time t using $\mathbf{y}_{0:t}^{(ij)}$, whereas the estimator generates $\hat{\mathbf{y}}_t^{(ij)}$ at time t using received messages $\mathbf{r}_{0:t}^{(ij)}$.

Finally, the desired result (5) is shown. Combining (33), (40), (25), (41), and (42), we show in Appendix D that

$$\sup_{t\geqslant 0} \mathbb{E}\left\{\left\|\hat{\mathbf{x}}_t^{(j)} - \mathbf{x}_t^{(j)}\right\|^2\right\} < \infty. \tag{48}$$

Note that $\varepsilon_t^{(j)} \leqslant \mathbb{E}\left\{\left\|\hat{\mathbf{x}}_t^{(j)} - \mathbf{x}_t^{(j)}\right\|^2\right\}$ since $\varepsilon_t^{(j)}$ is the minimum MSE over all distributed estimators for $\mathbf{x}_t^{(j)}$. Therefore, (6) holds, which is equivalent to (5).

E. Discussion

The tightness of the established sufficient condition can be seen by comparing it with a necessary condition established in a companion paper [56]. Specifically, the necessary condition consists of a subcondition on the sensing capability and a subcondition on the communication capability of the network. The sensing subcondition of the sufficient condition is the same as that of the necessary condition. Moreover, there is a gap in general between the communication subcondition of the sufficient condition and that of the necessary condition. In [56], a favorable property of such a gap in certain scenarios is shown: the thresholds determining the capacity region in the communication subcondition of the sufficient condition are extreme points of the capacity region in the necessary condition.

The designed distributed encoder and estimator are communication efficient from the following aspects. First, different from many existing methods for distributed learning and inference where nodes perform multiple rounds of communications at each time step, the designed encoding strategy does not require that each pair of nodes perform more than one round of communication at each time step, and thus the communication overhead is significantly reduced. Second, different from existing works that assume real vectors and matrices can be transmitted among nodes without any loss, this paper takes into account the constraint of each communication channel and designs encoding strategies for the transmission of informative messages via these channels. Third, the established sufficient condition can be used for the efficient allocation of communication resources in the network, which is important for the design of multiple access schemes. Specifically, given a network such that Subcondition (i) in Theorem 1 holds, we determine the amount of resource allocated to each communication channel so that (34) holds. Then Theorem 1 shows that distributed encoders and estimators can be designed to ensure that the total distributed inference MSE is bounded over time.

This paper addressed the boundedness of MMSE over time in distributed learning over networks. On the other hand, computing the MMSE or deriving a tight bound for it is a challenging problem. This is because the uncertainty in both sensing and communication needs to be considered and

the optimal real-time encoding and estimation need to be designed. Moreover, the derivation of error bound typically requires more assumptions on the probability distributions for system disturbance, observation noise, and communication channels than those adopted in this paper. Even though a tight error bound is difficult to obtain, some insights on the reduction of such error can be obtained from this paper. For example, Subcondition (ii) of Theorem 1 provides guidelines for efficient allocation of communication resources in the network, which is important for reducing the inference error. In particular, this subcondition shows that more resources are required for the message transmission to agents whose states are more sensitive to disturbance as well as to agents with less sensing capability and thus are more reliant on received messages for estimating their states.

VI. Case Studies

This section first explains the derived sufficient condition in an example network with three nodes. Then, the section presents numerical results when the designed distributed encoders and estimators are applied to NLN.

A. Sufficient Condition in a Network With Three Nodes

Consider a network of three nodes 1, 2, and 3, where node 1 is within the communication ranges of both nodes 2 and 3, as shown in Fig. 6. In other words, $\mathcal{V} = \{1,2,3\}$ and $\mathcal{E}_{\mathrm{u}} = \{(1,2),(1,3)\}$. Consider the case where node 1 is the only agent, i.e., $\mathcal{V}_{\mathrm{a}} = \{1\}$. The unknown state of each node is a real vector with three entries, i.e., $\mathbf{x}_t^{(j)} \in \mathbb{R}^3$ for $j \in \{1,2,3\}$. Consequently, the concatenated state $\mathbf{x}_t = \left[\left(\mathbf{x}_t^{(1)}\right)^{\mathrm{T}} \quad \left(\mathbf{x}_t^{(2)}\right)^{\mathrm{T}} \quad \left(\mathbf{x}_t^{(3)}\right)^{\mathrm{T}}\right]^{\mathrm{T}}$ is a 9-dimensional vector, and the subspace $\mathcal{X}^{(1)}$ corresponding to agent 1's unknown state is $\mathcal{X}^{(1)} = \mathcal{C}([\boldsymbol{e}_{1,9} \quad \boldsymbol{e}_{2,9} \quad \boldsymbol{e}_{3,9}])$. Matrices that determine the evolution of unknown states are given by $\boldsymbol{A}^{(j)} = \mathrm{diag}\{2,2,3\}$ for $j \in \{1,2,3\}$.

The sensor gain matrices for the three nodes are given by

$$\boldsymbol{\Gamma}^{(11)} = \mathbf{0}, \quad \boldsymbol{\Gamma}_1^{(12)} = \boldsymbol{\Gamma}_2^{(12)} = \mathbf{0}, \quad \boldsymbol{\Gamma}_1^{(13)} = \boldsymbol{\Gamma}_2^{(13)} = \mathbf{0} \tag{49a}$$

$$\boldsymbol{\Gamma}^{(22)} = \boldsymbol{I}, \quad \boldsymbol{\Gamma}_1^{(21)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad \boldsymbol{\Gamma}_2^{(21)} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \end{bmatrix} \tag{49b}$$

$$\boldsymbol{\Gamma}^{(33)} = \boldsymbol{I}, \quad \boldsymbol{\Gamma}_1^{(31)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \boldsymbol{\Gamma}_2^{(31)} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix}. \tag{49c}$$

Equation (49a) shows that both intra- and inter-node observations of node 1 contain only noise and do not provide useful information. As a result, node 1 must rely on messages received from nodes 2 and 3 for learning its state. In particular, these messages contain useful information extracted from observations obtained by nodes 2 and 3.

For this network, Subcondition (i) in Theorem 1 can be seen to hold. To check Subcondition (ii), consider an ordered tree $\mathcal{T}_1$ in which node 1 is the root, whereas nodes 2 and 3 are the first and second child of node 1, respectively. In other words, $\check{c}_1^{(1)} = 2$ and $\check{c}_2^{(1)} = 3$, where we recall that $\check{c}_n^{(1)}$ represents the nth child of node 1 for $n = 1,2$. Applying (63a) with $i = 2$ and $j = 1$ as well as noting that $\mathcal{V}_2(\{1\}) = \{2\}$ and $\mathcal{S}(\{1\}) = \{\mathbf{0}\}$ gives $\mathcal{G}_2^{(21)} = \mathcal{C}([\boldsymbol{e}_{1,9} \quad \boldsymbol{e}_{2,9}])$ and $\mathcal{G}_3^{(21)} = \{\mathbf{0}\}$. Substituting these into (34) and noting that $\mathcal{I}_{\boldsymbol{A}^{\mathrm{T}}}\left(\mathcal{G}_2^{(21)}\right) = \mathcal{G}_2^{(21)}$, $\mathcal{I}_{\boldsymbol{A}^{\mathrm{T}}}\left(\mathcal{G}_3^{(21)}\right) = \mathcal{G}_3^{(21)}$ gives $\gamma_{\mathcal{T}_1}^{(21)} = 2$. Similarly, applying (63b) with $i = 3$ gives $\mathcal{G}_2^{(31)} = \{\mathbf{0}\}$ and $\mathcal{G}_3^{(31)} = \mathcal{C}(\boldsymbol{e}_{3,9})$, and thus $\gamma_{\mathcal{T}_1}^{(31)} = \log 3$. Consequently, (34) becomes $\breve{C}^{(21)}(\alpha^{(1)}) > 2$ and $\breve{C}^{(31)}(\alpha^{(1)}) > \log 3$ with $\alpha^{(1)} = 32 \log 3$.

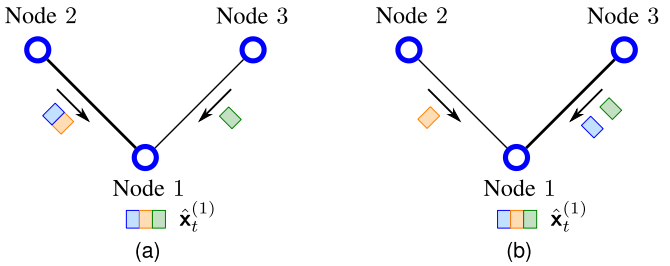


Fig. 6. Distributed encoding and estimation in a network of three nodes. (a): encoding and estimation based on $\mathcal{T}_1$. (b): encoding and estimation based on $\mathcal{T}_2$. Each line connecting a pair of nodes represents a communication channel to node 1 and its width indicates the threshold for the anytime capacity of this channel. The blue, orange, and green rectangles correspond to the first, second, and third entry of agent 1's state, respectively.

The designed distributed encoders and estimators based on $\mathcal{T}_1$ are explained in the following. Substituting $\mathcal{G}_2^{(21)} = \mathcal{C}([\boldsymbol{e}_{1,9} \quad \boldsymbol{e}_{2,9}])$ and $\mathcal{G}_3^{(21)} = \{\mathbf{0}\}$ into (30) gives $\mathcal{H}^{(21)} = \mathcal{C}([\boldsymbol{e}_{1,9} \quad \boldsymbol{e}_{2,9}])$. According to the encoding strategy presented in Section V-B, node 2 computes an estimator of $[\boldsymbol{e}_{1,9} \quad \boldsymbol{e}_{2,9}]^{\mathrm{T}}\mathbf{x}_t = \left[\left[\mathbf{x}_t^{(1)}\right]_1 \quad \left[\mathbf{x}_t^{(1)}\right]_2\right]^{\mathrm{T}}$ at each time t. Node 2 then generates encoded messages according to such an estimator, and transmits these messages to agent 1. Using the messages received from node 2, agent 1 estimates the first two entries $\left[\left[\mathbf{x}_t^{(1)}\right]_1 \quad \left[\mathbf{x}_t^{(1)}\right]_2\right]^{\mathrm{T}}$ of its unknown state. If the anytime capacity of the channel from node 2 to agent 1 satisfies $\breve{C}^{(21)}(\alpha^{(1)}) > 2$, then the error of the estimator obtained by agent 1 for these two entries is bounded over time. Analogously, $\mathcal{H}^{(31)} = \mathcal{C}(\boldsymbol{e}_{3,9})$, and agent 1 estimates $\boldsymbol{e}_{3,9}^{\mathrm{T}}\mathbf{x}_t = \left[\mathbf{x}_t^{(1)}\right]_3$, namely the third entry of its unknown state using messages received from node 3. If $\breve{C}^{(31)}(\alpha^{(1)}) > \log 3$, then the error of this estimator is also bounded over time. Finally, agent 1 linearly combines the estimators of the three entries of its unknown state to obtain a distributed estimator $\hat{\mathbf{x}}_t^{(1)}$ of $\mathbf{x}_t^{(1)}$. This is illustrated in Fig. (6a).

Consider another ordered tree $\mathcal{T}_2$ based on $\mathscr{G}_{\mathrm{u}}$ in which node 1 is the root, whereas nodes 3 and 2 are the first and second child of node 1, respectively. In other words, $\check{c}_1^{(1)} = 3$ and $\check{c}_2^{(1)} = 2$. Applying (63) gives $\mathcal{G}_2^{(31)} = \mathcal{C}(\boldsymbol{e}_{1,9})$, $\mathcal{G}_3^{(31)} = \mathcal{C}(\boldsymbol{e}_{3,9})$, $\mathcal{G}_2^{(21)} = \mathcal{C}(\boldsymbol{e}_{2,9})$, and $\mathcal{G}_3^{(21)} = \{\mathbf{0}\}$. Substituting these into (34) and (30) gives $\gamma_{\mathcal{T}_2}^{(21)} = 1$, $\gamma_{\mathcal{T}_2}^{(31)} = 1 + \log 3$, $\mathcal{H}^{(21)} = \mathcal{C}(\boldsymbol{e}_{2,9})$, and $\mathcal{H}^{(31)} = \mathcal{C}([\boldsymbol{e}_{1,9} \quad \boldsymbol{e}_{3,9}])$. If the designed encoders and estimators based on $\mathcal{T}_2$ are employed, then node 2 computes an estimator of $\boldsymbol{e}_{2,9}^{\mathrm{T}}\mathbf{x}_t = \left[\mathbf{x}_t^{(1)}\right]_2$ at each time t, generates encoded messages according to such an estimator, and transmits these messages to agent 1. Using the messages received from node 2, agent 1 estimates the second

entry $\left[\mathbf{x}_t^{(1)}\right]_2$ of its unknown state. Analogously, agent 1 estimates $[\boldsymbol{e}_{1,9} \quad \boldsymbol{e}_{3,9}]^{\mathrm{T}}\mathbf{x}_t = \left[\left[\mathbf{x}_t^{(1)}\right]_1 \quad \left[\mathbf{x}_t^{(1)}\right]_3\right]^{\mathrm{T}}$, namely the first and third entries of its unknown state using messages received from node 3. If the anytime capacities of the channels from node 2 and node 3 to agent 1 satisfy $\breve{C}^{(21)}(\alpha^{(1)}) > 1$ and $\breve{C}^{(31)}(\alpha^{(1)}) > 1 + \log 3$, respectively, then the errors of the estimators obtained by agent 1 for these three entries are all bounded over time. Finally, agent 1 linearly combines these estimators to obtain a distributed estimator $\hat{\mathbf{x}}_t^{(1)}$ of $\mathbf{x}_t^{(1)}$. This is illustrated in Fig. (6b).

Note that the thresholds for the anytime capacities of the two channels to agent 1 based on $\mathcal{T}_2$ are different from that based on $\mathcal{T}_1$. Specifically, the threshold $\gamma_{\mathcal{T}_2}^{(21)}$ for the channel from node 2 to agent 1 based on $\mathcal{T}_2$ is smaller than the threshold $\gamma_{\mathcal{T}_1}^{(21)}$ based on $\mathcal{T}_1$. To see this, note that the dimension of $\mathcal{H}^{(21)}$ based on $\mathcal{T}_2$ and $\mathcal{T}_1$ are 1 and 2, respectively. If $\mathcal{T}_2$ is employed for designing the distributed encoders and estimators, then node 2 transmits to agent 1 the information of an estimator for $\boldsymbol{e}_{2,9}^{\mathrm{T}}\mathbf{x}_t$, which is a vector with only one entry. By contrast, if $\mathcal{T}_1$ is employed, then node 2 needs to transmit to agent 1 the information of an estimator for $[\boldsymbol{e}_{1,9} \quad \boldsymbol{e}_{2,9}]^{\mathrm{T}}\mathbf{x}_t$, which is a vector with two entries. Consequently, node 2 needs to transmit less information to agent 1 if $\mathcal{T}_2$ is employed, and thus the threshold for the anytime capacity of the channel from node 2 to agent 1 is smaller. Similar arguments can be used for comparing thresholds $\gamma_{\mathcal{T}_2}^{(31)}$ and $\gamma_{\mathcal{T}_1}^{(31)}$.

In addition to $\mathcal{T}_1$ and $\mathcal{T}_2$, there are two other ordered trees $\mathcal{T}_3$ and $\mathcal{T}_4$ based on $\mathscr{G}_{\mathrm{u}}$, where the roots are node 2 and node 3, respectively. Similar calculations show that inequalities (34) based on $\mathcal{T}_3$ and $\mathcal{T}_4$ are the same as those based on $\mathcal{T}_2$ and $\mathcal{T}_1$, respectively. Combining the thresholds for anytime capacities corresponding to all the four ordered trees, Subcondition (ii) of Theorem 1 can be expressed as follows: the vector of anytime capacity $\left[\breve{C}^{(21)}\left(\alpha^{(1)}\right) \quad \breve{C}^{(31)}\left(\alpha^{(1)}\right)\right]^{\mathrm{T}}$ belongs to the region $\breve{\mathcal{R}}_{\mathcal{T}_1}^{(1)} \cup \breve{\mathcal{R}}_{\mathcal{T}_2}^{(1)}$, where

$$\breve{\mathcal{R}}_{\mathcal{T}_1}^{(1)} := \left\{ \left[c^{(21)} \quad c^{(31)}\right]^{\mathrm{T}} : c^{(21)} > 2,\, c^{(31)} > \log 3 \right\}$$
$$\breve{\mathcal{R}}_{\mathcal{T}_2}^{(1)} := \left\{ \left[c^{(21)} \quad c^{(31)}\right]^{\mathrm{T}} : c^{(21)} > 1,\, c^{(31)} > 1 + \log 3 \right\}.$$

Consequently, Theorem 1 states that if $\left[\breve{C}^{(21)}\left(\alpha^{(1)}\right) \; \breve{C}^{(31)}\left(\alpha^{(1)}\right)\right]^{\mathrm{T}} \in \breve{\mathcal{R}}_{\mathcal{T}_1}^{(1)} \cup \breve{\mathcal{R}}_{\mathcal{T}_2}^{(1)}$, then every node can construct an encoder for generating messages to each neighbor such that $\sup_{t \geqslant 0} F_t < \infty$.

B. Results for Network Localization and Navigation

This section evaluates via simulation the performance of the distributed encoder and estimator designed in Section V for NLN, an example application of distributed learning. The nodes in NLN are categorized as either anchors or agents, where anchors are static nodes and agents are mobile nodes. The aim of NLN is to infer in real time the positions of agents via sensing and communication.

In NLN, the state of a node consists of its position and other position-related quantities. Specifically, $\mathbf{x}_t^{(j)} := \left[\left(\mathbf{p}_t^{(j)}\right)^{\mathrm{T}} \quad \left(\mathbf{v}_t^{(j)}\right)^{\mathrm{T}}\right]^{\mathrm{T}}$, where $\mathbf{p}_t^{(j)} \in \mathbb{R}^3$ and $\mathbf{v}_t^{(j)} \in \mathbb{R}^3$ represent the 3-D position and velocity, respectively, of node j at time t. The evolution of $\mathbf{x}_t^{(j)}$ follows (1) with $\boldsymbol{A}^{(j)}$ given by $\boldsymbol{A}^{(j)} = \left[\begin{smallmatrix} \boldsymbol{I}_3 & \Delta \boldsymbol{I}_3 \\ \boldsymbol{0} & \boldsymbol{I}_3 \end{smallmatrix}\right]$, where Δ is the duration of a time step and is set to 0.2 second. Moreover, the disturbance $\boldsymbol{\zeta}_t^{(j)}$ is given by $\boldsymbol{\zeta}_t^{(j)} = \boldsymbol{B}\tilde{\boldsymbol{\zeta}}_t^{(j)}$, where $\boldsymbol{B} := \left[\frac{1}{2}(\Delta)^2 \boldsymbol{I}_3 \quad \Delta \boldsymbol{I}_3\right]^{\mathrm{T}}$ and $\tilde{\boldsymbol{\zeta}}_t^{(j)} \in \mathbb{R}^3$ is white noise. This is the discrete white noise acceleration model [70, Chapter 6.3.2] that has been widely used. In this section, $\tilde{\boldsymbol{\zeta}}_t^{(j)} = \boldsymbol{0}$ if node j is an anchor so that it remains static, otherwise $\tilde{\boldsymbol{\zeta}}_t^{(j)}$ follows Gaussian distribution with covariance matrix $0.04(\mathrm{m}^2/\mathrm{s}^4) \times \boldsymbol{I}_3$.

The considered intra- and inter-node observations are as follows. An intra-node observation $\mathbf{z}_t^{(jj)}$ obtained by node j is a measurement of its position at time t and is given by $\mathbf{z}_t^{(jj)} = \mathbf{p}_t^{(j)} + \mathbf{n}_t^{(jj)}$, where $\mathbf{n}_t^{(jj)}$ represents the observation noise. Such an observation can be obtained by using a global navigation satellite system receiver. An inter-node observation $\mathbf{z}_t^{(ji)}$ is a measurement of the displacement of node j with respect to node i at time t and is given by $\mathbf{z}_t^{(ji)} = \mathbf{p}_t^{(j)} - \mathbf{p}_t^{(i)} + \mathbf{n}_t^{(ji)}$, where $\mathbf{n}_t^{(jj)}$ represents the observation noise. Such an observation can be obtained by measuring the distance and relative angle between node j and i. In this section, both $\mathbf{n}_t^{(jj)}$ and $\mathbf{n}_t^{(jj)}$ follow Gaussian distribution with covariance matrix $0.04\,\mathrm{m}^2 \times \boldsymbol{I}_3$.

The network considered in this section consists of four anchors and multiple agents. The positions of the anchors are $[86.7 \quad 50 \quad 0]^{\mathrm{T}}\,\mathrm{m}$, $[-86.7 \quad 50 \quad 0]^{\mathrm{T}}\,\mathrm{m}$, $[0 \quad -100 \quad 0]^{\mathrm{T}}\,\mathrm{m}$, and $[0 \quad 0 \quad 100]^{\mathrm{T}}\,\mathrm{m}$, respectively, and the velocity of all the anchors are $\boldsymbol{0}\,\mathrm{m/s}$. The position of each agent at time 0 is randomly sampled from $[-75\,\mathrm{m}, 75\,\mathrm{m}]^3$, and the velocity of each agent at time 0 is randomly sampled from $[-0.5\,\mathrm{m/s}, 0.5\,\mathrm{m/s}]^3$.

Each anchor in the network is within the communication range of three randomly chosen agents, and each agent is in the communication range of three other randomly chosen agents. Each anchor obtains an intra-node observation at every time step. On the other hand, each agent does not obtain any intra-node observation. Instead, it obtains an inter-node observation with each neighbor at every time step. Furthermore, each node is connected with every neighbor via a noiseless digital channel, through which the node can transmit data under a rate constraint. For this type of channel, the data rate constraint coincides with both the Shannon capacity and the anytime capacity of the channel.

This section evaluates the average position MSE over all the agents using the designed distributed encoder and estimator. Specifically, the vector $\hat{\mathbf{p}}_t^{(j)}$ consisting of the first three entries of the designed estimator $\hat{\mathbf{x}}_t^{(j)}$ (see Section V) is employed as the distributed estimator of $\mathbf{p}_t^{(j)}$. The position MSE of node j's distributed estimator at time t is $\mathbb{E}\left\{\left\|\hat{\mathbf{p}}_t^{(j)} - \mathbf{p}_t^{(j)}\right\|^2\right\}$, and the average position MSE $\bar{\varepsilon}_t$ of distributed estimators over all the agents at time step t is given by

$$\bar{\varepsilon}_t := \frac{1}{|\mathcal{V}_{\mathrm{a}}|} \sum_{j \in \mathcal{V}_{\mathrm{a}}} \mathbb{E}\left\{\left\|\hat{\mathbf{p}}_t^{(j)} - \mathbf{p}_t^{(j)}\right\|^2\right\}. \tag{50}$$

The empirical value of the above average position MSE is obtained via Monte Carlo simulations.

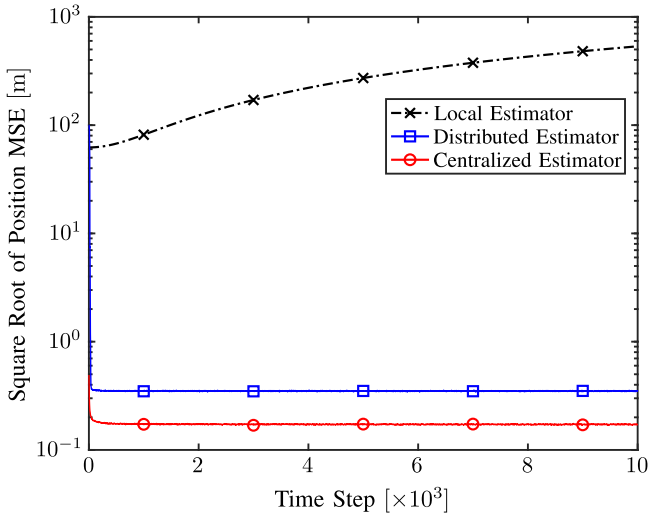


Fig. 7. Square roots of the average position MSEs at each time step for the distributed and centralized estimators.

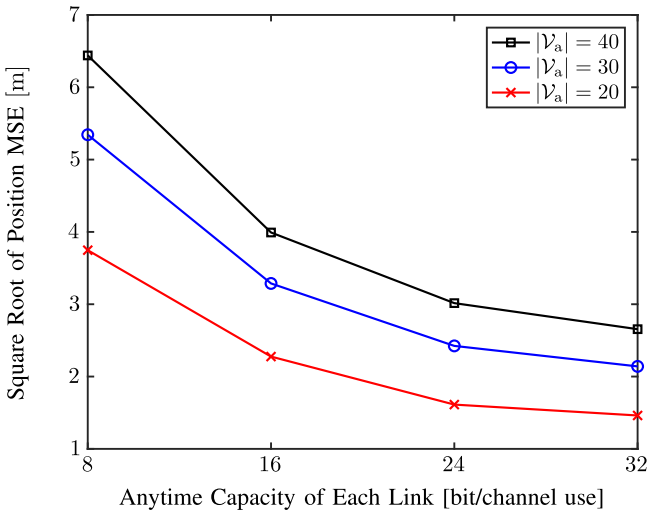


Fig. 8. Square roots of the average position MSEs over all time steps for the distributed estimators under different network parameters settings.

Recall that an ordered tree $\mathcal{T}$ based on the graph corresponding to the network is used for the design of distributed encoder and estimator. To construct $\mathcal{T}$ in the simulation, a random weight is first assigned to each edge of the graph and a minimum spanning tree of the resulting weighted undirected graph is constructed. In particular, the mean value of the weight for an edge between an anchor and an agent is set to be smaller than that for an edge between two agents. The purpose of this is to maximize the number of edges between anchors and agents in the spanning tree. Finally, an order of each node's children is specified to obtain the ordered tree $\mathcal{T}$.

First, the accuracy of the designed distributed estimators at each time step is evaluated and compared with two reference estimators: the local estimator and the centralized estimator. Specifically, the local estimator of agent j's position $\mathbf{p}_t^{(j)}$ is the MMSE estimator of $\mathbf{p}_t^{(j)}$ only using agent j's observations $\mathring{\mathbf{z}}_{0:t}^{(j)}$. This corresponds to the case where no communication among nodes is performed. The centralized estimator of $\mathbf{p}_t^{(j)}$ is the MMSE estimator of $\mathbf{p}_t^{(j)}$ using observations $\left\{\mathring{\mathbf{z}}_{0:t}^{(k)} : k \in \mathcal{V}\right\}$ obtained by all the nodes in the network. This corresponds to the case that each node can transmit all its observations ideally to a centralized processor so that this processor can estimate the positions of all the agents using observations obtained by the entire network. Both the local and the centralized estimators are computed using the Kalman filtering technique [71], [72], [73]. The average position MSEs of the local estimator and the centralized estimator are evaluated in the simulation, which are defined by substituting $\hat{\mathbf{p}}_t^{(j)}$ in (50) with the local and centralized estimator of $\mathbf{p}_t^{(j)}$, respectively.

Figure 7 shows the square roots of average position MSEs at each time step of the designed distributed estimator, local estimator, and centralized estimator in the scenario where the number of agents $|\mathcal{V}_\mathrm{a}| = 10$ and the anytime capacity of each channel is 8 bits per channel use. It can be observed that the error of both the distributed and centralized estimators are below one meter over the entire time horizon. In particular, the centralized estimator achieves smaller average position MSE since it can access observations of the entire network via ideal communication. On the other hand, the average position MSE of the local estimator increases fast with time, showing that each agent alone does not have enough sensing capability for localizing itself. Specifically, it can be verified that $\mathcal{X}^{(j)} \not\subseteq \mathcal{S}(\{j\})$ for all $j \in \mathcal{V}_\mathrm{a}$, namely the subspace corresponding to $\mathbf{x}_t^{(j)}$ is not contained in the observable subspace $\mathcal{S}(\{j\})$. As a result, agent j needs not only its local observations but also the messages received from other nodes in order to achieve bounded position MSE.

Next, the accuracy of the designed distributed estimators is evaluated under different network parameter settings. Figure 8 shows the square roots of the average position MSEs for different anytime capacities of each link and for different numbers of agents in the network. It can be observed that the average position MSE decreases with the anytime capacity of each link. This is because each node can transmit more data per channel use when the anytime capacity of the channel becomes higher, and thus the receiver node can extract more position information from its received messages.

Another observation from Fig. 8 is that the average position MSE increases with the number of agents in the network. This is because as the number of agents increases, more agents will have larger shortest-path distances to anchors in graph $\mathscr{G}_\mathrm{u}$, as each anchor is within the communication range of only three agents in the simulation. Since the position MSE of an agent increases with its shortest-path distance to an anchor, the average position MSE increases with the number of agents in the network. To see how the shortest-path distance to an anchor affects the position MSE of an agent, we compare the position estimation accuracy of an anchor with that of an agent being a neighbor of the anchor. Specifically, the anchor can estimate its position using its intra-node observations, which are measurements of the anchor's position. By contrast, the agent does not obtain intra-node observations. As a result, it estimates its position by combining the inter-node observations, which are measurements of the agent's displacement with respect to the

anchor, with the messages received from the anchor, which contains the anchor's position information. Due to the data rate constraint on the channel, the messages received by the agent contain less information than that available to the anchor, and thus the position MSE of the agent, whose shortest-path distance to the anchor is one, is larger than that of the anchor. Analogously, the position MSE of an agent with its shortest-path distance to an anchor being two is larger than that of an agent whose shortest-path distance to an anchor is one, and so on.

VII. Conclusion

This paper established a theoretical framework for communication-efficient distributed learning of time-varying states in complex networked systems without the help of central processors. The paper derived a sufficient condition under which the total distributed inference MSE of all the agent nodes are bounded over time. The sufficient condition consists of a subcondition on the network's sensing capabilities and one on the network's communication capabilities. In particular, the subcondition on the communication capabilities states that the anytime capacities of the communication channels be above certain thresholds, which are determined by the eigenvalues of the agents' dynamic matrices and the sensing capabilities of nodes in the network. The paper also developed real-time encoding strategies to achieve bounded total distributed inference MSE when the derived sufficient condition is satisfied. The paper provides guidelines for the design of communication-efficient distributed learning algorithms in complex networked systems. The results in this paper show the connection among learning, information, and control theories.

Appendix A
Proof of Lemma 1

Proof: Let $\boldsymbol{T}$ be a matrix such that its columns form an orthonormal basis of $\mathcal{C}\big(\boldsymbol{O}\big(\mathring{\boldsymbol{I}}^{(j)},\boldsymbol{A}\big)^{\mathrm{T}}\big)$. Consequently, there is a matrix $\boldsymbol{H}$ such that $\boldsymbol{H}^{\mathrm{T}}\boldsymbol{T}^{\mathrm{T}}=\boldsymbol{O}\big(\mathring{\boldsymbol{I}}^{(j)},\boldsymbol{A}\big)$. Define $\boldsymbol{\theta}_t:=\boldsymbol{T}^{\mathrm{T}}\mathbf{x}_t$. Combining (14) and (22) with the fact that $\mathcal{C}(\boldsymbol{T})$ is $\boldsymbol{A}^{\mathrm{T}}$-invariant, we obtain

$$\begin{aligned}\boldsymbol{\theta}_t&=\boldsymbol{T}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{T}\boldsymbol{\theta}_{t-1}+\boldsymbol{T}^{\mathrm{T}}\boldsymbol{\zeta}_t\\ \mathring{\mathbf{z}}_t^{(j)}&=\mathring{\boldsymbol{I}}^{(j)}\boldsymbol{T}\boldsymbol{\theta}_t+\mathring{\mathbf{n}}_t^{(j)}.\end{aligned}$$

Moreover, the pair $\big(\mathring{\boldsymbol{I}}^{(j)}\boldsymbol{T},\boldsymbol{T}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{T}\big)$ is observable, that is, matrix $\boldsymbol{M}$ defined as

$$\boldsymbol{M}:=\boldsymbol{O}\big(\mathring{\boldsymbol{I}}^{(j)}\boldsymbol{T},\boldsymbol{T}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{T}\big)^{\mathrm{T}}\,\boldsymbol{O}\big(\mathring{\boldsymbol{I}}^{(j)}\boldsymbol{T},\boldsymbol{T}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{T}\big)$$

is invertible. Define a linear estimator $\hat{\boldsymbol{\theta}}_t$ of $\boldsymbol{\theta}_t$ based on $\mathring{\mathbf{z}}_{0:t}^{(j)}$ as

$$\hat{\boldsymbol{\theta}}_t:=\big(\boldsymbol{T}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{T}\big)^{r-1}\boldsymbol{M}^{-1}\sum_{l=1}^{r}\big(\boldsymbol{T}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{T}\big)^{l}\big(\mathring{\boldsymbol{I}}^{(j)}\boldsymbol{T}\big)^{\mathrm{T}}\mathring{\mathbf{z}}_{t+l-r+1}^{(j)}$$

where r represents the rank of $\boldsymbol{O}\big(\mathring{\boldsymbol{I}}^{(j)}\boldsymbol{T},\boldsymbol{T}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{T}\big)$. Moreover, define $\hat{\boldsymbol{\xi}}_t^{(j)}:=\boldsymbol{H}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_t$. It can be verified that $\sup_{t\geqslant 0}\mathbb{E}\big\{\big\|\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t\big\|^a\big\}<\infty$. Combining this with $\big\|\hat{\boldsymbol{\xi}}_t^{(j)}-\boldsymbol{O}\big(\mathring{\boldsymbol{I}}^{(j)},\boldsymbol{A}\big)\mathbf{x}_t\big\|=\big\|\boldsymbol{H}^{\mathrm{T}}(\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t)\big\|\leqslant\big\|\boldsymbol{H}^{\mathrm{T}}\big\|\,\big\|\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t\big\|$, the desired result is obtained. □

Appendix B
Proof of Proposition 2

First, a lemma to be used is stated.

Lemma 2: For any three subspaces $\mathcal{Y}$, $\mathcal{U}$, and $\mathcal{W}$ of a real vector space, there exists a subspace denoted by $\mathcal{P}(\mathcal{Y},\mathcal{U},\mathcal{W})$ that satisfies the following relations.

$$(\mathcal{P}(\mathcal{Y},\mathcal{U},\mathcal{W})+\mathcal{U})\cap\mathcal{W}=(\mathcal{Y}+\mathcal{U})\cap\mathcal{W} \tag{51a}$$

$$\begin{aligned}\dim(\mathcal{P}(\mathcal{Y},\mathcal{U},\mathcal{W}))&=\dim\big((\mathcal{Y}+\mathcal{U})\cap\mathcal{W}\big)\\&\quad-\dim(\mathcal{U}\cap\mathcal{W})\end{aligned} \tag{51b}$$

$$\mathcal{P}(\mathcal{Y},\mathcal{U},\mathcal{W})\subseteq\mathcal{Y}\cap(\mathcal{U}+\mathcal{W})\,. \tag{51c}$$

Moreover, for any subspace $\tilde{\mathcal{Y}}$ that satisfies $(\tilde{\mathcal{Y}}+\mathcal{U})\cap\mathcal{W}=(\mathcal{Y}+\mathcal{U})\cap\mathcal{W}$, it holds that

$$\dim(\tilde{\mathcal{Y}})\geqslant\dim\big((\mathcal{Y}+\mathcal{U})\cap\mathcal{W}\big)-\dim(\mathcal{U}\cap\mathcal{W})\,. \tag{52}$$

Proof: Subspace $\mathcal{P}(\mathcal{Y},\mathcal{U},\mathcal{W})$ is constructed as follows. Let $\mathcal{V}$ be a subspace that satisfies

$$\begin{aligned}\mathcal{V}+(\mathcal{U}\cap\mathcal{W})&=(\mathcal{Y}+\mathcal{U})\cap\mathcal{W}\\ \dim(\mathcal{V})&=\dim\big((\mathcal{Y}+\mathcal{U})\cap\mathcal{W}\big)-\dim\big(\mathcal{U}\cap\mathcal{W}\big)\,.\end{aligned} \tag{53}$$

An example of $\mathcal{V}$ is the orthogonal complement of $\mathcal{U}\cap\mathcal{W}$ with respect to $(\mathcal{Y}+\mathcal{U})\cap\mathcal{W}$. Since $\mathcal{V}\subseteq\mathcal{Y}+\mathcal{U}$, there exist subspaces $\mathcal{Y}_0$ and $\mathcal{U}_0$ such that

$$\mathcal{Y}_0\subseteq\mathcal{Y},\quad\mathcal{U}_0\subseteq\mathcal{U},\ \mathcal{V}=\mathcal{Y}_0+\mathcal{U}_0\,. \tag{54}$$

Set subspace $\mathcal{P}(\mathcal{Y},\mathcal{U},\mathcal{W})$ to $\mathcal{P}(\mathcal{Y},\mathcal{U},\mathcal{W})=\mathcal{Y}_0$.

Equation (51) with $\mathcal{P}(\mathcal{Y},\mathcal{U},\mathcal{W})$ replaced by $\mathcal{Y}_0$ is shown next. First, (51a) is proved. On one hand, $\mathcal{Y}_0\subseteq\mathcal{Y}$ gives $(\mathcal{Y}_0+\mathcal{U})\cap\mathcal{W}\subseteq(\mathcal{Y}+\mathcal{U})\cap\mathcal{W}$. On the other hand, since $\mathcal{U}_0\subseteq\mathcal{U}$, we have $\mathcal{Y}_0+\mathcal{U}=\mathcal{Y}_0+\mathcal{U}_0+\mathcal{U}=\mathcal{V}+\mathcal{U}$, where (54) is used in the second equality. Therefore,

$$\begin{aligned}(\mathcal{Y}_0+\mathcal{U})\cap\mathcal{W}&=(\mathcal{V}+\mathcal{U})\cap\mathcal{W}\\&\supseteq(\mathcal{V}\cap\mathcal{W})+(\mathcal{U}\cap\mathcal{W})\\&=\mathcal{V}+(\mathcal{U}\cap\mathcal{W})=(\mathcal{Y}+\mathcal{U})\cap\mathcal{W}\end{aligned} \tag{55}$$

where the last but one equality is because $\mathcal{V}\subseteq\mathcal{W}$ as indicated by (53). Therefore, (51a) holds.

Next, (51b) is proved. On one hand, $\dim(\mathcal{Y}_0)\leqslant\dim(\mathcal{V})=\dim\big((\mathcal{Y}+\mathcal{U})\cap\mathcal{W}\big)-\dim\big(\mathcal{U}\cap\mathcal{W}\big)$ since $\mathcal{Y}_0\subseteq\mathcal{V}$ according to (54). On the other hand, to show that

$$\dim(\mathcal{Y}_0)\geqslant\dim\big((\mathcal{Y}+\mathcal{U})\cap\mathcal{W}\big)-\dim\big(\mathcal{U}\cap\mathcal{W}\big) \tag{56}$$

we only need to prove (52) for any $\tilde{\mathcal{Y}}$ satisfying $(\tilde{\mathcal{Y}}+\mathcal{U})\cap\mathcal{W}=(\mathcal{Y}+\mathcal{U})\cap\mathcal{W}$. Applying the equality $\dim(\mathcal{S}_1)+\dim(\mathcal{S}_2)=\dim(\mathcal{S}_1+\mathcal{S}_2)+\dim(\mathcal{S}_1\cap\mathcal{S}_2)$ for general subspaces $\mathcal{S}_1$ and $\mathcal{S}_2$, we obtain

$$\begin{aligned}&\dim(\tilde{\mathcal{Y}})+\dim(\mathcal{U}\cap\mathcal{W})\\&\quad=\dim(\tilde{\mathcal{Y}}+\mathcal{U}+\mathcal{W})+\dim\big((\tilde{\mathcal{Y}}+\mathcal{U})\cap\mathcal{W}\big)\\&\qquad+\dim(\tilde{\mathcal{Y}}\cap\mathcal{U})-\dim(\mathcal{U}+\mathcal{W})\\&\quad\geqslant\dim\big((\tilde{\mathcal{Y}}+\mathcal{U})\cap\mathcal{W}\big)=\dim\big((\mathcal{Y}+\mathcal{U})\cap\mathcal{W}\big)\,.\end{aligned} \tag{57}$$

This shows (52). Moreover, $\mathcal{Y}_0$ satisfies (55), and thus (56) holds. Therefore, (51b) is proved.

Finally, (51c) is proved. According to (54), $\mathcal{Y}_0\subseteq\mathcal{Y}$. To show $\mathcal{Y}_0\subseteq\mathcal{U}+\mathcal{W}$, recall that (51b) has been proved.

In other words, equality is achieved in (57) if $\tilde{\mathcal{Y}}$ is replaced by $\mathcal{Y}_0$ therein. This is true only if $\dim(\mathcal{Y}_0+\mathcal{U}+\mathcal{W}) = \dim(\mathcal{U}+\mathcal{W})$, which indicates $\mathcal{Y}_0 \subseteq \mathcal{U}+\mathcal{W}$, and thus (51c) is proved. □

Subspace $\mathcal{P}(\mathcal{Y},\mathcal{U},\mathcal{W})$ can be viewed as a generalization of a complementary subspace. In particular, setting $\mathcal{W}=\mathbb{R}^n$, then (51a) and (51b) become $\mathcal{P}(\mathcal{Y},\mathcal{U},\mathcal{W})+\mathcal{U}=\mathcal{Y}+\mathcal{U}$ and $\dim(\mathcal{P}(\mathcal{Y},\mathcal{U},\mathcal{W})) = \dim\big(\mathcal{Y}+\mathcal{U}\big)-\dim(\mathcal{U})$, respectively. This shows that $\mathcal{P}(\mathcal{Y},\mathcal{U},\mathcal{W})$ and $\mathcal{U}$ are complementary if they are both viewed as subspaces of $\mathcal{Y}+\mathcal{U}$.

Next, Proposition 2 is proved.

Proof: We construct $\mathcal{G}_\lambda^{(ij)} \subseteq \mathcal{M}_\lambda(\boldsymbol{A}^{\mathrm{T}})$ that satisfies the following properties

$$\mathcal{X}^{(j)} \subseteq \mathcal{S}(\{j\}) + \sum_{i\in\mathcal{N}^{(j)}} \sum_{\lambda\in\Lambda(\boldsymbol{A}^{(j)})} \mathcal{G}_\lambda^{(ij)} \quad \forall j\in\mathcal{V}_{\mathrm{a}} \tag{58a}$$

$$\sum_{\lambda\in\Lambda(\boldsymbol{A}^{(j)})} \mathcal{G}_\lambda^{(ij)} \subseteq \mathcal{S}(\{i\}) + \sum_{k\in\mathcal{N}^{(i)}\setminus\{j\}} \sum_{\lambda\in\Lambda(\boldsymbol{A}^{(i)})} \mathcal{G}_\lambda^{(ki)} \quad \forall j\in\mathcal{V},\, i\in\mathcal{N}^{(j)} \tag{58b}$$

$$\sum_{\lambda\in\Lambda(\boldsymbol{A}^{(j)})} \mathcal{G}_\lambda^{(ij)} \subseteq \mathcal{X}^{(i)} + \mathcal{X}^{(j)} \;\; \forall j\in\mathcal{V},\, i\in\mathcal{N}^{(j)}. \tag{58c}$$

Moreover, the dimensionality of $\mathcal{G}_\lambda^{(ij)}$ satisfies

$$\sum_{i\in\mathcal{N}_n^{(j)}} \dim(\mathcal{G}_\lambda^{(ij)}) = r_\lambda^{(j)}\big(\mathcal{N}_n^{(j)}\big) \quad \forall j\in\mathcal{V}_{\mathrm{a}},\, n\in\big\{1,2,\ldots,N^{(j)}\big\}. \tag{59}$$

Here, $\mathcal{N}_n^{(j)}$ is defined as

$$\mathcal{N}_n^{(j)} := \big\{\check{c}_n^{(j)}, \check{c}_{n+1}^{(j)}, \ldots, \check{c}_{N^{(j)}}^{(j)}\big\} \tag{60}$$

where we recall $\check{c}_{n'}^{(j)}$ represents either a child or the parent of node j in ordered tree $\mathcal{T}$ for $n' \in \big\{1,2,\ldots,N^{(j)}\big\}$ (see Section IV-B), and $r_\lambda^{(j)}(\cdot)$ in (59) is defined as

$$r_\lambda^{(j)}\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big) := \dim\big(\mathcal{X}^{(j)}\cap\mathcal{M}_\lambda(\boldsymbol{A}^{\mathrm{T}})\big) - \dim\big(\mathcal{S}\big(\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\big)\cap\mathcal{X}^{(j)}\cap\mathcal{M}_\lambda(\boldsymbol{A}^{\mathrm{T}})\big). \tag{61}$$

Note that subspaces $\mathcal{H}^{(ij)}$ defined as the right-hand side of (30) will then satisfy (29) since (58) implies (29). Here, we show how (58a) implies (29a). Applying $\mathcal{I}_{\boldsymbol{A}^{\mathrm{T}}}$ at both sides of (58a) and using (8), we obtain

$$\mathcal{I}_{\boldsymbol{A}^{\mathrm{T}}}\big(\mathcal{X}^{(j)}\big) \subseteq \mathcal{I}_{\boldsymbol{A}^{\mathrm{T}}}\big(\mathcal{S}(\{j\})\big) + \sum_{i\in\mathcal{N}^{(j)}} \sum_{\lambda\in\Lambda(\boldsymbol{A}^{(j)})} \mathcal{I}_{\boldsymbol{A}^{\mathrm{T}}}\big(\mathcal{G}_\lambda^{(ij)}\big) \tag{62}$$

Since $\boldsymbol{A}^{\mathrm{T}}$ is block-diagonal as shown in (13c), it can be verified that $\mathcal{I}_{\boldsymbol{A}^{\mathrm{T}}}\big(\mathcal{X}^{(j)}\big) = \mathcal{X}^{(j)}$. Moreover, $\mathcal{I}_{\boldsymbol{A}^{\mathrm{T}}}\big(\mathcal{S}(\{j\})\big) = \mathcal{S}(\{j\})$ as indicated by (27). Substituting these two equalities and (30) into (62), we obtain (29a). Similarly, applying $\mathcal{I}_{\boldsymbol{A}^{\mathrm{T}}}$ at both sides of (58b), we obtain (29b). Finally, if (58c) holds, then

$$\begin{aligned}\mathcal{H}^{(ij)} &= \mathcal{I}_{\boldsymbol{A}^{\mathrm{T}}}\Big(\sum_{\lambda\in\Lambda(\boldsymbol{A}^{(j)})} \mathcal{G}_\lambda^{(ij)}\Big)\\ &\subseteq \mathcal{I}_{\boldsymbol{A}^{\mathrm{T}}}\big(\mathcal{X}^{(i)}+\mathcal{X}^{(j)}\big)\\ &= \mathcal{I}_{\boldsymbol{A}^{\mathrm{T}}}\big(\mathcal{X}^{(i)}\big)+\mathcal{I}_{\boldsymbol{A}^{\mathrm{T}}}\big(\mathcal{X}^{(j)}\big) = \mathcal{X}^{(i)}+\mathcal{X}^{(j)}.\end{aligned}$$

The construction of $\mathcal{G}_\lambda^{(ij)}$ is presented as follows. For the case where node i is a child of node j, subspace $\mathcal{G}_\lambda^{(ij)}$ is given by as in (63), shown at the bottom of the page, where $\mathcal{P}(\cdot)$ represents a subspace introduced in Lemma 2.

For the case where node i is the parent of node j, subspace $\mathcal{G}_\lambda^{(ij)}$ is defined as in (64), shown at the bottom of the page. Note that (64) is a recursive definition: for a node i that is not the root of the ordered tree $\mathcal{T}$, constructing $\mathcal{G}_\lambda^{(ij)}$ requires $\mathcal{G}_\lambda^{(\check{c}_{N^{(i)}}^{(i)}\, i)}$, where $\check{c}_{N^{(i)}}^{(i)}$ is the parent of node i (recall Section V) and is an element of $\mathcal{N}^{(i)}\setminus\{j\}$. Therefore, $\mathcal{G}_\lambda^{(ij)}$ should be constructed first for the case where node i is the root of $\mathcal{T}$, namely the depth of i is zero.[3] Then $\mathcal{G}_\lambda^{(ij)}$ are constructed for nodes i with depth one in $\mathcal{T}$. Such a procedure is repeated until $\mathcal{G}_\lambda^{(ij)}$ have been constructed for all nodes i that have at least one child.

A proof that $\mathcal{G}_\lambda^{(ij)}$ constructed in (63) and (64) satisfy (58) and (59) can be found in [74, Appendix B.2.2], which is omitted here due to limitation of space. Recall that (58) implies (29), and thus Proposition 2 is proved. □

[3]The depth of a node is the length of the simple path from the node to the root of the tree [68, Appendix B].

$$\mathcal{G}_\lambda^{(ij)} := \mathcal{P}\Big(\mathcal{S}\big(\mathcal{V}_i(\{j\})\big)\cap\mathcal{M}_\lambda(\boldsymbol{A}^{\mathrm{T}}),\, \mathcal{S}(\{j\})\cap\mathcal{M}_\lambda(\boldsymbol{A}^{\mathrm{T}}),\, \mathcal{X}^{(j)}\cap\mathcal{M}_\lambda(\boldsymbol{A}^{\mathrm{T}})\Big) \quad \text{if } i=\check{c}_1^{(j)} \text{ and } i \text{ is a child of } j \tag{63a}$$

$$\mathcal{G}_\lambda^{(ij)} := \mathcal{P}\Big(\mathcal{S}\big(\mathcal{V}_i(\{j\})\big)\cap\mathcal{M}_\lambda(\boldsymbol{A}^{\mathrm{T}}),\, \mathcal{S}(\{j\})\cap\mathcal{M}_\lambda(\boldsymbol{A}^{\mathrm{T}}) + \sum_{k=1}^{n-1}\mathcal{G}_\lambda^{(\check{c}_k^{(j)}\, j)},\, \mathcal{X}^{(j)}\cap\mathcal{M}_\lambda(\boldsymbol{A}^{\mathrm{T}})\Big) \quad \text{if } i=\check{c}_n^{(j)} \text{ for } n>1 \text{ and } i \text{ is a child of } j. \tag{63b}$$

$$\mathcal{G}_\lambda^{(ij)} := \mathcal{P}\Big(\mathcal{S}(\{i\})\cap\mathcal{M}_\lambda(\boldsymbol{A}^{\mathrm{T}}) + \sum_{k\in\mathcal{N}^{(i)}\setminus\{j\}}\mathcal{G}_\lambda^{(ki)},\, \mathcal{S}(\{j\})\cap\mathcal{M}_\lambda(\boldsymbol{A}^{\mathrm{T}}) + \sum_{k\in\mathcal{N}^{(j)}\setminus\{i\}}\mathcal{G}_\lambda^{(kj)},\, \mathcal{X}^{(j)}\cap\mathcal{M}_\lambda(\boldsymbol{A}^{\mathrm{T}})\Big) \quad \text{if } i \text{ is the parent of } j. \tag{64}$$

APPENDIX C
LEMMAS FOR PROVING THEOREM 1

A few lemmas to be used for proving Theorem 1 are presented here. The first lemma shows a property of direct sum.

Lemma 3: For subspaces $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_m$ and subspaces $\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_m$ such that $\mathcal{S}_i \subseteq \mathcal{U}_i$ for all $i \in \{1, 2, \ldots, m\}$ and $\sum_{i=1}^m \mathcal{U}_i = \oplus_{i=1}^m \mathcal{U}_i$, it holds that $\left(\sum_{i=1}^m \mathcal{S}_i\right) \cap \mathcal{U}_j = \mathcal{S}_j$ for all $j \in \{1, 2, \ldots, m\}$.

Proof: On one hand, $\left(\sum_{i=1}^m \mathcal{S}_i\right) \cap \mathcal{U}_j \supseteq \sum_{i=1}^m (\mathcal{S}_i \cap \mathcal{U}_j) = \mathcal{S}_j$. On the other hand, consider an arbitrary $\boldsymbol{u} \in \left(\sum_{i=1}^m \mathcal{S}_i\right) \cap \mathcal{U}_j$, which can be written as $\boldsymbol{u} = \sum_{i=1}^m \boldsymbol{u}_i$ with $\boldsymbol{u}_i \in \mathcal{S}_i \subseteq \mathcal{U}_i$. It can be shown that $\boldsymbol{u} - \boldsymbol{u}_j \in \left(\sum_{i \neq j} \mathcal{U}_i\right) \cap \mathcal{U}_j$, and thus $\boldsymbol{u} - \boldsymbol{u}_j = \boldsymbol{0}$ by the property of direct sum. This shows that $\boldsymbol{u} \in \mathcal{S}_j$, and thus $\left(\sum_{i=1}^m \mathcal{S}_i\right) \cap \mathcal{U}_j \subseteq \mathcal{S}_j$. □

The next lemma shows the real Jordan canonical form of a real matrix.

Lemma 4: [75] For any real square matrix $\boldsymbol{F}$ with $\Lambda(\boldsymbol{F}) = \{\lambda_1, \lambda_2, \ldots, \lambda_m\}$, there exists a real matrix $\boldsymbol{M}$ that transforms $\boldsymbol{F}$ to its real Jordan canonical form $\boldsymbol{J}$ as

$$\boldsymbol{F} = \boldsymbol{M}\boldsymbol{J}\boldsymbol{M}^{-1} \tag{65}$$

where $\boldsymbol{J} = \operatorname{diag}\{\boldsymbol{J}_1, \boldsymbol{J}_2, \ldots, \boldsymbol{J}_m\}$. In particular, $\boldsymbol{J}_k$ is a real square matrix with $\mathcal{M}_{\lambda_k}(\boldsymbol{F})$ columns. In particular, if $\lambda_k \in \mathbb{R}$, then all the eigenvalues of $\boldsymbol{J}_k$ are λ_k. Otherwise eigenvalues of $\boldsymbol{J}_k$ are either λ_k or λ_k^*. In addition, $\boldsymbol{M}$ can be partitioned as $\boldsymbol{M} = [\boldsymbol{M}_1 \quad \boldsymbol{M}_2 \quad \cdots \quad \boldsymbol{M}_m]$ such that $\boldsymbol{M}_k$ has full column rank and forms a basis of $\mathcal{M}_{\lambda_k}(\boldsymbol{F})$ for $k = 1, 2, \ldots, m$.

The next lemma is on eigenvalues and determinants related to invariant subspaces.

Lemma 5: Let $\boldsymbol{F}$ be a real square matrix and let $\mathcal{Y}$ be an arbitrary $\boldsymbol{F}$-invariant subspace. Moreover, let $\boldsymbol{Y}$ be a real matrix whose columns form an orthonormal basis of $\mathcal{Y}$. Then the following equalities hold

$$\max_{\lambda \in \Lambda(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{F}\boldsymbol{Y})} |\lambda| = \max_{\lambda \in \Lambda(\boldsymbol{F})} |\lambda| \, \mathbb{1}_{[1,\infty)}\big(y(\lambda)\big) \tag{66a}$$

$$\left|\det(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{F}\boldsymbol{Y})\right| = \prod_{\lambda \in \Lambda(\boldsymbol{F})} |\lambda|^{y(\lambda)} \tag{66b}$$

where $y(\lambda)$ is defined as

$$y(\lambda) := \dim\big(\mathcal{Y} \cap \mathcal{M}_\lambda(\boldsymbol{F})\big).$$

Moreover, $\mathbb{1}_{[1,\infty)}\big(y(\lambda)\big) = 1$ if $y(\lambda) \geqslant 1$ and $\mathbb{1}_{[1,\infty)}\big(y(\lambda)\big) = 0$ otherwise.

Proof: See the proof for Lemma 12 in [74, Appendix B.1]. □

APPENDIX D
PROOF OF THEOREM 1

Proof: This appendix presents the expression of a_{ij}, shows that Subcondition (ii) of Theorem 1 guarantees (47), and proves (44), (48), as well as (42).

First, for any $j \in \mathcal{V}$ and $i \in \mathcal{N}^{(j)}$, scalar a_{ij} is defined as

$$a_{ij} = \begin{cases} 2^{D+2-h^{(i)}} & \text{if } j \text{ is the parent of } i \\ 2^{D+2-\tilde{d}_{\mathcal{T}}^{(ij)}} & \text{if } j \text{ is a child of } i. \end{cases} \tag{67}$$

Here, D represents the diameter of the ordered tree $\mathcal{T}$. Moreover, $h^{(i)}$ represents the height of node i in $\mathcal{T}$, i.e., the number of edges on the longest downward path from node i to a leaf node [68, Appendix B]. Variable $\tilde{d}_{\mathcal{T}}^{(ij)}$ represents the largest of the shortest-path distance on $\mathcal{T}$ between node i and any other node $k \in \mathcal{V}_i(\{j\})$, i.e., $\tilde{d}_{\mathcal{T}}^{(ij)} := \max_{k \in \mathcal{V}_i(\{j\})} d_{\mathcal{T}}(k, i)$, where $d_{\mathcal{T}}(k, i)$ represents the shortest-path distance between nodes k and i in $\mathcal{T}$, and $\mathcal{V}_i(\cdot)$ is defined in (21). Equation (67) indicates that $4 \leqslant a_{ij} \leqslant 2^{D+2}$, since $0 \leqslant h^{(i)} \leqslant D$ and $0 \leqslant \tilde{d}_{\mathcal{T}}^{(ij)} \leqslant D$.

Next, it is shown that Subcondition (ii) of Theorem 1 guarantees (47). To this end, using the fact that taking transpose of a matrix does not change its eigenvalues, and applying (66a) in Lemma 5 with $\boldsymbol{F} = \boldsymbol{A}^{\mathrm{T}}$ as well as $\mathcal{Y} = \mathcal{H}^{(ij)}$,

$$\begin{aligned} &\max_{\lambda \in \Lambda((\boldsymbol{H}^{(ij)})^{\mathrm{T}}\boldsymbol{A}\boldsymbol{H}^{(ij)})} |\lambda| \\ &\quad = \max_{\lambda \in \Lambda((\boldsymbol{H}^{(ij)})^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{H}^{(ij)})} |\lambda| \\ &\quad = \max_{\lambda \in \Lambda(\boldsymbol{A}^{\mathrm{T}})} |\lambda| \, \mathbb{1}_{[1,\infty)}\Big(\dim\big(\mathcal{H}^{(ij)} \cap \mathcal{M}_\lambda(\boldsymbol{A}^{\mathrm{T}})\big)\Big) \\ &\quad \leqslant \max_{\lambda \in \Lambda(\boldsymbol{A}^{(j)})} |\lambda| \end{aligned} \tag{68}$$

where (68) is because the following equality

$$\mathcal{H}^{(ij)} \cap \mathcal{M}_\lambda(\boldsymbol{A}^{\mathrm{T}}) = \begin{cases} \mathcal{I}_{\boldsymbol{A}^{\mathrm{T}}}(\mathcal{G}_\lambda^{(ij)}) & \text{if } \lambda \in \Lambda(\boldsymbol{A}^{(j)}) \\ \{\boldsymbol{0}\} & \text{otherwise} \end{cases} \tag{69}$$

which is obtained by applying Lemma 3. Combining (68) and $a_{ij} \leqslant 2^{D+2} < 2^{D+3}$ with (35) gives (47a).

Similarly, applying (66b) in Lemma 5 with $\boldsymbol{F} = \boldsymbol{A}^{\mathrm{T}}$ as well as $\mathcal{Y} = \mathcal{H}^{(ij)}$, and then taking the logarithm gives

$$\begin{aligned} &\log\Big(\Big|\det\Big(\big(\boldsymbol{H}^{(ij)}\big)^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{H}^{(ij)}\Big)\Big|\Big) \\ &\quad = \sum_{\lambda \in \Lambda(\boldsymbol{A}^{\mathrm{T}})} \dim\Big(\mathcal{H}^{(ij)} \cap \mathcal{M}_\lambda(\boldsymbol{A}^{\mathrm{T}})\Big) \log|\lambda|. \end{aligned} \tag{70}$$

Combining (34), (69), as well as (70), and using the fact that taking transpose of a matrix does not change its determinant, we obtain (47b).

Next, (44) is proved. Replacing i in (39) by l and combining the result with (43) gives

$$\begin{aligned} \boldsymbol{\omega}_t^{(kl)} = \mathbf{y}_t^{(kl)} &- \big(\boldsymbol{H}^{(kl)}\big)^{\mathrm{T}}\mathbf{x}_t + \big(\boldsymbol{H}^{(kl)}\big)^{\mathrm{T}}\boldsymbol{\zeta}_t \\ &+ \big(\boldsymbol{H}^{(kl)}\big)^{\mathrm{T}}\boldsymbol{A}\boldsymbol{H}^{(kl)}\Big(\big(\boldsymbol{H}^{(kl)}\big)^{\mathrm{T}}\mathbf{x}_{t-1} - \mathbf{y}_{t-1}^{(kl)}\Big). \end{aligned}$$

Taking the Euclidean norm of the above equality, raising the result to the power of a_{kl}, and then applying triangle inequality as well as the general inequality

$$\Big(\sum\nolimits_{l=1}^{L} y_l\Big)^x \leqslant L^x \sum\nolimits_{l=1}^{L} y_l^x \tag{71}$$

for an arbitrary positive integer L as well as positive real numbers y_l and x, we obtain as in (72), shown at the top of the next page. Taking expectation for both sides of (72) and using (42) as well as Assumption A3 in Section II, we obtain (44).

Next, (48) is proved. Subtracting (33) from (40), evaluating the squared Euclidean norm of the result, and

$$\left\|\boldsymbol{\omega}_t^{(kl)}\right\|^{a_{kl}} \leqslant 3^{a_{kl}}\left(\left\|\mathbf{y}_t^{(kl)}-\left(\boldsymbol{H}^{(kl)}\right)^{\mathrm{T}}\mathbf{x}_t\right\|^{a_{kl}}+\left\|\left(\boldsymbol{H}^{(kl)}\right)^{\mathrm{T}}\right\|^{a_{kl}}\left\|\boldsymbol{\zeta}_t\right\|^{a_{kl}}+\left\|\left(\boldsymbol{H}^{(kl)}\right)^{\mathrm{T}}\boldsymbol{A}\boldsymbol{H}^{(kl)}\right\|^{a_{kl}}\left\|\mathbf{y}_{t-1}^{(kl)}-\left(\boldsymbol{H}^{(kl)}\right)^{\mathrm{T}}\mathbf{x}_{t-1}\right\|^{a_{kl}}\right) \tag{72}$$

$$\begin{aligned}\mathbb{E}\left\{\left\|\hat{\mathbf{x}}_t^{(j)}-\mathbf{x}_t^{(j)}\right\|^2\right\} \leqslant \left(2N^{(j)}+1\right)^2\Bigg[& \left\|\left(\boldsymbol{\Phi}^{(jj)}\right)^{\mathrm{T}}\right\|^2\mathbb{E}\left\{\left\|\hat{\boldsymbol{\xi}}_t^{(j)}-\boldsymbol{O}\left(\mathring{\boldsymbol{I}}^{(j)},\boldsymbol{A}\right)\mathbf{x}_t\right\|^2\right\} \\ & +\sum_{i\in\mathcal{N}^{(j)}}\left\|\left(\boldsymbol{\Phi}^{(ij)}\right)^{\mathrm{T}}\right\|^2\left(\mathbb{E}\left\{\left\|\hat{\mathbf{y}}_t^{(ij)}-\mathbf{y}_t^{(ij)}\right\|^2\right\}+\mathbb{E}\left\{\left\|\mathbf{y}_t^{(ij)}-\left(\boldsymbol{H}^{(ij)}\right)^{\mathrm{T}}\mathbf{x}_t\right\|^2\right\}\right)\Bigg]\end{aligned} \tag{73}$$

applying triangle inequality,

$$\begin{aligned}\left\|\hat{\mathbf{x}}_t^{(j)}-\mathbf{x}_t^{(j)}\right\|^2 \leqslant \Bigg[& \left\|\left(\boldsymbol{\Phi}^{(jj)}\right)^{\mathrm{T}}\left(\hat{\boldsymbol{\xi}}_t^{(j)}-\boldsymbol{O}\left(\mathring{\boldsymbol{I}}^{(j)},\boldsymbol{A}\right)\mathbf{x}_t\right)\right\| \\ & +\sum_{i\in\mathcal{N}^{(j)}}\Bigg(\left\|\left(\boldsymbol{\Phi}^{(ij)}\right)^{\mathrm{T}}\left(\hat{\mathbf{y}}_t^{(ij)}-\mathbf{y}_t^{(ij)}\right)\right\| \\ & +\left\|\left(\boldsymbol{\Phi}^{(ij)}\right)^{\mathrm{T}}\left(\mathbf{y}_t^{(ij)}-\left(\boldsymbol{H}^{(ij)}\right)^{\mathrm{T}}\mathbf{x}_t\right)\right\|\Bigg)\Bigg]^2.\end{aligned}$$

Applying Cauchy-Schwarz inequality and the definition of spectral norm of a matrix, then taking expectation, we obtain as in (73), shown at the top of the page. Since $a_{ij}\geqslant 4$, applying Jensen's inequality gives

$$\begin{aligned}\mathbb{E}\left\{\left\|\mathbf{y}_t^{(ij)}-\left(\boldsymbol{H}^{(ij)}\right)^{\mathrm{T}}\mathbf{x}_t\right\|^{a_{ij}}\right\} & \\ \geqslant \mathbb{E}\left\{\left\|\mathbf{y}_t^{(ij)}-\left(\boldsymbol{H}^{(ij)}\right)^{\mathrm{T}}\mathbf{x}_t\right\|^2\right\}^{a_{ij}/2} & \end{aligned}$$

and thus (42) shows $\sup_{t\geqslant 0}\mathbb{E}\left\{\left\|\mathbf{y}_t^{(ij)}-\left(\boldsymbol{H}^{(ij)}\right)^{\mathrm{T}}\mathbf{x}_t\right\|^2\right\}<\infty$. Similarly, (41) shows that $\sup_{t\geqslant 0}\mathbb{E}\left\{\left\|\hat{\mathbf{y}}_t^{(ij)}-\mathbf{y}_t^{(ij)}\right\|^2\right\}<\infty$. Combining these two inequalities and (25) with (73) gives the desired result (48).

Finally, we prove (42), which is used for proving (44) and (48). Specifically, we show the construction of $\mathbf{y}_t^{(kl)}$ for the case where node l is the parent of node k in the ordered tree $\mathcal{T}$ and prove (42) via induction. For the base case of the induction, node k is a leaf in $\mathcal{T}$, i.e., node k has only one neighbor, which is its parent l. In this case, define $\mathbf{y}_t^{(kl)}:=\left(\boldsymbol{\Psi}_k^{(kl)}\right)^{\mathrm{T}}\hat{\boldsymbol{\xi}}_t^{(k)}$, where $\boldsymbol{\Psi}_k^{(kl)}$ is a matrix such that $\boldsymbol{H}^{(kl)}=\boldsymbol{O}\left(\mathring{\boldsymbol{I}}^{(k)},\boldsymbol{A}\right)^{\mathrm{T}}\boldsymbol{\Psi}_k^{(kl)}$. Note that this is consistent with the general definition (38) since $\mathcal{N}^{(k)}=\{l\}$. Replacing j and a in (25) by k and a_{kl}, respectively, and using the inequality

$$\left\|\mathbf{y}_t^{(kl)}-\left(\boldsymbol{H}^{(kl)}\right)^{\mathrm{T}}\mathbf{x}_t\right\| \leqslant \left\|\left(\boldsymbol{\Psi}_k^{(kl)}\right)^{\mathrm{T}}\right\|\left\|\hat{\boldsymbol{\xi}}_t^{(k)}-\boldsymbol{O}\left(\mathring{\boldsymbol{I}}^{(k)},\boldsymbol{A}\right)\mathbf{x}_t\right\|$$

due to the definition of matrix spectral norm, we obtain (42) for the base case of induction.

For the induction step, assume that for each child k' of node k, random vector $\mathbf{y}_t^{(k'k)}$ has been constructed with

$$\sup_{t\geqslant 0}\mathbb{E}\left\{\left\|\mathbf{y}_t^{(k'k)}-\left(\boldsymbol{H}^{(k'k)}\right)^{\mathrm{T}}\mathbf{x}_t\right\|^{a_{k'k}}\right\}<\infty. \tag{74}$$

Define $\boldsymbol{\omega}_t^{(k'k)}$ as (43). Using the induction hypothesis (74), we can show that $\sup_{t\geqslant 0}\mathbb{E}\left\{\left\|\boldsymbol{\omega}_t^{(k'k)}\right\|^{a_{k'k}}\right\}<\infty$. Therefore, Proposition 3 can be applied to show that there exist an encoder $\mathrm{Enc}\left(\mathbf{y}_t^{(k'k)},\left(\boldsymbol{H}^{(k'k)}\right)^{\mathrm{T}}\boldsymbol{A}\boldsymbol{H}^{(k'k)},a_{k'k},a_{k'k}/2\right)$ at node k' and an estimator $\hat{\mathbf{y}}_t^{(k'k)}$ of $\mathbf{y}_t^{(k'k)}$ at node k, respectively, such that $\sup_{t\geqslant 0}\mathbb{E}\left\{\left\|\hat{\mathbf{y}}_t^{(k'k)}-\mathbf{y}_t^{(k'k)}\right\|^{a_{k'k}/2}\right\}<\infty$. Define $\mathbf{y}_t^{(kl)}$ as (38). Applying triangle inequality and the general inequality (71), we can show (42) (see [74, Appendix B.2.3] for details), thus completing the induction. For the case where node l is a child of node k in the ordered tree $\mathcal{T}$, (42) can be proved similarly. This completes the proof. □

REFERENCES

[1] A. Conti et al., "Location awareness in beyond 5G networks," *IEEE Commun. Mag.*, vol. 59, no. 11, pp. 22–27, Nov. 2021.

[2] M. Z. Win et al., "Network localization and navigation via cooperation," *IEEE Commun. Mag.*, vol. 49, no. 5, pp. 56–62, May 2011.

[3] N. Patwari, J. N. Ash, S. Kyperountas, A. O. Hero, R. L. Moses, and N. S. Correal, "Locating the nodes: Cooperative localization in wireless sensor networks," *IEEE Signal Process. Mag.*, vol. 22, no. 4, pp. 54–69, Jul. 2005.

[4] A. A. Saucan and M. Z. Win, "Information-seeking sensor selection for ocean-of-things," *IEEE Internet Things J.*, vol. 7, no. 10, pp. 10072–10088, May 2020.

[5] S. G. Nagarajan, P. Zhang, and I. Nevat, "Geo-spatial location estimation for Internet of Things (IoT) networks with one-way time-of-arrival via stochastic censoring," *IEEE Internet Things J.*, vol. 4, no. 1, pp. 205–214, Feb. 2017.

[6] L. Chen et al., "Robustness, security and privacy in location-based services for future IoT: A survey," *IEEE Access*, vol. 5, pp. 8956–8977, 2017.

[7] M. Z. Win, F. Meyer, Z. Liu, W. Dai, S. Bartoletti, and A. Conti, "Efficient multi-sensor localization for the Internet of Things," *IEEE Signal Process. Mag.*, vol. 35, no. 5, pp. 153–167, Sep. 2018.

[8] M. Chen et al., "Distributed learning in wireless networks: Recent progress and future challenges," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 12, pp. 3579–3605, Dec. 2021.

[9] J. B. Predd, S. R. Kulkarni, and H. V. Poor, "Distributed learning in wireless sensor networks," *IEEE Signal Process. Mag.*, vol. 23, no. 4, pp. 56–69, Jul. 2006.

[10] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 50–60, May 2020.

[11] K. Ozkara, N. Singh, D. Data, and S. Diggavi, "QuPeD: Quantized personalization via distillation with applications to federated learning," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2021, pp. 1–15.

[12] O. A. Hanna, Y. H. Ezzeldin, C. Fragouli, and S. Diggavi, "Quantization of distributed data for learning," *IEEE J. Sel. Areas Inf. Theory*, vol. 2, no. 3, pp. 987–1001, Sep. 2021.

[13] M. Chen, N. Shlezinger, H. V. Poor, Y. C. Eldar, and S. Cui, "Communication-efficient federated learning," *Proc. Nat. Acad. Sci. USA*, vol. 118, no. 17, pp. 1–8, 2021.

[14] T. Sery, N. Shlezinger, K. Cohen, and Y. Eldar, "Over-the-air federated learning from heterogeneous data," *IEEE Trans. Signal Process.*, vol. 69, pp. 3796–3811, 2021.

[15] S. Wan, J. Lu, P. Fan, Y. Shao, C. Peng, and K. B. Letaief, "Convergence analysis and system design for federated learning over wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 12, pp. 3622–3639, Dec. 2021.

[16] Z. Zhu, S. Wan, P. Fan, and K. B. Letaief, "Federated multiagent actor–critic learning for age sensitive mobile-edge computing," *IEEE Internet Things J.*, vol. 9, no. 2, pp. 1053–1067, Jan. 2022.

[17] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, "Edge intelligence: Paving the last mile of artificial intelligence with edge computing," *Proc. IEEE*, vol. 107, no. 8, pp. 1738–1762, Aug. 2019.

[18] M. Zhang, J. Chen, S. He, L. Yang, X. Gong, and J. Zhang, "Privacy-preserving database assisted spectrum access for industrial Internet of Things: A distributed learning approach," *IEEE Trans. Ind. Electron.*, vol. 67, no. 8, pp. 7094–7103, Aug. 2020.

[19] P. Di Lorenzo, P. Banelli, S. Barbarossa, and S. Sardellitti, "Distributed adaptive learning of graph signals," *IEEE Trans. Signal Process.*, vol. 65, no. 16, pp. 4193–4208, Aug. 2017.

[20] A. Conti, S. Mazuelas, S. Bartoletti, W. C. Lindsey, and M. Z. Win, "Soft information for localization-of-things," *Proc. IEEE*, vol. 107, no. 11, pp. 2240–2264, Sep. 2019.

[21] M. Z. Win, Y. Shen, and W. Dai, "A theoretical foundation of network localization and navigation," *Proc. IEEE*, vol. 106, no. 7, pp. 1136–1165, Jul. 2018.

[22] X. Wang, L. Gao, and S. Mao, "BiLoc: Bi-modal deep learning for indoor localization with commodity 5 GHz WiFi," *IEEE Access*, vol. 5, pp. 4209–4220, 2017.

[23] A. Conti, M. Guerra, D. Dardari, N. Decarli, and M. Z. Win, "Network experimentation for cooperative localization,' *IEEE J. Sel. Areas Commun.*, vol. 30, no. 2, pp. 467–475, Feb. 2012.

[24] M. Z. Win, W. Dai, Y. Shen, G. Chrisikos, and H. V. Poor, "Network operation strategies for efficient localization and navigation," *Proc. IEEE*, vol. 106, no. 7, pp. 1224–1254, Jul. 2018.

[25] S. Das and J. M. F. Moura, "Consensus+innovations distributed Kalman filter with optimized gains," *IEEE Trans. Signal Process.*, vol. 65, no. 2, pp. 467–481, Jan. 2017.

[26] S. Das and J. M. F. Moura, "Distributed Kalman filtering with dynamic observations consensus," *IEEE Trans. Signal Process.*, vol. 63, no. 17, pp. 4458–4473, Sep. 2015.

[27] F. Zabini and A. Conti, "Inhomogeneous Poisson sampling of finite-energy signals with uncertainties in $\mathbb{R}^d$," *IEEE Trans. Signal Process.*, vol. 64, no. 18, pp. 4679–4694, Sep. 2016.

[28] G. Battistelli, L. Chisci, N. Forti, G. Pelosi, and S. Selleri, "Distributed finite-element Kalman filter for field estimation," *IEEE Trans. Autom. Control*, vol. 62, no. 7, pp. 3309–3322, Jul. 2017.

[29] P. K. Varshney, *Distributed Detection and Data Fusion*. Cham, Switzerland: Springer, 2012.

[30] P. Sharma, A.-A. Saucan, D. J. Bucci, and P. K. Varshney, "Decentralized Gaussian filters for cooperative self-localization and multi-target tracking," *IEEE Trans. Signal Process.*, vol. 67, no. 22, pp. 5896–5911, Nov. 2019.

[31] T. Vercauteren and X. Wang, "Decentralized sigma-point information filters for target tracking in collaborative sensor networks," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 2997–3009, Aug. 2005.

[32] S. Marano, V. Matta, and P. Willett, "Asymptotic design of quantizers for decentralized MMSE estimation," *IEEE Trans. Signal Process.*, vol. 55, no. 11, pp. 5485–5496, Nov. 2007.

[33] S. Marano, V. Matta, and P. Willett, "Distributed estimation in large wireless sensor networks via a locally optimum approach," *IEEE Trans. Signal Process.*, vol. 56, no. 2, pp. 748–756, Feb. 2008.

[34] Z. Liu, W. Dai, and M. Z. Win, "Mercury: An infrastructure-free system for network localization and navigation," *IEEE Trans. Mobile Comput.*, vol. 17, no. 5, pp. 1119–1133, May 2018.

[35] R. Olfati-Saber, J. A. Fax, and R. M. Murray, "Consensus and cooperation in networked multi-agent systems," *Proc. IEEE*, vol. 95, no. 1, pp. 215–233, Jan. 2007.

[36] F. S. Cattivelli and A. H. Sayed, "Diffusion LMS strategies for distributed estimation," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1035–1048, Mar. 2010.

[37] R. Carli, A. Chiuso, L. Schenato, and S. Zampieri, "Distributed Kalman filtering based on consensus strategies," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 4, pp. 622–633, May 2008.

[38] S. Kar, J. M. F. Moura, and H. V. Poor, "Distributed linear parameter estimation: Asymptotically efficient adaptive strategies," *SIAM J. Control Optim.*, vol. 51, no. 3, pp. 2200–2229, 2013.

[39] U. A. Khan and J. M. F. Moura, "Distributing the Kalman filter for large-scale systems," *IEEE Trans. Signal Process.*, vol. 56, no. 10, pp. 4919–4935, Oct. 2008.

[40] Y. Shen, S. Mazuelas, and M. Z. Win, "Network navigation: Theory and interpretation," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 9, pp. 1823–1834, Oct. 2012.

[41] X. Cao and T. Başar, "Decentralized online convex optimization with feedback delays," *IEEE Trans. Autom. Control*, vol. 67, no. 6, pp. 2889–2904, Jun. 2022.

[42] X. Cao and T. Başar, "Decentralized multi-agent stochastic optimization with pairwise constraints and quantized communications," *IEEE Trans. Signal Process.*, vol. 68, pp. 3296–3311, 2020.

[43] J. Schalkwijk and T. Kailath, "A coding scheme for additive noise channels with feedback—I: No bandwidth constraint," *IEEE Trans. Inf. Theory*, vol. IT-12, no. 2, pp. 172–182, Apr. 1966.

[44] S. K. Mitter, "Control with limited information," *Eur. J. Control*, vol. 7, nos. 2–3, pp. 122–131, Jan. 2001.

[45] V. S. Borkar and S. K. Mitter, "LQG control with communication constraints," in *Communications, Computation, Control, and Signal Processing*, A. Paulraj, V. Roychowdhury, and C. D. Schaper, Eds. Boston, MA, USA: Springer, 1997, pp. 365–373.

[46] S. Tatikonda and S. Mitter, "Control under communication constraints," *IEEE Trans. Autom. Control*, vol. 49, no. 7, pp. 1056–1068, Jul. 2004.

[47] R. W. Brocket and D. Liberzon, "Quantized feedback stabilization of linear systems," *IEEE Trans. Autom. Control*, vol. 45, no. 7, pp. 1279–1289, Jul. 2000.

[48] D. Liberzon and J. P. Hespanha, "Stabilization of nonlinear systems with limited information feedback," *IEEE Trans. Autom. Control*, vol. 50, no. 6, pp. 910–915, Jun. 2005.

[49] N. C. Martins, M. A. Dahleh, and N. Elia, "Feedback stabilization of uncertain systems in the presence of a direct link," *IEEE Trans. Autom. Control*, vol. 51, no. 3, pp. 438–447, Mar. 2006.

[50] P. Minero, M. Franceschetti, S. Dey, and G. N. Nair, "Data rate theorem for stabilization over time-varying feedback channels," *IEEE Trans. Autom. Control*, vol. 54, no. 2, pp. 243–255, Feb. 2009.

[51] G. N. Nair and R. J. Evans, "Stabilizability of stochastic linear systems with finite feedback data rates," *SIAM J. Control Optim.*, vol. 43, no. 2, pp. 413–436, Jul. 2004.

[52] G. N. Nair, F. Fagnani, S. Zampieri, and R. J. Evans, "Feedback control under data rate constraints: An overview," *Proc. IEEE*, vol. 95, no. 1, pp. 108–137, Jan. 2007.

[53] P. Tallapragada and J. Cortés, "Event-triggered stabilization of linear systems under bounded bit rates," *IEEE Trans. Autom. Control*, vol. 61, no. 6, pp. 1575–1589, Jun. 2016.

[54] S. Yüksel and T. Başar, *Stochastic Networked Control Systems: Stabilization and Optimization under Information Constraints*. New York, NY, USA: Springer, 2013.

[55] A. Sahai and S. Mitter, "The necessity and sufficiency of anytime capacity for stabilization of a linear system over a noisy communication link—Part I: Scalar systems," *IEEE Trans. Inf. Theory*, vol. 52, no. 8, pp. 3369–3395, Aug. 2006.

[56] Z. Liu, A. Conti, S. K. Mitter, and M. Z. Win, "Communication-efficient distributed learning over networks—Part II: Necessary conditions for accuracy," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 4, pp. 1102–1119, Apr. 2023.

[57] R. Vershynin, *High-Dimensional Probability: An Introduction With Applications in Data Science*. New York, NY, USA: Cambridge Univ. Press, 2018.

[58] K. Reif, S. Günther, E. Yaz, and R. Unbehauen, "Stochastic stability of the discrete-time extended Kalman filter," *IEEE Trans. Autom. Control*, vol. 44, no. 4, pp. 714–728, Apr. 1999.

[59] C. V. Rao, J. B. Rawlings, and D. Q. Mayne, "Constrained state estimation for nonlinear discrete-time systems: Stability and moving horizon approximations," *IEEE Trans. Autom. Control*, vol. 48, no. 2, pp. 246–258, Feb. 2003.

[60] S. Tatikonda and S. Mitter, "Control over noisy channels," *IEEE Trans. Autom. Control*, vol. 49, no. 7, pp. 1196–1201, Jul. 2004.

[61] I. Gohberg, P. Lancaster, and L. Rodman, *Invariant Subspaces of Matrices With Applications* (Classics in Applied Mathematics), no. 51. Philadelphia, PA, USA: SIAM, 2006.

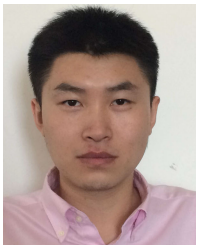
[62] F. M. Callier and C. A. Desoer, *Linear System Theory*. New York, NY, USA: Springer, 1991.

[63] P. Minero and M. Franceschetti, "Anytime capacity of a class of Markov channels," *IEEE Trans. Autom. Control*, vol. 62, no. 3, pp. 1356–1367, Mar. 2017.

[64] G. Como, F. Fagnani, and S. Zampieri, "Anytime reliable transmission of real-valued information through digital noisy channels," *SIAM J. Control Optim.*, vol. 48, no. 6, pp. 3903–3924, Jan. 2010.

[65] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear Estimation* (Prentice-Hall Information and System Sciences Series). Upper Saddle River, NJ, USA: Prentice-Hall, 2000.

[66] T. Kailath, *Linear Systems* (Prentice-Hall Information and System Sciences Series). Englewood Cliffs, NJ, USA: Prentice-Hall, 1980.
[67] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1979.
[68] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, 3rd ed. Cambridge, MA, USA: MIT Press, 2009.
[69] Z. Liu, A. Conti, S. K. Mitter, and M. Z. Win, "Filtering over non-Gaussian channels: The role of anytime capacity," *IEEE Control Syst. Lett.*, vol. 7, pp. 472–477, 2023.
[70] Y. Bar-Shalom, X.-R. Li, and T. Kirubarajan, *Estimation With Applications to Tracking and Navigation*. Hoboken, NJ, USA: Wiley, Jul. 2001.
[71] R. E. Kalman, "A new approach to linear filtering and prediction problems," *J. Basic Eng.*, vol. 82, no. 1, pp. 35–45, May 1960.
[72] T. Kailath, *Lectures on Wiener and Kalman Filtering*. New York, NY, USA: Springer, 1981.
[73] P. R. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification, and Adaptive Control*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1986.
[74] Z. Liu, "Decentralized inference and its application to network localization and navigation," Ph.D. dissertation, Dept. Aeronaut. Astronaut., Massachusetts Inst. Technol., Cambridge, MA, USA, May 2022.
[75] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 2nd ed. New York, NY, USA: Cambridge Univ. Press, 2013.



Zhenyu Liu (Member, IEEE) received the B.S. degree (Hons.) and M.S. degrees in electronic engineering from Tsinghua University, Beijing, China, in 2011 and 2014, respectively, the S.M. degree in aeronautics and astronautics and the Ph.D. degree in networks and statistics from the Massachusetts Institute of Technology (MIT) in 2022.

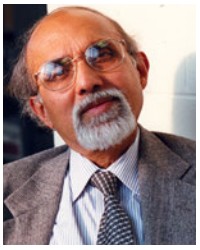
Since 2022, he has been a Post-Doctoral Associate in the wireless information and network sciences laboratory at MIT. His research interests include wireless communications, network localization, distributed inference, networked control, and quantum information science.

Dr. Liu received the first prize of the IEEE Communications Society's Student Competition in 2016 and 2019, the Research and Development 100 Award for Peregrine System in 2018, and the Best Paper Award at the IEEE Latin-American Conference on Communications in 2017.

Andrea Conti (Fellow, IEEE) is a Professor and founding director of the Wireless Communication and Localization Networks Laboratory at the University of Ferrara, Italy. Prior to joining the University of Ferrara, he was with CNIT and with IEIIT-CNR.

In Summer 2001, he was with the Wireless Systems Research Department at AT&T Research Laboratories. Since 2003, he has been a frequent visitor to the Wireless Information and Network Sciences Laboratory at the Massachusetts Institute of Technology, where he presently holds the Research Affiliate appointment. His research interests involve theory and experimentation of wireless communication and localization systems. His current research topics include network localization and navigation, distributed sensing, adaptive diversity communications, and quantum information science.

Dr. Conti has served as editor for IEEE journals and chaired international conferences. He was elected Chair of the IEEE Communications Society's Radio Communications Technical Committee and is Co-founder of the IEEE Quantum Communications & Information Technology Emerging Technical Subcommittee. He received the HTE Puskás Tivadar Medal, the IEEE Communications Society's Fred W. Ellersick Prize, and the IEEE Communications Society's Stephen O. Rice Prize in the field of Communications Theory. He is an elected Fellow of the IEEE and of the IET, and a member of Sigma Xi. He has been selected as an IEEE Distinguished Lecturer.



Sanjoy K. Mitter (Life Fellow, IEEE) received the Ph.D. degree in electrical engineering (automatic control) from the Imperial College London, London, U.K., in 1965.

He taught at Case Western Reserve University from 1965 to 1969. He joined the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, in 1969, where he has been a Professor of electrical engineering since 1973. He was the Director of the MIT Laboratory for Information and Decision Systems from 1981 to 1999. He has also been a Professor of mathematics at the Scuola Normale, Pisa, Italy, from 1986 to 1996. He has held visiting positions at Imperial College London; University of Groningen, The Netherlands; INRIA, France; Tata Institute of Fundamental Research, India; ETH, Zurich, Switzerland; and several American universities. He was the McKay Professor at the University of California, Berkeley, CA, USA, in March 2000, and held the Russell-Severance-Springer Chair in Fall 2003. His current research interests include communication and control in a networked environment, the relationship of statistical and quantum physics to information theory and control, and autonomy and adaptiveness for integrative organization.

Dr. Mitter received the AACC Richard E. Bellman Control Heritage Award in 2007 and the IEEE Eric E. Sumner Award in 2015. He is a member of the National Academy of Engineering. He has received the 2000 IEEE Control Systems Award. Moe Z. Win (Fellow, IEEE) is a Professor at the Massachusetts Institute of Technology (MIT) and the founding director of the Wireless Information and Network Sciences Laboratory. Prior to joining MIT, he was with AT&T Research Laboratories and with NASA Jet Propulsion Laboratory.

Moe Z. Win (Fellow, IEEE) is a Professor at the Massachusetts Institute of Technology (MIT) and the founding director of the Wireless Information and Network Sciences Laboratory. Prior to joining MIT, he was with AT&T Research Laboratories and with NASA Jet Propulsion Laboratory.

His research encompasses fundamental theories, algorithm design, and network experimentation for a broad range of real-world problems. His current research topics include ultra-wideband systems, network localization and navigation, network interference exploitation, and quantum information science. He has served the IEEE Communications Society as an elected Member-at-Large on the Board of Governors, as elected Chair of the Radio Communications Committee, and as an IEEE Distinguished Lecturer. Over the last two decades, he held various editorial positions for IEEE journals and organized numerous international conferences. Recently, he has served on the SIAM Diversity Advisory Committee.

Dr. Win is an elected Fellow of the AAAS, the EURASIP, the IEEE, and the IET. He was honored with two IEEE Technical Field Awards: the IEEE Kiyo Tomiyasu Award (2011) and the IEEE Eric E. Sumner Award (2006, jointly with R. A. Scholtz). His publications, co-authored with students and colleagues, have received several awards. Other recognitions include the MIT Everett Moore Baker Award (2022), the IEEE Vehicular Technology Society James Evans Avant Garde Award (2022), the IEEE Communications Society Edwin H. Armstrong Achievement Award (2016), the Cristoforo Colombo International Prize for Communications (2013), the Copernicus Fellowship (2011) and the *Laurea Honoris Causa* (2008) from the Università degli Studi di Ferrara, and the U.S. Presidential Early Career Award for Scientists and Engineers (2004). He is an ISI Highly Cited Researcher.