Communication-Efficient Distributed Learning Over Networks—Part II: Necessary 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 crucial for many applications such as localization and tracking, autonomy, and crowd sensing. This paper investigates communication-efficient distributed learning of time-varying states over networks. Specifically, the paper considers a network of nodes that infer their current states in a decentralized manner using observations obtained via local sensing and messages obtained via noisy inter-node communications. The paper derives a necessary condition in terms of the sensing and communication capabilities of the network for the boundedness of the learning error over time. The necessary condition is compared with the sufficient condition established in a companion paper and the gap between the two conditions is discussed. The paper provides guidelines for efficient management of the sensing and communication resources for distributed learning in complex networked systems.**

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

I. Introduction

DISTRIBUTED learning aims to estimate states of a networked system by exploiting the sensing, communication, and computing capabilities of nodes in the network [1], [2], [3], [4], [5], [6], [7], [8], [9], [10]. Distributed learning is essential for many emerging applications such as localization and tracking [11], [12], [13], [14], [15], [16], autonomy [17], [18], [19], Internet-of-Things (IoT) [20], [21], [22], and environmental monitoring [23], [24], [25], [26]. This paper investigates communication-efficient distributed learning of time-varying states in complex networked systems, where each node aims to learn a state using its own sensing data and the information exchanged with its neighbors, i.e., other nodes within its communication range. Such a collaborative learning task is referred to as distributed inference in this paper. An application of distributed inference is localization and navigation, where nodes track their positions in real time by sensing the environment and communicating with each other [27], [28], [29], [30], [31]. An important difference between distributed inference and federated learning [8], [9], [10], an emerging paradigm for multi-agent learning, is that the former task does not require any central processor for coordinating the nodes in the network.

Distributed inference is challenging due to the limitations on the sensing and communication capabilities of nodes in the network. In particular, a node typically obtains partial and noisy observations of its unknown state, which can be insufficient for achieving desirable accuracy. Moreover, a node can transmit only a limited number of messages to other nodes at each time and these messages can be corrupted or dropped during transmission. Consequently, communication-efficient strategies for generating the transmitted messages are required so that these messages contain the most useful information for the destination nodes and are robust to corruptions and communication failures. Another challenge to communication-efficient distributed inference over networks is that the latency incurred by message generation and state estimation can reduce the learning accuracy significantly as the state is rapidly time-varying.

A variety of distributed techniques have been proposed for learning [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], optimization [32], [33], [34], and inference [35], [36], [37], [38], [39]. In particular, many of the existing distributed methods require iterative communications among the nodes in the network, such as consensus [40], [41], [42], diffusion [43], [44], [45], and loopy belief propagation [46], [47], [48]. Moreover, different methods have been proposed for improving the communication efficiency in distributed learning, inference, and control for systems with resource constraints. These methods include employing event-triggered communications [49], [50], [51], optimal resource allocation schemes [52], [53], [54], and efficient message quantization and encoding strategies [55], [56], [57]. In particular, real-time encoding techniques are important for the learning and control of dynamic systems under communication constraints due to the negative effect of latency on the performance [58], [59], [60], [61], [62], [63], [64], [65], [66], [67].

In the existing works on distributed learning and inference, many papers do not consider the channel noise or communication failure, or they make strong assumptions such

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 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.2023.3242738.

Digital Object Identifier 10.1109/JSAC.2023.3242738

as the noise follows Gaussian distributions or has bounded supports. Moreover, the iterative mechanisms for distributed learning and inference can incur significant communication overhead and result in intolerable latency. These limitations call for further investigation of distributed inference over networks under a general system model with communication efficiency demand.

An important question for distributed learning is what are the requirements on the sensing and communication capabilities of the network for obtaining desirable inference accuracy? A related question is how to identify the bottleneck of the network and manage the limited sensing and communication resources efficiently? A deep understanding of these questions can benefit the design and operation of networked systems for learning and inference applications. The paper aims to establish a theoretical foundation for communication-efficient distributed learning of time-varying states over networks. The key contributions of this paper are as follows:

- we formulate a communication-efficient distributed inference problem over networks by considering general models for state disturbance, observation noise, and communication channels;
- we derive a necessary condition in terms of the network's sensing and communication capabilities under which the distributed inference error of the network is bounded over time; and
- we compare the derived necessary condition with a sufficient condition established in a companion paper and present a favorable property of the gap between regions of channel capacities for the two conditions.

The remaining sections are organized as follows: Section II presents the problem formulation. Section III describes preliminaries on invariant subspaces and real generalized eigenspaces. Section IV presents a necessary condition under which the distributed inference error is bounded over time. Section V compares the necessary condition with the sufficient condition derived in a companion paper [68]. Finally, Section VI concludes the paper.

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 $\boldsymbol{\mathsf{x}}$ and $\boldsymbol{x}$, respectively. The expectation and covariance matrix of $\boldsymbol{\mathsf{x}}$ are denoted by $\mathbb{E}\{\boldsymbol{\mathsf{x}}\}$ and $\mathbb{V}\{\boldsymbol{\mathsf{x}}\}$, respectively, whereas the conditional expectation of $\boldsymbol{\mathsf{x}}$ given a random vector $\boldsymbol{\mathsf{y}}$ is denoted by $\mathbb{E}\{\boldsymbol{\mathsf{x}}\,|\,\boldsymbol{\mathsf{y}}\}$. The differential entropy of $\boldsymbol{\mathsf{x}}$ and conditional differential entropy of $\boldsymbol{\mathsf{x}}$ given $\boldsymbol{\mathsf{y}}$ are denoted by $h(\boldsymbol{\mathsf{x}})$ and $h(\boldsymbol{\mathsf{x}}\,|\,\boldsymbol{\mathsf{y}})$, respectively. The mutual information between $\boldsymbol{\mathsf{x}}$ and $\boldsymbol{\mathsf{y}}$ is denoted by $I(\boldsymbol{\mathsf{x}};\boldsymbol{\mathsf{y}})$, whereas $I(\boldsymbol{\mathsf{x}};\boldsymbol{\mathsf{y}}\,|\,\boldsymbol{\mathsf{z}})$ represents the conditional mutual information between $\boldsymbol{\mathsf{x}}$ and $\boldsymbol{\mathsf{y}}$ given $\boldsymbol{\mathsf{z}}$. Given a discrete-time stochastic process $\{\boldsymbol{\mathsf{x}}_t\}_{t\geqslant 0}$, the notation $\boldsymbol{\mathsf{x}}_{s:t}$ represents the vertical concatenation of $\boldsymbol{\mathsf{x}}_s, \boldsymbol{\mathsf{x}}_{s+1}, \ldots, \boldsymbol{\mathsf{x}}_t$ 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 is denoted by $\boldsymbol{0}_{m\times n}$; when $n = 1$, the m-dimensional vector of zeros is simply denoted by $\boldsymbol{0}_m$; 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.

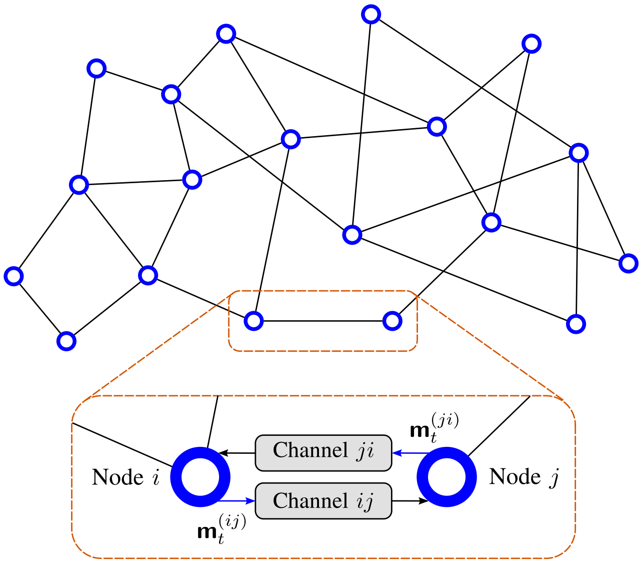


Fig. 1. Distributed inference via sensing and communication in a network.

II. Problem Formulation

Consider a network represented by an undirected graph $\mathscr{G}_{\mathrm{u}} = \{\mathcal{V}, \mathcal{E}_{\mathrm{u}}\}$, where $\mathcal{V}$ represents the vertex set consisting of all the nodes in the network and $\mathcal{E}_{\mathrm{u}}$ represents the edge set. In particular, there is an edge between nodes i and j, namely $(i, j) \in \mathcal{E}_{\mathrm{u}}$, if the two nodes are within the communication range of each other, as shown in Fig. 1. In this case, node i is referred to as a neighbor of node j and vice versa. The set of all the neighbors of node j is denoted by $\mathcal{N}_{\mathrm{u}}^{(j)}$.

Each node in the network is associated with a time varying state. In particular, the state $\boldsymbol{\mathsf{x}}_t^{(j)} \in \mathbb{R}^d$ of node j at

TABLE I

NOTATIONS AND DEFINITIONS OF IMPORTANT QUANTITIES

Notation	Definition	Notation	Definition
$\mathcal{V}$	set of all the nodes in the network	v	the number of nodes in the network
$\mathcal{N}_\mathrm{u}^{(j)}$	set of neighbors of node j	$\mathbf{x}_t^{(j)}$	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
$\varepsilon_t^{(j)}$	individual distributed inference mean-square error of agent j at time t	$\mathcal{I}_{\boldsymbol{F}}(\mathcal{Y})$	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
$\mathcal{N}_\mathrm{s}^{(j)}$	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 $\mathscr{G}_\mathrm{u}$ 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	$\mathring{\mathbf{r}}_t^{(j)}$	messages received by node j from its neighbors in $\mathcal{N}^{(j)}$ at time t
$\mathring{\boldsymbol{\Gamma}}^{(j)}$	sensor gain matrix corresponding to $\mathring{\mathbf{z}}_t^{(j)}$	$\boldsymbol{O}(\mathring{\boldsymbol{\Gamma}}^{(j)},\boldsymbol{A})$	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 $\mathcal{V}$	$C^{(ij)}$	Shannon capacity for the channel from node i to node j

time t 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 deterministic matrix called dynamic matrix of node j; $\boldsymbol{\zeta}_t^{(j)}$ represents the disturbance to $\mathbf{x}_t^{(j)}$ and is modeled as a zero-mean random vector. Each node is equipped with a sensor for observing its own state and the states of its neighbors, as shown in Fig. 2. In particular, a node j can obtain an intra-node observation $\mathbf{z}_t^{(jj)}$ and an inter-node observation $\mathbf{z}_t^{(ji)}$ with each neighbor at every time t. These observations satisfy

$$\mathbf{z}_t^{(jj)} = \boldsymbol{\Gamma}^{(jj)}\mathbf{x}_t^{(j)} + \mathbf{n}_t^{(jj)} \tag{2a}$$

$$\mathbf{z}_t^{(ji)} = \boldsymbol{\Gamma}_1^{(ji)}\mathbf{x}_t^{(j)} + \boldsymbol{\Gamma}_2^{(ji)}\mathbf{x}_t^{(i)} + \mathbf{n}_t^{(ji)} \tag{2b}$$

$$\text{for } t = 0,1,\ldots$$

where $\boldsymbol{\Gamma}^{(jj)}$, $\boldsymbol{\Gamma}_1^{(ji)}$, and $\boldsymbol{\Gamma}_2^{(ji)}$ are deterministic matrices. In particular, $\boldsymbol{\Gamma}^{(jj)}$ is the sensor gain matrix for intra-node observations obtained by node j, whereas $\boldsymbol{\Gamma}_1^{(ji)}$ and $\boldsymbol{\Gamma}_2^{(ji)}$ are sensor gain matrices for inter-node observations obtained by node j with node i. Moreover, $\mathbf{n}_t^{(jj)}$ and $\mathbf{n}_t^{(ji)}$ in (2) are zero-mean random vectors representing observation noise. The observations $\mathring{\mathbf{z}}_t^{(j)}$ obtained by node j at time t consist of $\mathbf{z}_t^{(jj)}$ and $\big\{\mathbf{z}_t^{(ji)} : i \in \mathcal{N}_\mathrm{u}^{(j)}\big\}$, i.e.,

$$\mathring{\mathbf{z}}_t^{(j)} := \Big[\big(\mathbf{z}_t^{(jj)}\big)^\mathrm{T} \quad \big(\mathbf{z}_t^{(j\,i_1)}\big)^\mathrm{T} \quad \big(\mathbf{z}_t^{(j\,i_2)}\big)^\mathrm{T} \ \cdots \ \big(\mathbf{z}_t^{(j\,i_{N^{(j)}})}\big)^\mathrm{T}\Big]^\mathrm{T}. \tag{3}$$

where $i_1, i_2, \ldots, i_{N^{(j)}}$ are all the elements of $\mathcal{N}_\mathrm{u}^{(j)}$ and $N^{(j)} := \big|\mathcal{N}_\mathrm{u}^{(j)}\big|$.

Sensing capabilities of different nodes in a networked system can vary significantly due to device heterogeneity. In particular, the sensing capability of node j is characterized by sensor gain matrices $\boldsymbol{\Gamma}^{(jj)}$, $\boldsymbol{\Gamma}_1^{(ji)}$, and $\boldsymbol{\Gamma}_2^{(ji)}$. For example, if $\boldsymbol{\Gamma}_1^{(ji)} = \boldsymbol{\Gamma}_2^{(ji)} = \mathbf{0}$ for a particular j, then $\mathbf{z}_t^{(ji)} = \mathbf{n}_t^{(ji)}$, i.e., $\mathbf{z}_t^{(ji)}$ contains only noise. This corresponds to the case where node j's sensor does not have enough capability for sensing the state of its neighbor i.

Nodes in the network infer their states collaboratively by transmitting encoded messages to each other. At every time step, each node generates an encoded message in real time for each of its neighbors based on the local observations and received messages obtained by the node. Specifically, let $\mathbf{m}_t^{(ji)}$ represent the encoded message transmitted by node j to its neighbor node i at time t, and let $\mathbf{r}_t^{(ji)}$ represent the corresponding message received by node i. The transmitted message $\mathbf{m}_t^{(ji)}$ can be written as

$$\mathbf{m}_t^{(ji)} = \boldsymbol{\mu}_t^{(ji)}\big(\mathring{\mathbf{z}}_{0:t}^{(j)}, \mathring{\mathbf{r}}_{0:t-1}^{(j)}\big) \tag{4}$$

where $\mathring{\mathbf{r}}_t^{(j)}$ represents the messages received by node j from all its neighbors at time $t = 0,1,\ldots$, i.e.,

$$\mathring{\mathbf{r}}_t^{(j)} := \Big[\big(\mathbf{r}_t^{(i_1\,j)}\big)^\mathrm{T} \quad \big(\mathbf{r}_t^{(i_2\,j)}\big)^\mathrm{T} \ \cdots \ \big(\mathbf{r}_t^{(i_{N^{(j)}}\,j)}\big)^\mathrm{T}\Big]^\mathrm{T} \tag{5}$$

and $\boldsymbol{\mu}_t^{(ji)}$ is a deterministic function referred to as the encoding function of node j for node i at time t. The sequence $\boldsymbol{\mu}_0^{(ji)}, \boldsymbol{\mu}_1^{(ji)}, \ldots$ is called the encoding strategy of node j for node i. Note that the received message $\mathbf{r}_t^{(ji)}$ can be different from the transmitted message $\mathbf{m}_t^{(ji)}$ due to channel noise or communication failure.

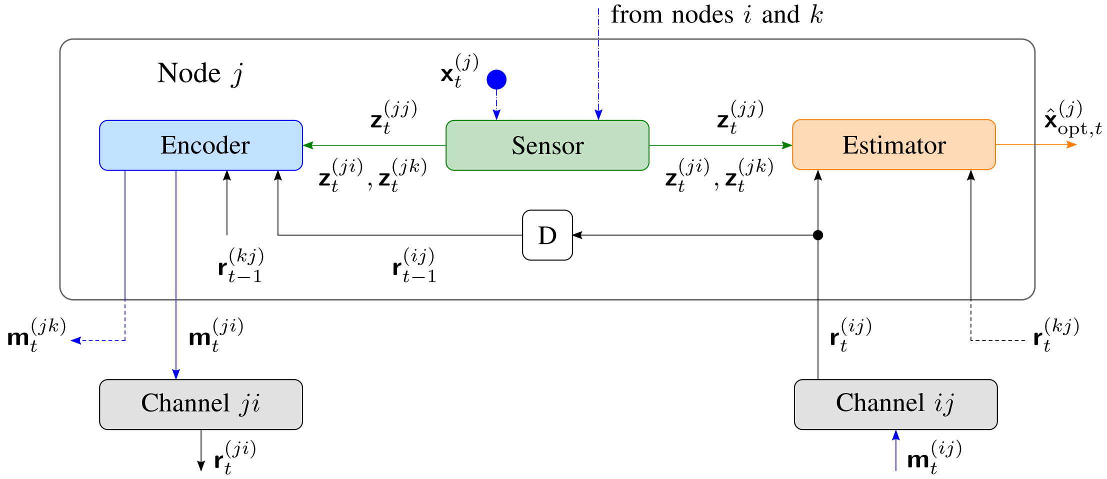


Fig. 2. Block diagram of node j: the sensor observes $\mathbf{z}_t^{(jj)}$, $\mathbf{z}_t^{(ji)}$, and $\mathbf{z}_t^{(jk)}$ at time t. The observations and received messages are used by the encoder for generating transmitted messages $\mathbf{m}_t^{(ji)}$ as well as $\mathbf{m}_t^{(jk)}$ and by the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

A set $\mathcal{V}_{\mathrm{a}} \subseteq \mathcal{V}$ of nodes named agents aim to learn their states in real time. Indeed, $\mathcal{V}_{\mathrm{a}}$ is an arbitrary non-empty subset of $\mathcal{V}$. For example, $\mathcal{V}_{\mathrm{a}}$ can contain a single node or it can contain all the nodes in the network. An agent node j computes an estimator of $\mathbf{x}_t^{(j)}$ at each time t using the local observations $\mathring{\mathbf{z}}_{0:t}^{(j)}$ and received messages $\mathring{\mathbf{r}}_{0:t}^{(j)}$ obtained up to time t. This paper adopts the mean-square error (MSE) as the metric for the learning performance of each agent. Given $\mathring{\mathbf{z}}_{0:t}^{(j)}$ and $\mathring{\mathbf{r}}_{0:t}^{(j)}$, the optimal distributed estimator at agent j with the minimum MSE is the minimum-mean-square-error (MMSE) estimator $\hat{\mathbf{x}}_{\mathrm{opt},t}^{(j)}$ given by $\hat{\mathbf{x}}_{\mathrm{opt},t}^{(j)} := \mathbb{E}\{\mathbf{x}_t^{(j)} \,|\, \mathring{\mathbf{z}}_{0:t}^{(j)}, \mathring{\mathbf{r}}_{0:t}^{(j)}\}$. See Fig. 2 for a block diagram of the sensing, encoding, and inference performed by node j.

Encoding strategies are critical for the accuracy of distributed estimators as they determine whether the received messages provide useful information for inferring the states and whether they are robust to channel noise and communication failure. To maximize the distributed learning performance, non-linear schemes need to be considered for encoding functions $\boldsymbol{\mu}_t^{(ji)}$. Consequently, the distributed learning problem described above is non-linear and is challenging to solve. Furthermore, the distributed learning problem is different from the problem of inferring a global unknown state by a multi-node network, which has been investigated in the literature. In particular, many existing works employ a consensus mechanism, which involves the exchange of information about the global state among nodes, in order to achieve consistent estimation results across the network. Different from those works, this paper does not consider consensus mechanisms as each node aims to infer its own state.

We focus on a connected graph $\mathscr{G}_{\mathrm{u}}$, namely there exists a path between an arbitrary pair of vertices in the graph. The generalization to cases where $\mathscr{G}_{\mathrm{u}}$ is not connected is straightforward: each connected component of the graph is treated as a separate network and the results presented in this paper are applied for each network. The following assumptions are made on the initial state $\mathbf{x}_0^{(j)}$, state disturbance $\boldsymbol{\zeta}_t^{(j)}$, observation noise $\mathbf{n}_t^{(jj)}$ and $\mathbf{n}_t^{(ji)}$, and communication channels.

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 matrix $\boldsymbol{B}$ consisting of d rows and vector $\boldsymbol{b} \in \mathbb{R}^d$ such that $\boldsymbol{b} \notin \mathcal{C}(\boldsymbol{B})$, there exists a number $\underline{h}^{(j)}(\boldsymbol{b}, \boldsymbol{B}) > -\infty$ such that $h\big(\boldsymbol{b}^{\mathrm{T}}\boldsymbol{\zeta}_t^{(j)} \,\big|\, \boldsymbol{B}^{\mathrm{T}}\boldsymbol{\zeta}_t^{(j)}\big) \geqslant \underline{h}^{(j)}(\boldsymbol{b}, \boldsymbol{B})$ for all $t \geqslant 0$.

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

A5. The channel between each pair of nodes is memoryless.

Assumption A3 states that if $\boldsymbol{b} \notin \mathcal{C}(\boldsymbol{B})$, then there is uncertainty in $\boldsymbol{b}^{\mathrm{T}}\boldsymbol{\zeta}_t^{(j)}$ even though the linear combinations $\boldsymbol{B}^{\mathrm{T}}\boldsymbol{\zeta}_t^{(j)}$ of $\boldsymbol{\zeta}_t^{(j)}$ are known. This is a mild assumption. As an example, this assumption holds if entries of $\boldsymbol{\zeta}_t^{(j)}$ are independent and have Gaussian or uniform distributions with variances bounded from below by a positive constant for all $t \geqslant 0$.

In the considered distributed learning problem, both the transmitter and the receiver know the channel state information. However, the channel is not assumed to be reciprocal, namely the condition of the channel from node i to its neighbor node j can be different than that from node j to node i. Furthermore, the paper considers the scenario where the communication in the network is coordinated via a multiple access scheme to avoid interference.

Note that the models of the disturbance, noise, and channels used by the paper are general. For the ease of presentation, the paper considers scenarios where the magnitudes of all the eigenvalues of $\boldsymbol{A}^{(j)}$ are no smaller than one for all $j \in \mathcal{V}$. Results in this paper can be extended to scenarios without any assumption on the eigenvalues of $\boldsymbol{A}^{(j)}$.

The learning objective function F_t for the network at time t is the total distributed inference MSE of all the agents.

In particular, F_t is defined as

$$F_t := \sum_{j\in\mathcal{V}_\mathrm{a}} \varepsilon_t^{(j)} \tag{6}$$

where $\varepsilon_t^{(j)}$ is the individual distributed inference MSE of agent j at time t defined as

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

Here, $\hat{\mathbf{x}}_{\mathrm{opt},t}^{(j)}$ represents the distributed estimator of $\mathbf{x}_t^{(j)}$ at node j with minimum MSE and is given by the conditional expectation $\hat{\mathbf{x}}_{\mathrm{opt},t}^{(j)} := \mathbb{E}\{\mathbf{x}_t^{(j)} \,|\, \mathring{\mathbf{z}}_{0:t}^{(j)}, \mathring{\mathbf{r}}_{0:t}^{(j)}\}$. The error $\varepsilon_t^{(j)}$ is affected by the quality of $\mathring{\mathbf{z}}_{0:t}^{(j)}$ and $\mathring{\mathbf{r}}_{0:t}^{(j)}$. In particular, $\mathring{\mathbf{r}}_{0:t}^{(j)}$ is determined by the encoding strategies employed by nodes in the network.

This paper investigates fundamental limits of distributed inference when the optimal encoding strategies are employed. Specifically, the paper establishes a necessary condition under which the sequence $\{F_t\}_{t\geqslant 0}$ is bounded over time, i.e.,

$$\sup_{t\geqslant 0}\, F_t < \infty\,. \tag{7}$$

This is equivalent to

$$\sup_{t\geqslant 0}\, \varepsilon_t^{(j)} < \infty \quad \forall j \in \mathcal{V}_\mathrm{a}\,. \tag{8}$$

Boundedness of inference error is an important property, which has been studied in the literature [23], [69], [70].

In the distributed learning problem investigated by this paper, communication efficiency is accounted for as in the following. First, the communication channel between each pair of nodes is used at most only once. This is more efficient compared to many existing works on distributed learning that require multiple iterations of communication among nodes at each time step. Second, the limitation on the communication capability over each channel is taken into account in our distributed learning problem. In fact, the derived necessary condition contains a subcondition on the Shannon capacities of communication channels in the network. By contrast, many existing works do not consider limitations on the capacities of channels and assume that real vectors or matrices can be transmitted among nodes with no loss. Even if such transmission can be achieved approximately, its communication overhead and resource consumption can be significantly high.

III. Preliminaries

First, the definition of invariant subspaces is presented. 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 $\boldsymbol{u} \in \mathcal{Y}$. Define subspace $\mathcal{I}_{\boldsymbol{F}}(\mathcal{Y})$ as

$$\mathcal{I}_{\boldsymbol{F}}(\mathcal{Y}) := \mathcal{C}\big(\big[\boldsymbol{Y} \quad \boldsymbol{F}\boldsymbol{Y} \quad \cdots \quad \boldsymbol{F}^{n-1}\boldsymbol{Y}\big]\big) \tag{9}$$

where $\boldsymbol{Y}$ is a matrix whose columns form a basis of $\mathcal{Y}$. Subspace $\mathcal{I}_{\boldsymbol{F}}(\mathcal{Y})$ is called the minimum $\boldsymbol{F}$-invariant subspace over $\mathcal{Y}$ [71, Chapter 2], since $\mathcal{I}_{\boldsymbol{F}}(\mathcal{Y})$ is $\boldsymbol{F}$-invariant, contains $\mathcal{Y}$, and is not a proper subset of any $\boldsymbol{F}$-invariant subspace containing $\mathcal{Y}$. Invariant subspaces are used in Section IV and Appendix B for establishing a necessary condition for the total distributed inference MSE to be bounded by performing linear transformations of the unknown states.

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

$$\Lambda(\boldsymbol{F}) := \big\{\lambda : \mathrm{Im}(\lambda) \geqslant 0;\, \exists \boldsymbol{u} \in \mathbb{C}^n, \boldsymbol{u} \neq \boldsymbol{0} \text{ s.t. } \boldsymbol{F}\boldsymbol{u} = \lambda\boldsymbol{u}\big\} \tag{10}$$

where $\mathrm{Im}(\lambda)$ represents the imaginary part of λ. A real generalized eigenspace $\mathcal{M}_\lambda(\boldsymbol{F})$ associated with an arbitrary element λ of $\Lambda(\boldsymbol{F})$ is defined as

$$\mathcal{M}_\lambda(\boldsymbol{F}) := \begin{cases} \mathcal{M}_\lambda^\mathrm{C}(\boldsymbol{F}) \cap \mathbb{R}^n & \text{if } \lambda \in \mathbb{R} \\ \big(\mathcal{M}_\lambda^\mathrm{C}(\boldsymbol{F}) + \mathcal{M}_{\lambda^*}^\mathrm{C}(\boldsymbol{F})\big) \cap \mathbb{R}^n & \text{otherwise} \end{cases} \tag{11}$$

where

$$\mathcal{M}_\lambda^\mathrm{C}(\boldsymbol{F}) := \big\{\boldsymbol{u} \in \mathbb{C}^n : \big(\boldsymbol{F} - \lambda\boldsymbol{I}\big)^n\boldsymbol{u} = \boldsymbol{0}\big\} \tag{12}$$

represents the generalized eigenspace of $\boldsymbol{F}$ associated with λ over $\mathbb{C}$. The set $\mathcal{M}_\lambda(\boldsymbol{F})$ is a subspace of $\mathbb{R}^n$ and is $\boldsymbol{F}$-invariant. Set $\mathcal{M}_\lambda(\boldsymbol{F})$ is referred to as the real generalized eigenspace of $\boldsymbol{F}$ associated with λ. 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

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

In Proposition 1, $\oplus$ represents the direct sum of subspaces [72, Chapter 4]. This proposition is adapted from the decomposition theorem in [72, Chapter 4.3].

Using real generalized eigenspaces of a real square matrix helps to convert the matrix to its real Jordan canonical form via a similarity transformation. Such a transformation is useful for analyzing dynamic systems and is also used in this paper.

IV. Necessary Condition for the Boundedness of Total Distributed Inference MSE

In this section, we consider the scenario where there is no cycle in $\mathscr{G}_\mathrm{u}$. Moreover, we omit the subscript u in $\mathcal{N}_\mathrm{u}^{(j)}$ and write it as $\mathcal{N}^{(j)}$ in the rest of the paper. Some definitions are given below for simplicity of the presentation. Specifically, define

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

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

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

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

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

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

where for a given integer $m > 0$, $\boldsymbol{e}_{k,m}$ represents a unit vector of m entries with its kth entry being one and other entries being zero for $1 \leqslant k \leqslant m$. Define a subspace $\mathcal{X}^{(j)} \subseteq \mathbb{R}^{dv}$ as

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

Recall that node j aims to infer $\mathbf{x}_t^{(j)}$, which is equivalent to estimating the projection of the concatenated state $\mathbf{x}_t$ onto $\mathcal{X}^{(j)}$. Therefore, $\mathcal{X}^{(j)}$ is a subspace that corresponds to $\mathbf{x}_t^{(j)}$, and we refer to $\mathcal{X}^{(j)}$ as node j's state subspace. Note that the problem formulation is not affected by the introduction of $\mathbf{x}_t$ and the distributed inference problem without any central processor is not changed.

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

$$\mathbf{z}_t^{(jj)} = \mathring{\boldsymbol{\Gamma}}^{(jj)}\mathbf{x}_t + \mathbf{n}_t^{(jj)} \tag{17a}$$

$$\mathbf{z}_t^{(ji)} = \mathring{\boldsymbol{\Gamma}}^{(ji)}\mathbf{x}_t + \mathbf{n}_t^{(ji)} \tag{17b}$$

where $\mathring{\boldsymbol{\Gamma}}^{(jj)}$ and $\mathring{\boldsymbol{\Gamma}}^{(ji)}$ are defined as

$$\mathring{\boldsymbol{\Gamma}}^{(jj)} := \boldsymbol{e}_{j,v}^{\mathrm{T}} \otimes \boldsymbol{\Gamma}^{(jj)} \tag{18a}$$

$$\mathring{\boldsymbol{\Gamma}}^{(ji)} := \boldsymbol{e}_{j,v}^{\mathrm{T}} \otimes \boldsymbol{\Gamma}_1^{(ji)} + \boldsymbol{e}_{i,v}^{\mathrm{T}} \otimes \boldsymbol{\Gamma}_2^{(ji)}. \tag{18b}$$

Combining (3) and (17b), $\mathring{\mathbf{z}}_t^{(j)}$ can be written as

$$\mathring{\mathbf{z}}_t^{(j)} = \mathring{\boldsymbol{\Gamma}}^{(j)}\mathbf{x}_t + \mathring{\mathbf{n}}_t^{(j)} \tag{19}$$

where

$$\mathring{\boldsymbol{\Gamma}}^{(j)} := \Big[\big(\mathring{\boldsymbol{\Gamma}}^{(jj)}\big)^{\mathrm{T}}\ \big(\mathring{\boldsymbol{\Gamma}}^{(j\,i_1)}\big)^{\mathrm{T}}\ \big(\mathring{\boldsymbol{\Gamma}}^{(j\,i_2)}\big)^{\mathrm{T}} \cdots \big(\mathring{\boldsymbol{\Gamma}}^{(j\,i_{N^{(j)}})}\big)^{\mathrm{T}}\Big]^{\mathrm{T}} \tag{20}$$

$$\mathring{\mathbf{n}}_t^{(j)} := \Big[\big(\mathbf{n}_t^{(jj)}\big)^{\mathrm{T}}\ \ \big(\mathbf{n}_t^{(j\,i_1)}\big)^{\mathrm{T}}\ \ \big(\mathbf{n}_t^{(j\,i_2)}\big)^{\mathrm{T}}\ \cdots \big(\mathbf{n}_t^{(j\,i_{N^{(j)}})}\big)^{\mathrm{T}}\Big]^{\mathrm{T}}. \tag{21}$$

Here, indices $i_1, i_2, \ldots, i_{N^{(j)}}$ are all the elements of $\mathcal{N}^{(j)}$.

Entries of $\mathring{\mathbf{z}}_t^{(j)}$ can be viewed as observations of $\mathbf{x}_t$. According to (19), the observability matrix corresponding to observations obtained by node j is $\boldsymbol{O}\big(\mathring{\boldsymbol{\Gamma}}^{(j)}, \boldsymbol{A}\big)$, where $\boldsymbol{O}(\boldsymbol{G}, \boldsymbol{F})$ is defined as [73], [74], and [75]

$$\boldsymbol{O}(\boldsymbol{G}, \boldsymbol{F}) := \big[\boldsymbol{G}^{\mathrm{T}}\quad \boldsymbol{F}^{\mathrm{T}}\boldsymbol{G}^{\mathrm{T}}\ \ \cdots\ \ (\boldsymbol{F}^{k-1})^{\mathrm{T}}\boldsymbol{G}^{\mathrm{T}}\big]^{\mathrm{T}} \tag{22}$$

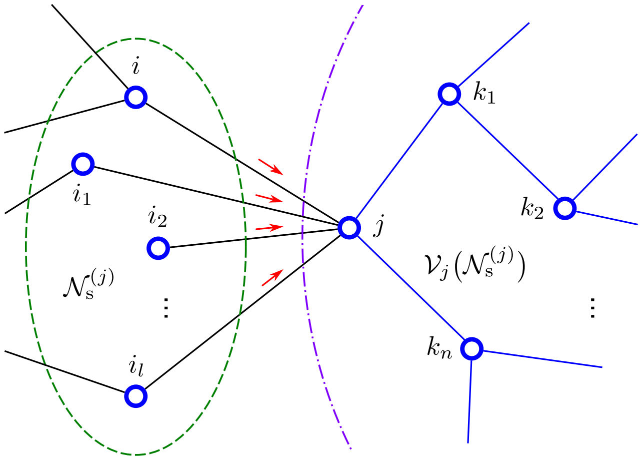
for general matrices $\boldsymbol{F} \in \mathbb{R}^{k\times k}$ and $\boldsymbol{G}$ with k columns. Define the observable subspace $\mathcal{S}(\mathcal{V}_0)$ corresponding to observations obtained by nodes in $\mathcal{V}_0$ as

$$\mathcal{S}(\mathcal{V}_0) := \mathcal{C}\Big(\boldsymbol{O}\big(\big[\mathring{\boldsymbol{\Gamma}}^{(j)}\big]_{j\in\mathcal{V}_0}, \boldsymbol{A}\big)^{\mathrm{T}}\Big) \tag{23}$$

where $\big[\mathring{\boldsymbol{\Gamma}}^{(j)}\big]_{j\in\mathcal{V}_0}$ represents the vertical concatenation of $\mathring{\boldsymbol{\Gamma}}^{(j)}$ for all $j \in \mathcal{V}_0$. The next lemma shows that using observations obtained by nodes in $\mathcal{V}_0$, we can compute an estimator of $\boldsymbol{H}^{\mathrm{T}}\mathbf{x}_t$ with bounded MSE over time for any $\boldsymbol{H}$ whose columns belong to $\mathcal{S}(\mathcal{V}_0)$. This is why we call $\mathcal{S}(\mathcal{V}_0)$ an observable subspace.

Lemma 1: For any real matrix $\boldsymbol{H}$ whose columns belong to $\mathcal{S}(\mathcal{V}_0)$, an estimator $\hat{\boldsymbol{\upbeta}}_t$ of $\boldsymbol{H}^{\mathrm{T}}\mathbf{x}_t$ can be constructed at each time $t \geqslant 0$ using observations $\big\{\mathring{\mathbf{z}}_{0:t}^{(j)} : j \in \mathcal{V}_0\big\}$ obtained by nodes in $\mathcal{V}_0$ so that the MSE of $\hat{\boldsymbol{\upbeta}}_t$ is bounded over time, i.e.,

$$\sup_{t\geqslant 0}\ \mathbb{E}\Big\{\big\|\hat{\boldsymbol{\upbeta}}_t - \boldsymbol{H}^{\mathrm{T}}\mathbf{x}_t\big\|^2\Big\} < \infty\,. \tag{24}$$



i
i_1
i_2
i_l
j
k_1
k_2
k_n
$\mathcal{N}_{\mathrm{s}}^{(j)}$
$\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)$
$\vdots$
$\vdots$

Fig. 3. Set $\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)$ in part of the network. In particular, $\mathcal{N}_{\mathrm{s}}^{(j)}$ consists of the nodes inside the dashed green ellipse, whereas $\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)$ consists of the nodes to the right of the dash-dotted purple curve. Messages received by node j from all the nodes in $\mathcal{N}_{\mathrm{s}}^{(j)}$ are represented by red arrows.

Proof: See Appendix A. □

The observable subspace $\mathcal{S}(\mathcal{V}_0)$ is $\boldsymbol{A}^{\mathrm{T}}$-invariant, namely

$$\mathcal{I}_{\boldsymbol{A}^{\mathrm{T}}}\big(\mathcal{S}(\mathcal{V}_0)\big) = \mathcal{S}(\mathcal{V}_0)\,. \tag{25}$$

For the special case where $\mathcal{V}_0 = \{j\}$, the observable subspace $\mathcal{S}(\{j\})$ is given by $\mathcal{S}(\{j\}) = \mathcal{C}\big(\boldsymbol{O}\big(\mathring{\boldsymbol{\Gamma}}^{(j)}, \boldsymbol{A}\big)^{\mathrm{T}}\big)$.

Finally, for any $j \in \mathcal{V}$ and subset $\mathcal{N}_{\mathrm{s}}^{(j)}$ of $\mathcal{N}^{(j)}$, define $\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big) \subseteq \mathcal{V}$ as a subset consisting of node j and nodes that are connected to j if edges (j, l) are removed from the graph for all $l \in \mathcal{N}_{\mathrm{s}}^{(j)}$. Mathematically,

$$\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big) := \{j\} \cup \Big\{k \in \mathcal{V} : k \leftrightarrow j \text{ in graph } \big\{\mathcal{V},\, \mathcal{E}_{\mathrm{u}} \setminus \{(j,l) : l \in \mathcal{N}_{\mathrm{s}}^{(j)}\}\big\}\Big\} \tag{26}$$

where $k \leftrightarrow j$ represents that there is a path between nodes k and j. The definition of $\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)$ is illustrated in Fig. 3.

For the case where there is no cycle in $\mathscr{G}_{\mathrm{u}}$, a necessary condition for $\{F_t\}_{t\geqslant 0}$ to be bounded, i.e., $\sup_{t\geqslant 0} F_t < \infty$, is given below.

Theorem 1: Consider the distributed learning problem presented in Section II. If there is no cycle in the graph $\mathscr{G}_{\mathrm{u}}$, then a necessary condition for achieving $\sup_{t\geqslant 0} F_t < \infty$ is that both of the following two subconditions hold.

(i) For every $j \in \mathcal{V}_{\mathrm{a}}$, the following relationship holds

$$\mathcal{X}^{(j)} \subseteq \mathcal{S}(\mathcal{V})\,. \tag{27}$$

(ii) For any $j \in \mathcal{V}_{\mathrm{a}}$ and non-empty subset $\mathcal{N}_{\mathrm{s}}^{(j)}$ of $\mathcal{N}^{(j)}$ such that $\mathcal{X}^{(j)} \not\subseteq \mathcal{S}\big(\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\big)$, the Shannon capacity $C^{(ij)}$ for the channel from node $i \in \mathcal{N}_{\mathrm{s}}^{(j)}$ to node j satisfies

$$\sum_{i\in\mathcal{N}_{\mathrm{s}}^{(j)}} C^{(ij)} > \sum_{\lambda\in\Lambda(\boldsymbol{A}^{(j)})} r_\lambda^{(j)}\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\log|\lambda| \tag{28}$$

where $r_\lambda^{(j)}\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)$ is defined as

$$\begin{aligned} r_\lambda^{(j)}\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big) &:= \dim\big(\mathcal{X}^{(j)} \cap \mathcal{M}_\lambda(\boldsymbol{A}^{\mathrm{T}})\big) \\ &\quad - \dim\big(\mathcal{S}\big(\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\big) \cap \mathcal{X}^{(j)} \cap \mathcal{M}_\lambda(\boldsymbol{A}^{\mathrm{T}})\big)\,. \end{aligned} \tag{29}$$

Proof: An outline for the proof of Subcondition (ii) is provided here, and Subcondition (i) can be proved similarly. A detailed proof for the theorem is presented in Appendix B.

Consider an auxiliary processor that can obtain two collections of data from two disjoint subsets of nodes in the network. The first collection of data is from $\mathcal{N}_{\mathrm{s}}^{(j)}$ and consists of messages received by agent j from nodes in $\mathcal{N}_{\mathrm{s}}^{(j)}$. The second collection of data is from subset $\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)$ and consists of all the observations obtained by nodes in $\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)$ as well as all the messages received via edges that connect these nodes. Note that this processor has access to more data than agent j does. In the proof, we construct a random vector $\boldsymbol{\theta}_t$, which is a linear function of $\mathbf{x}_t$, such that the auxiliary processor can compute an estimator of $\boldsymbol{\theta}_t$ whose MSE is bounded over time. In particular, the auxiliary processor needs to extract information of $\boldsymbol{\theta}_t$ contained in messages received by agent j from nodes in $\mathcal{N}_{\mathrm{s}}^{(j)}$. To obtain enough information, the channels from nodes in $\mathcal{N}_{\mathrm{s}}^{(j)}$ to agent j need to be sufficiently reliable, which leads to the requirement (28) on the total capacities of these channels.

Next, the data available to the auxiliary processor as well as the construction of $\boldsymbol{\theta}_t$ are presented, and the proof of (28) is briefly described. Without loss of generality, let elements of $\mathcal{N}_{\mathrm{s}}^{(j)}$ and $\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)$ be (see Fig. 3)

$$\begin{aligned}\mathcal{N}_{\mathrm{s}}^{(j)} &= \{i, i_1, i_2, \ldots, i_l\}\\ \mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big) &= \{j, k_1, k_2, \ldots, k_n\}\,.\end{aligned} \tag{30}$$

Let $\mathring{\mathbf{r}}_t^{(j)}\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)$ represent all the messages received by node j from nodes in set $\mathcal{N}_{\mathrm{s}}^{(j)}$ at time t, i.e.,

$$\mathring{\mathbf{r}}_t^{(j)}\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big) := \Big[\big(\mathbf{r}_t^{(i\,j)}\big)^{\mathrm{T}} \;\; \big(\mathbf{r}_t^{(i_1\,j)}\big)^{\mathrm{T}} \;\; \big(\mathbf{r}_t^{(i_2\,j)}\big)^{\mathrm{T}} \cdots \big(\mathbf{r}_t^{(i_l\,j)}\big)^{\mathrm{T}}\Big]^{\mathrm{T}}. \tag{31}$$

Received messages $\mathring{\mathbf{r}}_t^{(j)}\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)$ are represented by red arrows in Fig. 3. The collection of data obtained by the auxiliary processor from $\mathcal{N}_{\mathrm{s}}^{(j)}$ up to time t is $\mathring{\mathbf{r}}_{0:t}^{(j)}\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)$, which is defined as

$$\begin{aligned}\mathring{\mathbf{r}}_{0:t}^{(j)}\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big) := \bigg[\Big(\mathring{\mathbf{r}}_0^{(j)}\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\Big)^{\mathrm{T}} \quad \Big(\mathring{\mathbf{r}}_1^{(j)}\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\Big)^{\mathrm{T}}&\\ \cdots \; \Big(\mathring{\mathbf{r}}_t^{(j)}\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\Big)^{\mathrm{T}}\bigg]^{\mathrm{T}}&.\end{aligned}$$

Moreover, define the set $\mathcal{E}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)$ of edges that connect nodes in $\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)$ as

$$\begin{aligned}\mathcal{E}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big) := \big\{(k,l) : k \in \mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big),\, l \in \mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big),&\\ (k,l) \in \mathcal{E}_{\mathrm{u}}\big\}&\end{aligned} \tag{32}$$

where we recall that $\mathcal{E}_{\mathrm{u}}$ represents the edge set of $\mathscr{G}_{\mathrm{u}}$. Edges in $\mathcal{E}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)$ are represented by solid blue lines to the right of the dash-dotted purple curve in Fig. 3. At time t, the observations $\breve{\mathbf{z}}_t$ obtained by nodes in $\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)$ as well as the messages $\breve{\mathbf{r}}_t$ received via edges in $\mathcal{E}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)$ are then given by

$$\breve{\mathbf{z}}_t := \Big[\big(\mathring{\mathbf{z}}_t^{(j)}\big)^{\mathrm{T}} \quad \big(\mathring{\mathbf{z}}_t^{(k_1)}\big)^{\mathrm{T}} \quad \big(\mathring{\mathbf{z}}_t^{(k_2)}\big)^{\mathrm{T}} \;\cdots\; \big(\mathring{\mathbf{z}}_t^{(k_n)}\big)^{\mathrm{T}}\Big]^{\mathrm{T}} \tag{33}$$

$$\breve{\mathbf{r}}_t := \big[\mathbf{r}_t^{(kl)}\big]_{(k,l)\in\mathcal{E}_j(\mathcal{N}_{\mathrm{s}}^{(j)})} \tag{34}$$

where $\mathring{\mathbf{z}}_t^{(j)}$ is defined in (3), and $\big[\mathbf{r}_t^{(kl)}\big]_{(k,l)\in\mathcal{E}_j(\mathcal{N}_{\mathrm{s}}^{(j)})}$ represents the vertical concatenation of all the $\mathbf{r}_t^{(kl)}$ and $\mathbf{r}_t^{(lk)}$ with $(k,l) \in \mathcal{E}_j(\mathcal{N}_{\mathrm{s}}^{(j)})$. Consequently, the collection of data obtained by the auxiliary processor from $\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)$ up to time t consists of $\breve{\mathbf{z}}_{0:t}$ and $\breve{\mathbf{r}}_{0:t}$. In summary, the data obtained by the auxiliary processor up to time t consist of $\mathring{\mathbf{r}}_{0:t}^{(j)}\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)$ as well as $\breve{\mathbf{z}}_{0:t}$ and $\breve{\mathbf{r}}_{0:t}$.

Random vector $\boldsymbol{\theta}_t$ is defined as $\boldsymbol{\theta}_t := \boldsymbol{T}_\theta^{\mathrm{T}}\mathbf{x}_t$, where $\boldsymbol{T}_\theta$ is a matrix with orthonormal columns such that $\mathcal{C}(\boldsymbol{T}_\theta)$ is the orthogonal complement of $\mathcal{S}\big(\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\big)$ with respect to $\mathcal{X}^{(j)} + \mathcal{S}\big(\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\big)$. In other words, $\boldsymbol{T}_\theta^{\mathrm{T}}\boldsymbol{u} = \mathbf{0}$ for any $\boldsymbol{u} \in \mathcal{S}\big(\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\big)$ and $\mathcal{C}(\boldsymbol{T}_\theta) \subseteq \mathcal{X}^{(j)} + \mathcal{S}\big(\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\big)$. The intuition for constructing $\boldsymbol{\theta}_t$ in such a manner is given below. First, if $\sup_{t\geqslant 0}\varepsilon_t^{(j)} < \infty$, then the auxiliary processor can compute an MMSE estimator of $\boldsymbol{\theta}_t$ whose MSE is bounded over time, i.e.,

$$\sup_{t\geqslant 0}\mathbb{E}\Big\{\big\|\boldsymbol{\theta}_t - \mathbb{E}\big\{\boldsymbol{\theta}_t \,\big|\, \breve{\mathbf{z}}_{0:t}, \breve{\mathbf{r}}_{0:t}, \mathring{\mathbf{r}}_{0:t}^{(j)}\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\big\}\big\|^2\Big\} < \infty\,. \tag{35}$$

Second, the auxiliary processor needs to rely on messages $\mathring{\mathbf{r}}_{0:t}^{(j)}\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)$ received from $\mathcal{N}_{\mathrm{s}}^{(j)}$ for achieving (35), as the data $\breve{\mathbf{z}}_{0:t}$ and $\breve{\mathbf{r}}_{0:t}$ obtained from $\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)$ are less informative for inferring $\boldsymbol{T}_\theta^{\mathrm{T}}\mathbf{x}_t$. This is because $\mathcal{C}(\boldsymbol{T}_\theta)$ is orthogonal to $\mathcal{S}\big(\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\big)$, which is the observable subspace corresponding to observations obtained by nodes in $\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)$.

Inequality (28) is proved using the maximum differential entropy lemma [76, Chapter 2], which establishes an inequality between MSE and conditional entropy power. Specifically, the conditional entropy power $N(\mathbf{x}\,|\,\mathbf{y})$ of a general random vector $\mathbf{x} \in \mathbb{R}^n$ given a general random vector $\mathbf{y}$ is defined as

$$N(\mathbf{x}\,|\,\mathbf{y}) := \exp\big\{2h(\mathbf{x}\,|\,\mathbf{y})/n\big\}\,. \tag{36}$$

The maximum differential entropy lemma gives

$$\begin{aligned}&\mathbb{E}\Big\{\big\|\boldsymbol{\theta}_t - \mathbb{E}\big\{\boldsymbol{\theta}_t \,\big|\, \breve{\mathbf{z}}_{0:t}, \breve{\mathbf{r}}_{0:t}, \mathring{\mathbf{r}}_{0:t}^{(j)}\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\big\}\big\|^2\Big\}\\ &\quad \geqslant \frac{1}{2\pi e}N\big(\boldsymbol{\theta}_t \,\big|\, \breve{\mathbf{z}}_{0:t}, \breve{\mathbf{r}}_{0:t}, \mathring{\mathbf{r}}_{0:t}^{(j)}\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\big) \geqslant \frac{1}{2\pi e}N\big(\boldsymbol{\theta}_t \,\big|\, \boldsymbol{\psi}_t\big)\end{aligned} \tag{37}$$

where $\boldsymbol{\psi}_t$ is defined in (74) and contains all the entries of $\breve{\mathbf{z}}_{0:t}$, $\breve{\mathbf{r}}_{0:t}$, and $\mathring{\mathbf{r}}_{0:t}^{(j)}\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)$. The second inequality in (37) is because conditioning reduces differential entropy. Combining (37) with (35) gives $\sup_{t\geqslant 0} N\big(\boldsymbol{\theta}_t \,\big|\, \boldsymbol{\psi}_t\big) < \infty$. In Appendix B, a recursive expression for $\big\{N\big(\boldsymbol{\theta}_t \,\big|\, \boldsymbol{\psi}_t\big)\big\}_{t\geqslant 0}$ is derived based on the fact that both $\mathcal{X}^{(j)}$ and $\mathcal{S}\big(\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\big)$ are $\boldsymbol{A}^{\mathrm{T}}$-invariant. Using the recursive expression, it is shown there that $\sup_{t\geqslant 0} N\big(\boldsymbol{\theta}_t \,\big|\, \boldsymbol{\psi}_t\big) < \infty$ only if

$$\sum_{i\in\mathcal{N}_{\mathrm{s}}^{(j)}} C^{(ij)} > \log\big|\det\big(\boldsymbol{T}_\theta^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{T}_\theta\big)\big|\,. \tag{38}$$

It is also shown in Appendix B that

$$\log\big|\det\big(\boldsymbol{T}_\theta^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{T}_\theta\big)\big| \;=\; \sum_{\lambda\in\Lambda(\boldsymbol{A}^{(j)})} r_\lambda^{(j)}\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\log|\lambda|\,. \tag{39}$$

Combining (38) with (39) gives the desired result (28). □

Subcondition (i) in Theorem 1 is on the sensing capability of the network. The sensing subcondition can be interpreted by considering a centralized MMSE estimator of $\mathbf{x}_t^{(j)}$ based on observations obtained by all the nodes in the network up to time t. This is equivalent to the scenario where the communications among nodes are ideal so that each node can transmit infinite amount of information to its neighbors with no loss. As a result, each node can disseminate all its observations via such communications to the entire network. Consequently, each agent can estimate its state using observations obtained by all the nodes in the network. Note that the MSE of the centralized MMSE estimator for $\mathbf{x}_t^{(j)}$ is no larger than the individual distributed inference MSE $\varepsilon_t^{(j)}$ as the centralized MMSE estimator uses more observations than the distributed MMSE estimator. Consequently, if $\varepsilon_t^{(j)}$ is bounded over time, namely (8) holds, then the MSE of this centralized MMSE estimator is also bounded over time. This requires that relationship (27) hold, namely the subspace $\mathcal{X}^{(j)}$ corresponding to the unknown state $\mathbf{x}_t^{(j)}$ be contained in the observable subspace $\mathcal{S}(\mathcal{V})$ corresponding to observations obtained by all the nodes. Note that $\mathcal{S}(\mathcal{V})$ is determined only by the sensor gain matrices and the matrix $\boldsymbol{A}$ affecting the evolution of each node's state.

Subcondition (ii) of Theorem 1 is on the communication capability of the network. In particular, the communication subcondition requires that the total Shannon capacity of the channels from nodes in $\mathcal{N}_\mathrm{s}^{(j)}$ to node j be above a threshold specified by the right-hand side of (28). This threshold is determined by two factors: each eigenvalue λ of dynamic matrix $\boldsymbol{A}^{(j)}$ and the value of $r_\lambda^{(j)}\big(\mathcal{N}_\mathrm{s}^{(j)}\big)$. In particular, λ determines the dynamics of node j's state. As λ increases, the variation of node j's state between two time steps becomes more significant, and thus node j needs more information to keep track of such variation. Consequently, the threshold on the total Shannon capacity becomes larger.

The quantity $r_\lambda^{(j)}\big(\mathcal{N}_\mathrm{s}^{(j)}\big)$ affecting the threshold in Subcondition (ii) is interpreted as follows. Consider the inference of $\mathbf{x}_t^{(j)}$ by the auxiliary processor described in the proof outline for Theorem 1. Note that this processor has access to more data than node j and thus can construct an estimator of $\mathbf{x}_t^{(j)}$ with smaller error compared to the distributed estimator at node j. Recall that inferring $\mathbf{x}_t^{(j)}$ is equivalent to estimating the projection of $\mathbf{x}_t$ onto node j's state subspace $\mathcal{X}^{(j)}$, which is $\boldsymbol{A}^\mathrm{T}$-invariant. Applying Proposition 1 as well as using the relationship between $\Lambda(\boldsymbol{A}^\mathrm{T})$ and $\Lambda(\boldsymbol{A}^{(j)})$, subspace $\mathcal{X}^{(j)}$ can be decomposed as

$$\mathcal{X}^{(j)} = \bigoplus_{\lambda\in\Lambda(\boldsymbol{A}^{(j)})} \big(\mathcal{X}^{(j)} \cap \mathcal{M}_\lambda(\boldsymbol{A}^\mathrm{T})\big).$$

In other words, $\mathcal{X}^{(j)}$ is decomposed into its intersection with each real generalized eigenspace of $\boldsymbol{A}^\mathrm{T}$. In order to estimate $\mathbf{x}_t^{(j)}$, the processor needs to learn the projection of $\mathbf{x}_t$ onto the intersection $\mathcal{X}^{(j)} \cap \mathcal{M}_\lambda(\boldsymbol{A}^\mathrm{T})$ for each $\lambda \in \Lambda(\boldsymbol{A}^{(j)})$. Recall that the auxiliary processor can learn such projection from two disjoint sets of nodes: $\mathcal{N}_\mathrm{s}^{(j)}$ and $\mathcal{V}_j\big(\mathcal{N}_\mathrm{s}^{(j)}\big)$. In particular, from the second set $\mathcal{V}_j\big(\mathcal{N}_\mathrm{s}^{(j)}\big)$, the processor can learn the projection of $\mathbf{x}_t$ onto $\mathcal{S}\big(\mathcal{V}_j\big(\mathcal{N}_\mathrm{s}^{(j)}\big)\big)\cap\mathcal{X}^{(j)}\cap\mathcal{M}_\lambda(\boldsymbol{A}^\mathrm{T})$, namely the intersection of the observable subspace corresponding to observations obtained by nodes in $\mathcal{V}_j\big(\mathcal{N}_\mathrm{s}^{(j)}\big)$, node j's state subspace, and the real generalized eigenspace corresponding to eigenvalue λ. However, in order to learn the projection of $\mathbf{x}_t$ onto the complementary subspace of $\mathcal{S}\big(\mathcal{V}_j\big(\mathcal{N}_\mathrm{s}^{(j)}\big)\big)\cap\mathcal{X}^{(j)}\cap\mathcal{M}_\lambda(\boldsymbol{A}^\mathrm{T})$ with respect to $\mathcal{X}^{(j)}\cap\mathcal{M}_\lambda(\boldsymbol{A}^\mathrm{T})$, the processor has to employ data from the first set of nodes $\mathcal{N}_\mathrm{s}^{(j)}$. The dimension of this complementary subspace is $r_\lambda^{(j)}\big(\mathcal{N}_\mathrm{s}^{(j)}\big)$. Consequently, the processor needs more data from $\mathcal{N}_\mathrm{s}^{(j)}$ if $r_\lambda^{(j)}\big(\mathcal{N}_\mathrm{s}^{(j)}\big)$ increases. The data from $\mathcal{N}_\mathrm{s}^{(j)}$ consist of messages received by node j from nodes in $\mathcal{N}_\mathrm{s}^{(j)}$. Consequently, the capacities of the channels from nodes in $\mathcal{N}_\mathrm{s}^{(j)}$ to node j are required to be larger if $r_\lambda^{(j)}\big(\mathcal{N}_\mathrm{s}^{(j)}\big)$ increases. This explains the effect of $r_\lambda^{(j)}\big(\mathcal{N}_\mathrm{s}^{(j)}\big)$ on the threshold for the total Shannon capacity in Subcondition (ii).

The quantity $r_\lambda^{(j)}\big(\mathcal{N}_\mathrm{s}^{(j)}\big)$ is determined by the sensing capabilities of nodes in $\mathcal{V}_j\big(\mathcal{N}_\mathrm{s}^{(j)}\big)$: this value is smaller if the sensing capabilities of these nodes are improved. Specifically, it can be seen from (29) that $r_\lambda^{(j)}\big(\mathcal{N}_\mathrm{s}^{(j)}\big)$ is non-increasing as the observable subspace $\mathcal{S}\big(\mathcal{V}_j\big(\mathcal{N}_\mathrm{s}^{(j)}\big)\big)$ becomes larger, i.e., the sensing capabilities of nodes in $\mathcal{V}_j\big(\mathcal{N}_\mathrm{s}^{(j)}\big)$ are improved. Recall that the auxiliary processor relies on two information sources: observations obtained by nodes in $\mathcal{V}_j\big(\mathcal{N}_\mathrm{s}^{(j)}\big)$ and messages received from nodes in $\mathcal{N}_\mathrm{s}^{(j)}$. As the sensing capabilities of nodes in $\mathcal{V}_j\big(\mathcal{N}_\mathrm{s}^{(j)}\big)$ improve, the observations obtained by these nodes become more informative, and thus the auxiliary processor relies less on messages received from nodes in $\mathcal{N}_\mathrm{s}^{(j)}$. As a result, the requirement on the quality of channels from nodes in $\mathcal{N}_\mathrm{s}^{(j)}$ to node j is less stringent, which allows the total Shannon capacity of channels from nodes in $\mathcal{N}_\mathrm{s}^{(j)}$ to node j to be smaller.

Remark 1: Theorem 1 provides insights for the design of communication-efficient distributed learning in complex networked systems. Specifically, if (27) or (28) does not hold, then the total distributed inference MSE will be unbounded over time irrespective of the encoding strategies employed by the network. To avoid this, more resources need to be provided to the network to improve its sensing and communication capabilities. For example, if (28) does not hold for a particular j and $\mathcal{N}_\mathrm{s}^{(j)}$, then either additional sensing resources should be allocated to nodes in $\mathcal{V}_j\big(\mathcal{N}_\mathrm{s}^{(j)}\big)$, or additional communication resources should be provided to nodes in $\mathcal{N}_\mathrm{s}^{(j)}$. Specifically, additional sensing resources allocated to nodes in $\mathcal{V}_j\big(\mathcal{N}_\mathrm{s}^{(j)}\big)$ increase the information in observations made by these nodes, whereas additional communication resources for nodes in $\mathcal{N}_\mathrm{s}^{(j)}$ increase the information in messages transmitted by these nodes. Theorem 1 helps with the identification of the network's bottleneck as well as the allocation of sensing and communication resources.

In addition to the boundedness of MSE in the distributed learning problem, another important aspect is to compute the MSE or derive a tight bound for it. This is more challenging compared to studying the inference error for centralized learning. In order to derive a tight error bound for distributed learning, we not only need to account for the system disturbance and observation uncertainty, but also need to investigate the optimal real-time encoding strategy for each node in the network. Such strategies are typically difficult to find, especially for a network with multiple nodes and a general topology. On the other hand, intuition on reducing the MSE for distributed estimators can be obtained from the derived necessary condition. One example is the allocation of sensing and communication resources for nodes in $\mathcal{V}_j\big(\mathcal{N}_\text{s}^{(j)}\big)$ and $\mathcal{N}_\text{s}^{(j)}$, as described in the previous paragraph.

V. Comparison of Necessary and Sufficient Conditions

This section compares the necessary condition for distributed inference established in this paper with the sufficient condition presented in a companion paper [68] and provides a case study for demonstrating these conditions.

A. Comparison Between the Necessary Condition and the Sufficient Condition

The sufficient condition for F_t to be bounded over time established in [68, Theorem 1] is presented here for self containment. The sufficient condition requires the notion of ordered trees based on graph $\mathscr{G}_\text{u}$. Specifically, an ordered tree $\mathcal{T}$ is constructed by first finding a spanning tree of $\mathscr{G}_\text{u}$, then assigning a node as the root of the spanning tree, and finally specifying an order for all the children of each node. Let $\mathcal{N}^{(j)}$ represent the set of neighbors for node j in $\mathcal{T}$ and define $N^{(j)} := \big|\mathcal{N}^{(j)}\big|$.[1] Moreover, let $\breve{c}_n^{(j)}$ represent the nth child of node j for $n = 1, 2, \ldots, N^{(j)} - 1$. If node j is the root of $\mathcal{T}$, then $\breve{c}_{N^{(j)}}^{(j)}$ represents the $N^{(j)}$th child of node j. Otherwise $\breve{c}_{N^{(j)}}^{(j)}$ represents the parent of node j. The sufficient condition is presented in the next theorem.

Theorem 2: Consider the distributed learning problem presented in Section II. If Assumptions A1-A5 in [68] hold,[2] then a sufficient condition for $\sup_{t\geqslant 0} F_t < \infty$ is given as follows: there exists an ordered tree $\mathcal{T}$ based on $\mathscr{G}_\text{u}$ such that both of the following two subconditions hold:

(i) Relationship (27) holds for every $j \in \mathcal{V}_\text{a}$.

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

$$\breve{C}^{(ij)}\big(\alpha^{(j)}\big) \;>\; \sum_{\lambda\in\mathit{\Lambda}(\boldsymbol{A}^{(j)})} \dim\!\big(\mathcal{I}_{\boldsymbol{A}^\text{T}}\big(\mathcal{G}_\lambda^{(ij)}\big)\big)\log|\lambda| =: \gamma_{\mathcal{T}}^{(ij)} \qquad \forall j\in\mathcal{V},\, i\in\mathcal{N}^{(j)} \tag{40}$$

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

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

where D is the diameter of $\mathcal{T}$.[3] Here, $\mathcal{G}_\lambda^{(ij)}$ is a subspace of $\mathbb{R}^{dv}$ whose definition is given in [68]. In particular, $\mathcal{G}_\lambda^{(ij)}$ satisfies

$$\sum_{i\in\mathcal{N}_n^{(j)}}\dim\!\big(\mathcal{G}_\lambda^{(ij)}\big) = r_\lambda^{(j)}\big(\mathcal{N}_n^{(j)}\big) \qquad \forall j\in\mathcal{V}_\text{a},\; n\in\big\{1,2,\ldots,N^{(j)}\big\} \tag{41}$$

where $\mathcal{N}_n^{(j)} \subseteq \mathcal{N}^{(j)}$ and is defined as $\mathcal{N}_n^{(j)} := \big\{\breve{c}_n^{(j)}, \breve{c}_{n+1}^{(j)}, \ldots, \breve{c}_{N^{(j)}}^{(j)}\big\}$, and $r_\lambda^{(j)}(\cdot)$ is defined as (29). The operator $\mathcal{I}$ in (40) is defined in (9).

Both the necessary condition and sufficient condition consist of two parts. The first part is on the sensing capability of the network and it is the same for both the necessary condition and the sufficient condition. The second part is on the communication capability of the network, where the necessary condition and the sufficient condition differ. Specifically, for each agent node j, the necessary condition specifies a region for the Shannon capacities of channels from all its neighbors, whereas the sufficient condition specifies a region for the anytime capacities of these channels. The gap between these two regions satisfies a favorable property in certain scenarios, which is shown next.

Consider the vector $\boldsymbol{C}^{(j)} \in \mathbb{R}^{N^{(j)}}$ that represents the Shannon capacities of the channels to agent node j from all its neighbors. Mathematically,

$$\boldsymbol{C}^{(j)} := \Big[C^{(i_1\,j)} \quad C^{(i_2\,j)} \quad \cdots \quad C^{(i_{N^{(j)}}\,j)}\Big]^\text{T}$$

where $i_1, i_2, \ldots, i_{N^{(j)}}$ are all the neighbors of node j, i.e., $\mathcal{N}^{(j)} = \big\{i_1, i_2, \ldots, i_{N^{(j)}}\big\}$.[4] Then (28) in Theorem 1, together with the constraint that $C^{(ij)} \geqslant 0$, specifies a region $\mathcal{R}^{(j)} \subseteq \mathbb{R}^{N^{(j)}}$ where $\boldsymbol{C}^{(j)}$ must belong to if the total distributed inference MSE is bounded over time. Specifically, $\mathcal{R}^{(j)}$ is given by (42), as shown at the bottom of the page.

Analogously, consider the vector $\breve{\boldsymbol{C}}^{(j)}\big(\alpha^{(j)}\big)$ that represents the $\alpha^{(j)}$-anytime capacities of the channels to node j from all its neighbors, i.e.,

$$\breve{\boldsymbol{C}}^{(j)}\big(\alpha^{(j)}\big) := \Big[\breve{C}^{(i_1\,j)}\big(\alpha^{(j)}\big) \quad \breve{C}^{(i_2\,j)}\big(\alpha^{(j)}\big) \quad \cdots \quad \breve{C}^{(i_{N^{(j)}}\,j)}\big(\alpha^{(j)}\big)\Big]^\text{T}.$$

Then (40) specifies a region for $\breve{\boldsymbol{C}}^{(j)}\big(\alpha^{(j)}\big)$ that ensures bounded total distributed inference MSE. Specifically, for any

$$\mathcal{R}^{(j)} := \Big\{\boldsymbol{c} = \big[c^{(i_1\,j)} \quad c^{(i_2\,j)} \quad \cdots \quad c^{(i_{N^{(j)}}\,j)}\big]^\text{T} : c^{(ij)} \geqslant 0 \quad \forall i\in\mathcal{N}^{(j)},\ \text{and} \quad \sum_{i\in\mathcal{N}_\text{s}^{(j)}} c^{(ij)} \;>\; \sum_{\lambda\in\mathit{\Lambda}(\boldsymbol{A}^{(j)})} r_\lambda^{(j)}\big(\mathcal{N}_\text{s}^{(j)}\big)\log|\lambda| \quad \forall\mathcal{N}_\text{s}^{(j)}\subseteq\mathcal{N}^{(j)}\ \text{s.t.}\ \mathcal{X}^{(j)}\not\subseteq\mathcal{S}\big(\mathcal{V}_j\big(\mathcal{N}_\text{s}^{(j)}\big)\big)\Big\} \tag{42}$$

[1] If there is no cycle in $\mathscr{G}_\text{u}$, then a spanning tree of $\mathscr{G}_\text{u}$ is $\mathscr{G}_\text{u}$ itself, and $\mathcal{N}^{(j)} = \mathcal{N}_\text{u}^{(j)}$.

[2] Assumptions A1, A2, and A4 in [68] are the same as those in this paper, whereas Assumptions A3 and A5 are different.

[3] The diameter of a tree is defined as the largest of shortest-path distances between all pairs of nodes in the tree [77, Chapter 22].

[4] Note that $\breve{c}_1^{(j)}, \breve{c}_2^{(j)}, \ldots, \breve{c}_{N^{(j)}}^{(j)}$ is a permutation of $i_1, i_2, \ldots, i_{N^{(j)}}$, where the permutation is determined by the specification of the ordered tree $\mathcal{T}$.

ordered tree $\mathcal{T}$ based on $\mathscr{G}_{\mathrm{u}}$, define

$$\breve{\mathcal{R}}_{\mathcal{T}}^{(j)} := \Big\{ \boldsymbol{c} = \Big[c^{(i_1 j)} \quad c^{(i_2 j)} \quad \cdots \quad c^{(i_{N^{(j)}}\, j)}\Big]^{\mathrm{T}} : c^{(ij)} > \gamma_{\mathcal{T}}^{(ij)} \quad \forall i \in \mathcal{N}^{(j)} \Big\} \tag{43}$$

where $\gamma_{\mathcal{T}}^{(ij)}$ is defined in (40). If $\breve{\boldsymbol{C}}^{(j)}\big(\alpha^{(j)}\big)$ is within region $\breve{\mathcal{R}}_{\mathcal{T}}^{(j)}$ and Subcondition (i) of Theorem 2 holds, then encoding strategies can be designed for each node in the network so that the total distributed inference MSE of agents is bounded over time.

In general, there is a gap between the capacity region $\mathcal{R}^{(j)}$ for the necessary condition and region $\breve{\mathcal{R}}_{\mathcal{T}}^{(j)}$ for the sufficient condition. To characterize such a gap, some definitions are introduced here. Define matrix $\breve{\boldsymbol{A}}^{(ij)} \in \mathbb{R}^{(dv)\times(dv)}$ as

$$\breve{\boldsymbol{A}}^{(ij)} := \big(\boldsymbol{e}_{i,v}\boldsymbol{e}_{i,v}^{\mathrm{T}}\big) \otimes \boldsymbol{A}^{(i)} + \big(\boldsymbol{e}_{j,v}\boldsymbol{e}_{j,v}^{\mathrm{T}}\big) \otimes \boldsymbol{A}^{(j)}. \tag{44}$$

In other words, $\breve{\boldsymbol{A}}^{(ij)}$ is a block-diagonal matrix with the ith and jth block on its main diagonal being $\boldsymbol{A}^{(i)}$ and $\boldsymbol{A}^{(j)}$, respectively, and other blocks being zero matrices. In addition, let $\bar{\mathcal{R}}^{(j)}$ represent the closure of $\mathcal{R}^{(j)}$. Such a closure is a polyhedron since all constraints in (42) are linear [78, Chapter 2]. Moreover, let

$$\boldsymbol{\gamma}_{\mathcal{T}}^{(j)} := \Big[\gamma_{\mathcal{T}}^{(i_1 j)} \quad \gamma_{\mathcal{T}}^{(i_2 j)} \quad \cdots \quad \gamma_{\mathcal{T}}^{(i_{N^{(j)}}\, j)}\Big]^{\mathrm{T}} \tag{45}$$

represent the vector that determines $\breve{\mathcal{R}}_{\mathcal{T}}^{(j)}$. The next proposition presents scenarios in which $\boldsymbol{\gamma}_{\mathcal{T}}^{(j)}$ is an extreme point of $\bar{\mathcal{R}}^{(j)}$.

Proposition 2: Consider the scenario where there is no cycle in the graph $\mathscr{G}_{\mathrm{u}}$. For any ordered tree $\mathcal{T}$ based on $\mathscr{G}_{\mathrm{u}}$, if the equality $\mathcal{I}_{(\breve{\boldsymbol{A}}^{(ij)})^{\mathrm{T}}}\big(\mathcal{G}_{\lambda}^{(ij)}\big) = \mathcal{G}_{\lambda}^{(ij)}$ holds for all $i \in \mathcal{N}^{(j)}$ and $\lambda \in \Lambda(\boldsymbol{A}^{(j)})$, then $\boldsymbol{\gamma}_{\mathcal{T}}^{(j)}$ is an extreme point of $\bar{\mathcal{R}}^{(j)}$.

Proof: See Appendix C. □

The equality $\mathcal{I}_{(\breve{\boldsymbol{A}}^{(ij)})^{\mathrm{T}}}\big(\mathcal{G}_{\lambda}^{(ij)}\big) = \mathcal{G}_{\lambda}^{(ij)}$ in Proposition 2 holds if, for example, both $\boldsymbol{A}^{(i)}$ and $\boldsymbol{A}^{(j)}$ are diagonalizable and only have real eigenvalues.

Remark 2: It is a favorable property that $\boldsymbol{\gamma}_{\mathcal{T}}^{(j)}$ is an extreme point of $\bar{\mathcal{R}}^{(j)}$. To see this, consider the scenario where the vector $\boldsymbol{C}^{(j)}$ of Shannon capacities and vector $\breve{\boldsymbol{C}}^{(j)}\big(\alpha^{(j)}\big)$ of anytime capacities coincide. This is the case, for example, if node j receives messages from its neighbors via noiseless digital channels. Moreover, consider the case where $\boldsymbol{C}^{(j)}$ equals the vector $\boldsymbol{\gamma}_{\mathcal{T}}^{(j)}$. According to the sufficient condition shown in Theorem 2, if Subcondition (i) is satisfied, then node j can design a distributed estimator whose MSE is bounded over time. However, such an estimator would be impossible to design if any entry of $\boldsymbol{C}^{(j)}$ is reduced by a positive amount and the other entries do not change. In other words, we cannot reduce the capacity of the channel to node j from one of its neighbors without increasing the capacities of channels from its other neighbors. This is because the capacity vector obtained by reducing one of the entries in $\boldsymbol{C}^{(j)}$ no longer belongs to the region $\mathcal{R}^{(j)}$, since $\boldsymbol{C}^{(j)} = \boldsymbol{\gamma}_{\mathcal{T}}^{(j)}$ is an extreme point for the closure $\bar{\mathcal{R}}^{(j)}$ of $\mathcal{R}^{(j)}$. Consequently, the necessary condition is not satisfied, and it is impossible to design a distributed estimator whose MSE is bounded over time.

An example comparing the capacity regions for the necessary condition and for the sufficient condition is presented in Section V-B.

B. Case Study

Consider a network of three nodes 1, 2, and 3, where node 1 is within the communication ranges of the other two nodes. Consequently, the vertex set and edge set corresponding to this network are given by $\mathcal{V} = \{1,2,3\}$ and $\mathcal{E}_{\mathrm{u}} = \{(1,2),(1,3)\}$, respectively. The dimensionality of each node's unknown state is $d = 3$. Consequently, the concatenated state $\mathbf{x}_t \in \mathbb{R}^9$. Moreover, node 1 is the only agent and its state subspace is $\mathcal{X}^{(1)} = \mathcal{C}([\boldsymbol{e}_{1,9} \quad \boldsymbol{e}_{2,9} \quad \boldsymbol{e}_{3,9}])$. Matrix $\boldsymbol{A}^{(j)}$ in (1) is set to

$$\boldsymbol{A}^{(j)} = \operatorname{diag}\{2,2,3\} \quad \text{for } j \in \{1,2,3\}. \tag{46}$$

The sensor gain matrices of node 1 are given by $\boldsymbol{\Gamma}^{(11)} = \boldsymbol{0}$ and $\boldsymbol{\Gamma}_1^{(1i)} = \boldsymbol{\Gamma}_2^{(1i)} = \boldsymbol{0}$ for $i \in \{2,3\}$. In this case, the observations of agent 1 contain only noise and this agent must rely on messages received from nodes 2 and 3 for leaning its state. The sensor gain matrices of nodes 2 and 3 are given by

$$\boldsymbol{\Gamma}^{(22)} = \boldsymbol{I}, \quad \boldsymbol{\Gamma}_1^{(21)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad \boldsymbol{\Gamma}_2^{(21)} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \end{bmatrix} \tag{47a}$$

$$\boldsymbol{\Gamma}^{(33)} = \boldsymbol{I}, \quad \boldsymbol{\Gamma}_1^{(31)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \boldsymbol{\Gamma}_2^{(31)} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix}. \tag{47b}$$

First, the necessary condition given by Theorem 1 is presented. Combining (20), (23), (46), and (47) gives $\mathcal{S}(\mathcal{V}) = \mathbb{R}^9$. Therefore, $\mathcal{X}^{(1)} \subseteq \mathcal{S}(\mathcal{V})$, showing that Subcondition (i) in Theorem 1 holds for $j = 1$. Subcondition (ii) in Theorem 1 indicates that the vector of Shannon capacity $\boldsymbol{C}^{(1)} = \big[C^{(21)} \quad C^{(31)}\big]^{\mathrm{T}}$ must belong to the region $\mathcal{R}^{(1)}$ given by (42). Choose $\mathcal{N}_{\mathrm{s}}^{(1)}$ as $\mathcal{N}_{\mathrm{s}}^{(1)} = \{2\}$, and thus $\mathcal{V}_1\big(\mathcal{N}_{\mathrm{s}}^{(1)}\big) = \{1,3\}$. Combining (46) and (47) gives

$$\mathcal{S}(\{1,3\}) = \mathcal{S}(\{3\}) = \mathcal{C}([\boldsymbol{e}_{1,9} \quad \boldsymbol{e}_{3,9} \quad \boldsymbol{e}_{7,9} \quad \boldsymbol{e}_{8,9} \quad \boldsymbol{e}_{9,9}]) \tag{48}$$

and thus $\mathcal{X}^{(1)} \not\subseteq \mathcal{S}(\{1,3\})$. Moreover, (46) shows that

$$\mathcal{X}^{(1)} \cap \mathcal{M}_2(\boldsymbol{A}^{\mathrm{T}}) = \mathcal{C}([\boldsymbol{e}_{1,9} \quad \boldsymbol{e}_{2,9}])$$
$$\mathcal{X}^{(1)} \cap \mathcal{M}_3(\boldsymbol{A}^{\mathrm{T}}) = \mathcal{C}(\boldsymbol{e}_{3,9}).$$

Therefore,

$$\mathcal{S}(\{1,3\}) \cap \mathcal{X}^{(1)} \cap \mathcal{M}_2(\boldsymbol{A}^{\mathrm{T}}) = \mathcal{C}(\boldsymbol{e}_{1,9}) \tag{49a}$$

$$\mathcal{S}(\{1,3\}) \cap \mathcal{X}^{(1)} \cap \mathcal{M}_3(\boldsymbol{A}^{\mathrm{T}}) = \mathcal{C}(\boldsymbol{e}_{3,9}). \tag{49b}$$

Substituting (49a), $\mathcal{V}_1\big(\mathcal{N}_{\mathrm{s}}^{(1)}\big) = \{1,3\}$, and $\dim\big(\mathcal{X}^{(1)} \cap \mathcal{M}_2(\boldsymbol{A}^{\mathrm{T}})\big) = 2$ into (29) gives $r_2^{(j)}(\{2\}) = 1$. Similar calculation gives $r_3^{(j)}(\{2\}) = 0$. Therefore, the constraint in (42) for $\mathcal{N}_{\mathrm{s}}^{(1)} = \{2\}$ becomes $c^{(21)} > \log 2 = 1$. In other words, the Shannon capacity of the channel from node 2 to node 1 is required to be larger than 1 bit per channel use. This is because node 1 relies on messages received from node 2 for inferring the second entry $\big[\mathbf{x}_t^{(1)}\big]_2$ of its unknown state $\mathbf{x}_t^{(1)}$. To see this, note from (48) that $\mathcal{S}(\{1,3\}) \cap \mathcal{X}^{(1)} = \mathcal{C}([\boldsymbol{e}_{1,9} \quad \boldsymbol{e}_{3,9}])$. This shows that all the observations

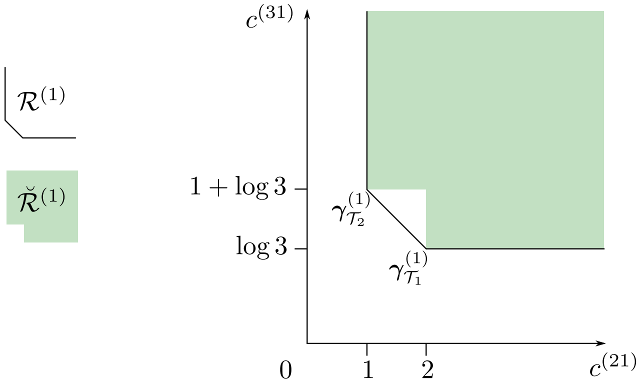


Fig. 4. Regions of channel capacities for the necessary and sufficient conditions: the region $\mathcal{R}^{(1)}$ for the necessary condition is the area to the upper-right of the line segments, whereas the region $\breve{\mathcal{R}}^{(1)}$ for the sufficient condition is the shaded area. Vectors $\gamma_{\mathcal{T}_1}^{(1)}$ and $\gamma_{\mathcal{T}_2}^{(1)}$ determining $\breve{\mathcal{R}}^{(1)}$ are extreme points of the closure of $\mathcal{R}^{(1)}$.

obtained by nodes 1 and 3 provide information only of $[\boldsymbol{e}_{1,9}\quad \boldsymbol{e}_{3,9}]^{\mathrm{T}}\mathbf{x}_t = \big[[\mathbf{x}_t^{(1)}]_1\quad [\mathbf{x}_t^{(1)}]_3\big]^{\mathrm{T}}$, namely the first and third entries of node 1's state. Consequently, node 1 relies on messages received from node 2 for inferring $[\mathbf{x}_t^{(1)}]_2$, which leads to the requirement on the capacity for the channel from node 2 to node 1. Similarly, setting $\mathcal{N}_{\mathrm{s}}^{(1)} = \{3\}$ gives $c^{(31)} > \log 3$, as node 1 relies on messages received from node 3 for inferring $[\mathbf{x}_t^{(1)}]_3$. This can be seen by noting $\mathcal{S}(\{1,2\}) \cap \mathcal{X}^{(1)} = \mathcal{C}([\boldsymbol{e}_{1,9}\quad \boldsymbol{e}_{2,9}])$. Finally, setting $\mathcal{N}_{\mathrm{s}}^{(1)} = \{2,3\}$ gives $c^{(21)} + c^{(31)} > 2 + \log 3$. Therefore, $\mathcal{R}^{(1)}$ is given by

$$\mathcal{R}^{(1)} := \Big\{\big[c^{(21)}\quad c^{(31)}\big]^{\mathrm{T}} : c^{(21)} > 1,\ c^{(31)} > \log 3, \\ c^{(21)} + c^{(31)} > 2 + \log 3\Big\}. \tag{50}$$

Region $\mathcal{R}^{(1)}$ is the area to the upper-right of the line segments in Fig. 4. The closure $\bar{\mathcal{R}}^{(1)}$ of $\mathcal{R}^{(1)}$ is obtained by substituting each greater-than sign in (50) with a greater-than-or-equal-to sign and $\bar{\mathcal{R}}^{(1)}$ is a polygon.

For the sufficient condition given by Theorem 2, it was shown in [68] that if the vector of anytime capacity $\breve{\boldsymbol{C}}^{(1)}(\alpha^{(1)}) := \big[\breve{C}^{(21)}(\alpha^{(1)})\quad \breve{C}^{(31)}(\alpha^{(1)})\big]^{\mathrm{T}}$ belongs to $\breve{\mathcal{R}}^{(1)} := \breve{\mathcal{R}}_{\mathcal{T}_1}^{(1)} \cup \breve{\mathcal{R}}_{\mathcal{T}_2}^{(1)}$ for $\alpha^{(1)} = 32\log 3$, then the network has sufficient communication capabilities to ensure that the distributed MSE of agent 1 is bounded over time. Specifically, $\breve{\mathcal{R}}_{\mathcal{T}_l}^{(1)}$ is given by

$$\breve{\mathcal{R}}_{\mathcal{T}_l}^{(1)} := \Big\{\big[c^{(21)}\quad c^{(31)}\big]^{\mathrm{T}} : c^{(21)} > \gamma_{\mathcal{T}_l}^{(21)},\ c^{(31)} > \gamma_{\mathcal{T}_l}^{(31)}\Big\} \\ \text{for } l \in \{1,2\}$$

where $\gamma_{\mathcal{T}_1}^{(21)} = 2$, $\gamma_{\mathcal{T}_1}^{(31)} = \log 3$, $\gamma_{\mathcal{T}_2}^{(21)} = 1$, and $\gamma_{\mathcal{T}_2}^{(31)} = 1 + \log 3$. Vectors $\gamma_{\mathcal{T}_1}^{(1)} := \big[\gamma_{\mathcal{T}_1}^{(21)}\quad \gamma_{\mathcal{T}_1}^{(31)}\big]^{\mathrm{T}}$ and $\gamma_{\mathcal{T}_2}^{(1)} := \big[\gamma_{\mathcal{T}_2}^{(21)}\quad \gamma_{\mathcal{T}_2}^{(31)}\big]^{\mathrm{T}}$ are also shown in the figure and they are extreme points of $\bar{\mathcal{R}}^{(1)}$. The gap between $\mathcal{R}^{(1)}$ and $\breve{\mathcal{R}}^{(1)}$ is the white triangle in Fig. 4 whose vertices are $\gamma_{\mathcal{T}_1}^{(1)}$, $\gamma_{\mathcal{T}_2}^{(1)}$, and $\big[2\quad 1+\log 3\big]^{\mathrm{T}}$. Note that the gap would be larger if $\gamma_{\mathcal{T}_1}^{(1)}$ and $\gamma_{\mathcal{T}_2}^{(1)}$ are not extreme points of $\bar{\mathcal{R}}^{(1)}$.

We comment on the resemblance of the region (50) to the rate region for a distributed source coding problem described as follows. Consider two correlated data sources 1 and 2. In particular, source 1 generates a sequence $\mathbf{s}_{1:n}^{(1)}$ consisting of n independent, identically distributed (IID) random variables $\mathrm{s}_1^{(1)}, \mathrm{s}_2^{(1)}, \ldots, \mathrm{s}_n^{(1)}$. A source encoder at source 1 generates a codeword according to $\mathbf{s}_{1:n}^{(1)}$ at rate $r^{(1)}$, namely the codeword is represented by $nr^{(1)}$ bits in total. Analogously, source 2 generates a sequence $\mathbf{s}_{1:n}^{(2)}$ consisting of IID random variables $\mathrm{s}_1^{(2)}, \mathrm{s}_2^{(2)}, \ldots, \mathrm{s}_n^{(2)}$, and another source encoder generates a codeword at rate $r^{(2)}$ according to $\mathbf{s}_{1:n}^{(2)}$. A decoder estimates both sequences $\mathbf{s}_{1:n}^{(1)}$ and $\mathbf{s}_{1:n}^{(2)}$ using the codewords generated by both source encoders. If at least one of the two estimated sequences does not equal its actual value, then the decoder is said to have made an error. An important question is under what conditions the error probability vanishes as n approaches infinity. This problem has been solved in the literature. In particular, a necessary condition is that the rate $\big[r^{(1)}\quad r^{(2)}\big]^{\mathrm{T}}$ satisfies the following constraints [79], [80]

$$r^{(1)} \geqslant H\big(\mathrm{s}_1^{(1)}\,\big|\,\mathrm{s}_1^{(2)}\big), \quad r^{(2)} \geqslant H\big(\mathrm{s}_1^{(2)}\,\big|\,\mathrm{s}_1^{(1)}\big) \tag{51a}$$

$$r^{(1)} + r^{(2)} \geqslant H\big(\mathrm{s}_1^{(1)}, \mathrm{s}_1^{(2)}\big) \tag{51b}$$

where $H(\mathrm{x}\,|\,\mathrm{y})$ represents the conditional entropy of x given y for two general discrete random variables x and y, whereas $H(\mathrm{x}, \mathrm{y})$ represents the entropy of x and y. Note that the above constraints resemble the constraints that determine the region $\mathcal{R}^{(1)}$ in (50) and its closure $\bar{\mathcal{R}}^{(1)}$ for the distributed learning problem, with $r^{(1)}$ and $r^{(2)}$ corresponding to $c^{(21)}$ and $c^{(31)}$, respectively. Furthermore, the constraints (51) were shown by Slepian and Wolf to be not only necessary but also sufficient [80]. Therefore, there is no gap between the rate regions for the necessary condition and for the sufficient condition in the distributed source coding problem. Of course, the distributed source coding problem is significantly different from the distributed learning problem in this paper, which involves real-time sensing, communication, and inference in a network with multiple nodes.

VI. Conclusion

The paper derived a necessary condition under which the total distributed inference MSE of all the agents is bounded over time. The necessary condition consists of two subconditions: one is on the sensing capabilities and the other is on the communication capabilities of the network. The paper compares the established necessary condition with the sufficient condition presented in a companion paper, which also consists of a subcondition on sensing capabilities and a subcondition on communication capabilities. In particular, the subcondition on the sensing capabilities is the same for the necessary condition and the sufficient condition. For the subcondition on the communication capabilities, the vectors determining the capacity region for the sufficient condition are extreme points of the capacity region for the necessary condition in certain scenarios. The paper provides insights for the design of accurate and communication-efficient distributed learning in multi-agent networks.

APPENDIX A
LEMMAS USED FOR THE PROOF OF THEOREM 1

This appendix proves Lemma 1 and presents a few lemmas used for proving Theorem 1. First, Lemma 1 is proved.

Proof: Extending Lemma 1 of [68], we can show that there is an estimator $\hat{\boldsymbol{\xi}}_t$ of $\boldsymbol{O}\big(\big[\mathring{\boldsymbol{\Gamma}}^{(j)}\big]_{j\in\mathcal{V}_0}, \boldsymbol{A}\big)\mathbf{x}_t$ satisfying

$$\sup_{t\geqslant 0} \mathbb{E}\Big\{\big\|\hat{\boldsymbol{\xi}}_t - \boldsymbol{O}\big(\big[\mathring{\boldsymbol{\Gamma}}^{(j)}\big]_{j\in\mathcal{V}_0}, \boldsymbol{A}\big)\mathbf{x}_t\big\|^2\Big\} < \infty\,. \tag{52}$$

Since $\mathcal{C}(\boldsymbol{H}) \subseteq \mathcal{S}(\mathcal{V}_0) = \mathcal{C}\Big(\boldsymbol{O}\big(\big[\mathring{\boldsymbol{\Gamma}}^{(j)}\big]_{j\in\mathcal{V}_0}, \boldsymbol{A}\big)^{\mathrm{T}}\Big)$, there exists a matrix $\boldsymbol{\Phi}$ such that $\boldsymbol{H} = \boldsymbol{O}\big(\big[\mathring{\boldsymbol{\Gamma}}^{(j)}\big]_{j\in\mathcal{V}_0}, \boldsymbol{A}\big)^{\mathrm{T}}\boldsymbol{\Phi}$. Define estimator $\hat{\boldsymbol{\beta}}_t := \boldsymbol{\Phi}^{\mathrm{T}}\hat{\boldsymbol{\xi}}_t$, and its MSE satisfies

$$\begin{aligned}&\mathbb{E}\Big\{\big\|\boldsymbol{\Phi}^{\mathrm{T}}\hat{\boldsymbol{\xi}}_t - \boldsymbol{H}^{\mathrm{T}}\mathbf{x}_t\big\|^2\Big\}\\ &\qquad\leqslant \big\|\boldsymbol{\Phi}^{\mathrm{T}}\big\|^2\,\mathbb{E}\Big\{\big\|\hat{\boldsymbol{\xi}}_t - \boldsymbol{O}\big(\big[\mathring{\boldsymbol{\Gamma}}^{(j)}\big]_{j\in\mathcal{V}_0}, \boldsymbol{A}\big)\mathbf{x}_t\big\|^2\Big\}\,.\end{aligned}$$

Combining this with (52) gives the desired result (24). □

Next, two lemmas used for the Proof of Theorem 1 are presented. The first lemma shows a property of direct sum.

Lemma 2 [68]: For subspaces $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_m$ and $\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_m$ 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 $\big(\sum_{i=1}^m \mathcal{S}_i\big) \cap \mathcal{U}_j = \mathcal{S}_j$ for all $j \in \{1, 2, \ldots, m\}$.

Corollary 1: For any real square matrix $\boldsymbol{F}$ as well as subspaces $\mathcal{V}_i$ such that $\mathcal{V}_i = \oplus_{\mu\in\Lambda(\boldsymbol{F})}\big(\mathcal{V}_i \cap \mathcal{M}_\mu(\boldsymbol{F})\big)$ for $i = 1, 2$, the following holds for all $\lambda \in \Lambda(\boldsymbol{F})$

$$(\mathcal{V}_1 + \mathcal{V}_2) \cap \mathcal{M}_\lambda(\boldsymbol{F}) = \big(\mathcal{V}_1 \cap \mathcal{M}_\lambda(\boldsymbol{F})\big) + \big(\mathcal{V}_2 \cap \mathcal{M}_\lambda(\boldsymbol{F})\big). \tag{53}$$

Proof: Without loss of generality, suppose $\Lambda(\boldsymbol{F}) = \{\lambda_1, \lambda_2, \ldots, \lambda_m\}$, and $\lambda = \lambda_j$ for some $1 \leqslant j \leqslant m$. By assumption,

$$\begin{aligned}\mathcal{V}_1 + \mathcal{V}_2 &= \Big(\sum_{i=1}^m \mathcal{V}_1 \cap \mathcal{M}_{\lambda_i}(\boldsymbol{F})\Big) + \Big(\sum_{i=1}^m \mathcal{V}_2 \cap \mathcal{M}_{\lambda_i}(\boldsymbol{F})\Big)\\ &= \sum_{i=1}^m \big(\mathcal{V}_1 \cap \mathcal{M}_{\lambda_i}(\boldsymbol{F})\big) + \big(\mathcal{V}_2 \cap \mathcal{M}_{\lambda_i}(\boldsymbol{F})\big)\,.\end{aligned}$$

Applying Lemma 2 with $\mathcal{S}_i = \big(\mathcal{V}_1 \cap \mathcal{M}_{\lambda_i}(\boldsymbol{F})\big) + \big(\mathcal{V}_2 \cap \mathcal{M}_{\lambda_i}(\boldsymbol{F})\big)$ and $\mathcal{U}_i = \mathcal{M}_{\lambda_i}(\boldsymbol{F})$ for all $i \in \{1, 2, \ldots, m\}$ gives the desired result. □

The second lemma is on the determinant of the product of matrices that satisfy certain conditions.

Lemma 3: Let $\boldsymbol{F} = \mathrm{diag}\{\boldsymbol{F}_1, \boldsymbol{F}_2\}$ represent an invertible real matrix with $\boldsymbol{F}_1 \in \mathbb{R}^{n_1\times n_1}$ and let $\mathcal{Y}$ be an $\boldsymbol{F}$-invariant subspace. Let $\boldsymbol{T}$ be a matrix whose columns form an orthonormal basis for the orthogonal complement of $\mathcal{Y}$ with respect to $\mathcal{Y} + \mathcal{C}\big([\boldsymbol{I}_{n_1} \ \ \mathbf{0}]^{\mathrm{T}}\big)$. Then

$$\big|\det(\boldsymbol{T}^{\mathrm{T}}\boldsymbol{F}\boldsymbol{T})\big| = \prod_{\lambda\in\Lambda(\boldsymbol{F}_1)} |\lambda|^{r(\lambda)}$$

where

$$\begin{aligned}r(\lambda) :=\ &\dim\big(\mathcal{M}_\lambda(\boldsymbol{F}_1)\big)\\ &- \dim\big(\mathcal{Y} \cap \mathcal{C}\big([\boldsymbol{I}_{n_1} \ \ \mathbf{0}]^{\mathrm{T}}\big) \cap \mathcal{M}_\lambda(\boldsymbol{F})\big)\,.\end{aligned} \tag{54}$$

Proof: Define $\mathcal{E} := \mathcal{C}\big([\boldsymbol{I}_{n_1} \ \ \mathbf{0}]^{\mathrm{T}}\big)$ and let $\boldsymbol{Y}$ represent a matrix whose columns form an orthonormal basis for $\mathcal{Y}$. Then $\boldsymbol{Y}_0 := [\boldsymbol{Y} \ \ \boldsymbol{T}]$ is a matrix whose columns form an orthonormal basis for $\mathcal{Y} + \mathcal{E}$. Since $\boldsymbol{T}^{\mathrm{T}}\boldsymbol{Y} = \mathbf{0}$ and $\mathcal{Y}$ is $\boldsymbol{F}$-invariant, it holds that $\boldsymbol{T}^{\mathrm{T}}\boldsymbol{F}\boldsymbol{Y} = \mathbf{0}$. Therefore, $\boldsymbol{Y}_0^{\mathrm{T}}\boldsymbol{F}\boldsymbol{Y}_0$ can be partitioned as

$$\boldsymbol{Y}_0^{\mathrm{T}}\boldsymbol{F}\boldsymbol{Y}_0 = \begin{bmatrix}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{F}\boldsymbol{Y} & \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{F}\boldsymbol{T}\\ \mathbf{0} & \boldsymbol{T}^{\mathrm{T}}\boldsymbol{F}\boldsymbol{T}\end{bmatrix}. \tag{55}$$

Recall that $\mathcal{Y}$ is $\boldsymbol{F}$-invariant. Moreover, $\mathcal{E}$ can be shown to be $\boldsymbol{F}$-invariant as $\boldsymbol{F}$ is block-diagonal. Consequently, $\mathcal{C}(\boldsymbol{Y}_0) = \mathcal{Y} + \mathcal{E}$ is also $\boldsymbol{F}$-invariant. Applying Lemma 5 in [68] gives $\big|\det(\boldsymbol{Y}_0^{\mathrm{T}}\boldsymbol{F}\boldsymbol{Y}_0)\big| = \prod_{\lambda\in\Lambda(\boldsymbol{F})}|\lambda|^{y_0(\lambda)}$ and $\big|\det(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{F}\boldsymbol{Y})\big| = \prod_{\lambda\in\Lambda(\boldsymbol{F})}|\lambda|^{y(\lambda)}$, where

$$\begin{aligned}y_0(\lambda) &:= \dim\big((\mathcal{Y} + \mathcal{E}) \cap \mathcal{M}_\lambda(\boldsymbol{F})\big)\\ y(\lambda) &:= \dim\big(\mathcal{Y} \cap \mathcal{M}_\lambda(\boldsymbol{F})\big)\,.\end{aligned}$$

Combining this with (55) gives

$$\big|\det(\boldsymbol{T}^{\mathrm{T}}\boldsymbol{F}\boldsymbol{T})\big| = \frac{|\det(\boldsymbol{Y}_0^{\mathrm{T}}\boldsymbol{F}\boldsymbol{Y}_0)|}{|\det(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{F}\boldsymbol{Y})|} = \prod_{\lambda\in\Lambda(\boldsymbol{F})}|\lambda|^{y_0(\lambda)-y(\lambda)}. \tag{56}$$

Since both $\mathcal{Y}$ and $\mathcal{E}$ are $\boldsymbol{F}$-invariant, applying Proposition 1 and Corollary 1 gives $(\mathcal{Y} + \mathcal{E}) \cap \mathcal{M}_\lambda(\boldsymbol{F}) = \big(\mathcal{Y} \cap \mathcal{M}_\lambda(\boldsymbol{F})\big) + \big(\mathcal{E} \cap \mathcal{M}_\lambda(\boldsymbol{F})\big)$. Combining this with the property that $\dim(\mathcal{Y}_1 + \mathcal{Y}_2) = \dim(\mathcal{Y}_1) + \dim(\mathcal{Y}_2) - \dim(\mathcal{Y}_1 \cap \mathcal{Y}_2)$ for two general subspaces $\mathcal{Y}_1$ and $\mathcal{Y}_2$ gives

$$\begin{aligned}y_0(\lambda) - y(\lambda) = \ &\dim\big(\mathcal{E} \cap \mathcal{M}_\lambda(\boldsymbol{F})\big)\\ &- \dim\big(\mathcal{Y} \cap \mathcal{E} \cap \mathcal{M}_\lambda(\boldsymbol{F})\big)\,.\end{aligned} \tag{57}$$

Since $\boldsymbol{F}$ is block diagonal, it holds $\Lambda(\boldsymbol{F}) = \Lambda(\boldsymbol{F}_1) \cup \Lambda(\boldsymbol{F}_2)$. Moreover, for any $\lambda \in \Lambda(\boldsymbol{F}) \setminus \Lambda(\boldsymbol{F}_1)$, it can be shown that $\mathcal{E} \cap \mathcal{M}_\lambda(\boldsymbol{F}) = \{\mathbf{0}\}$. Combining this with (54) and (57) gives

$$y_0(\lambda) - y(\lambda) = \begin{cases} r(\lambda) & \text{if } \lambda \in \Lambda(\boldsymbol{F}_1)\\ 0 & \text{otherwise.}\end{cases}$$

Substituting this into (56) gives the desired result. □

APPENDIX B
PROOF OF THEOREM 1

Proof: To facilitate the presentation of the proof, the following notation is introduced: let $\epsilon(\boldsymbol{\theta}; \boldsymbol{\varphi})$ represent the MMSE for estimating $\boldsymbol{\theta}$ using $\boldsymbol{\varphi}$, i.e.,

$$\epsilon(\boldsymbol{\theta}; \boldsymbol{\varphi}) := \mathbb{E}\Big\{\big\|\boldsymbol{\theta} - \mathbb{E}\{\boldsymbol{\theta} \,|\, \boldsymbol{\varphi}\}\big\|^2\Big\} \tag{58}$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$ are two general random vectors. A property of $\epsilon(\cdot; \cdot)$ is that $\epsilon(\boldsymbol{\theta}; \boldsymbol{\varphi}_1) \leqslant \epsilon(\boldsymbol{\theta}; \boldsymbol{\varphi}_2)$ if $\boldsymbol{\varphi}_1$ contains all the entries of $\boldsymbol{\varphi}_2$.

Consider the estimation of $\mathbf{x}_t^{(j)}$ using $\mathring{\mathbf{r}}_{0:t}^{(j)}\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)$, $\breve{\mathbf{z}}_{0:t}$, and $\breve{\mathbf{r}}_{0:t}$, where $\mathring{\mathbf{r}}_t^{(j)}\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)$, $\breve{\mathbf{z}}_t$, and $\breve{\mathbf{r}}_t$ are defined in (31), (33), and (34), respectively. The MMSE $\epsilon\big(\mathbf{x}_t^{(j)}; \breve{\mathbf{z}}_{0:t}, \breve{\mathbf{r}}_{0:t}, \mathring{\mathbf{r}}_{0:t}^{(j)}\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\big)$ satisfies

$$\begin{aligned}\epsilon\big(\mathbf{x}_t^{(j)}; \breve{\mathbf{z}}_{0:t}, \breve{\mathbf{r}}_{0:t}, \mathring{\mathbf{r}}_{0:t}^{(j)}\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\big) &\leqslant \epsilon\big(\mathbf{x}_t^{(j)}; \mathbf{z}_{0:t}^{(j)}, \mathring{\mathbf{r}}_{0:t}^{(j)}\big(\mathcal{N}^{(j)}\big)\big)\\ &= \varepsilon_t^{(j)}.\end{aligned} \tag{59}$$

To see the inequality in (59), note that $j \in \mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)$ always holds, and thus $\breve{\mathbf{z}}_{0:t}$ contains all the entries of $\mathring{\mathbf{z}}_{0:t}^{(j)}$.

In addition, $\mathring{\mathbf{r}}_{0:t}^{(j)}\big(\mathcal{N}^{(j)}\big) = \mathring{\mathbf{r}}_{0:t}^{(j)}\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big) + \mathring{\mathbf{r}}_{0:t}^{(j)}\big(\mathcal{N}^{(j)} \setminus \mathcal{N}_{\mathrm{s}}^{(j)}\big)$, and $\check{\mathbf{r}}_{0:t}$ contains all the entries of $\mathring{\mathbf{r}}_{0:t}^{(j)}\big(\mathcal{N}^{(j)} \setminus \mathcal{N}_{\mathrm{s}}^{(j)}\big)$. Therefore, $\big[\check{\mathbf{r}}_{0:t}^{\mathrm{T}} \quad \mathring{\mathbf{r}}_{0:t}^{(j)}\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)^{\mathrm{T}}\big]^{\mathrm{T}}$ contains all the entries of $\mathring{\mathbf{r}}_{0:t}^{(j)}\big(\mathcal{N}^{(j)}\big)$. Using the property of $\epsilon(\cdot;\cdot)$, the inequality in (59) can be seen to hold. The equality in (59) is due to the definition of $\varepsilon_t^{(j)}$. According to (59), if (8) holds, then

$$\sup_{t\geqslant 0}\, \epsilon\Big(\mathbf{x}_t^{(j)};\, \check{\mathbf{z}}_{0:t}, \check{\mathbf{r}}_{0:t}, \mathring{\mathbf{r}}_{0:t}^{(j)}\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\Big) < \infty\,. \tag{60}$$

The theorem is proved by showing that if (60) holds, then both subconditions need to be satisfied. We begin with the proof for Subcondition (ii). To prove this subcondition, we show that if (60) holds and $\mathcal{X}^{(j)} \nsubseteq \mathcal{S}\big(\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\big)$, then (28) must be satisfied. Specifically, this inequality is shown via a similarity transformation of the state. To this end, let $\boldsymbol{T}$ be an orthonormal matrix and partition it as

$$\boldsymbol{T} = \begin{bmatrix}\boldsymbol{T}_0 & \boldsymbol{T}_\theta & \boldsymbol{T}_\varphi\end{bmatrix} \tag{61}$$

where $\boldsymbol{T}_\varphi$ satisfies

$$\mathcal{C}(\boldsymbol{T}_\varphi) = \mathcal{S}\big(\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\big)$$

and $\mathcal{C}(\boldsymbol{T}_\theta)$ is the orthogonal complement of $\mathcal{S}\big(\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\big)$ with respect to $\mathcal{X}^{(j)} + \mathcal{S}\big(\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\big)$. Recall from (25) that $\mathcal{C}(\boldsymbol{T}_\varphi) = \mathcal{S}\big(\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\big)$ is $\boldsymbol{A}^{\mathrm{T}}$-invariant. Combining this with $\boldsymbol{T}_\theta^{\mathrm{T}}\boldsymbol{T}_\varphi = \boldsymbol{T}_0^{\mathrm{T}}\boldsymbol{T}_\varphi = \mathbf{0}$ gives

$$\boldsymbol{T}_\varphi^{\mathrm{T}}\boldsymbol{A}\boldsymbol{T}_\theta = \big(\boldsymbol{T}_\theta^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{T}_\varphi\big)^{\mathrm{T}} = \mathbf{0} \tag{62a}$$

$$\boldsymbol{T}_\varphi^{\mathrm{T}}\boldsymbol{A}\boldsymbol{T}_0 = \big(\boldsymbol{T}_0^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{T}_\varphi\big)^{\mathrm{T}} = \mathbf{0}\,. \tag{62b}$$

Moreover, $\mathcal{C}\big([\boldsymbol{T}_\theta \quad \boldsymbol{T}_\varphi]\big) = \mathcal{X}^{(j)} + \mathcal{S}\big(\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\big)$ is also $\boldsymbol{A}^{\mathrm{T}}$-invariant, as both $\mathcal{X}^{(j)}$ and $\mathcal{S}\big(\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\big)$ are $\boldsymbol{A}^{\mathrm{T}}$-invariant. Combining this with $\boldsymbol{T}_0^{\mathrm{T}}[\boldsymbol{T}_\theta \quad \boldsymbol{T}_\varphi] = \mathbf{0}$ gives

$$\boldsymbol{T}_\theta^{\mathrm{T}}\boldsymbol{A}\boldsymbol{T}_0 = \big(\boldsymbol{T}_0^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{T}_\theta\big)^{\mathrm{T}} = \mathbf{0}\,. \tag{63}$$

Left multiplying $\boldsymbol{T}^{\mathrm{T}}$ to (14) and applying (62) as well as (63) gives

$$\begin{bmatrix}\boldsymbol{\uptheta}_t\\ \boldsymbol{\upvarphi}_t\end{bmatrix} = \begin{bmatrix}\boldsymbol{A}_\theta & \boldsymbol{A}_{\theta\varphi}\\ \mathbf{0} & \boldsymbol{A}_\varphi\end{bmatrix}\begin{bmatrix}\boldsymbol{\uptheta}_{t-1}\\ \boldsymbol{\upvarphi}_{t-1}\end{bmatrix} + \begin{bmatrix}\boldsymbol{T}_\theta^{\mathrm{T}}\\ \boldsymbol{T}_\varphi^{\mathrm{T}}\end{bmatrix}\boldsymbol{\upzeta}_t \tag{64}$$

where

$$\boldsymbol{\uptheta}_t := \boldsymbol{T}_\theta^{\mathrm{T}}\mathbf{x}_t, \qquad \boldsymbol{\upvarphi}_t := \boldsymbol{T}_\varphi^{\mathrm{T}}\mathbf{x}_t \tag{65a}$$

$$\boldsymbol{A}_\theta := \boldsymbol{T}_\theta^{\mathrm{T}}\boldsymbol{A}\boldsymbol{T}_\theta, \quad \boldsymbol{A}_{\theta\varphi} := \boldsymbol{T}_\theta^{\mathrm{T}}\boldsymbol{A}\boldsymbol{T}_\varphi \tag{65b}$$

$$\boldsymbol{A}_\varphi := \boldsymbol{T}_\varphi^{\mathrm{T}}\boldsymbol{A}\boldsymbol{T}_\varphi. \tag{65c}$$

We show

$$\sup_{t\geqslant 0}\, \epsilon\Big(\boldsymbol{\uptheta}_t;\, \check{\mathbf{z}}_{0:t}, \check{\mathbf{r}}_{0:t}, \mathring{\mathbf{r}}_{0:t}^{(j)}\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\Big) < \infty \tag{66}$$

by considering the MMSE estimator $\hat{\boldsymbol{\uptheta}}_t$ of $\boldsymbol{\uptheta}_t$ using $\check{\mathbf{z}}_{0:t}$, $\check{\mathbf{r}}_{0:t}$, and $\mathring{\mathbf{r}}_{0:t}^{(j)}\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)$, i.e.,

$$\hat{\boldsymbol{\uptheta}}_t := \mathbb{E}\Big\{\boldsymbol{\uptheta}_t \,\Big|\, \check{\mathbf{z}}_{0:t}, \check{\mathbf{r}}_{0:t}, \mathring{\mathbf{r}}_{0:t}^{(j)}\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\Big\}.$$

Specifically, recall that $\mathcal{C}(\boldsymbol{T}_\theta) \subseteq \mathcal{X}^{(j)} + \mathcal{S}\big(\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\big)$ and $\mathcal{C}(\boldsymbol{e}_{j,v} \otimes \boldsymbol{I}_d) = \mathcal{X}^{(j)}$. Define $\mathring{\boldsymbol{\Gamma}}\big(\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\big)$ as the vertical concatenation of $\mathring{\boldsymbol{\Gamma}}^{(k)}$ for all $k \in \mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)$, i.e.,

$$\mathring{\boldsymbol{\Gamma}}\big(\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\big) := \Big[\big(\mathring{\boldsymbol{\Gamma}}^{(j)}\big)^{\mathrm{T}} \quad \big(\mathring{\boldsymbol{\Gamma}}^{(k_1)}\big)^{\mathrm{T}} \quad \big(\mathring{\boldsymbol{\Gamma}}^{(k_2)}\big)^{\mathrm{T}} \\ \cdots \quad \big(\mathring{\boldsymbol{\Gamma}}^{(k_n)}\big)^{\mathrm{T}}\Big]^{\mathrm{T}} \tag{67}$$

where the elements of $\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)$ are given by (30). Combining definitions (67) and (23) gives $\mathcal{S}\big(\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\big) = \mathcal{C}\big(\boldsymbol{O}\big(\mathring{\boldsymbol{\Gamma}}\big(\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\big), \boldsymbol{A}\big)^{\mathrm{T}}\big)$. Consequently, there exist two real matrices $\boldsymbol{G}_1$ and $\boldsymbol{G}_2$ such that

$$\boldsymbol{T}_\theta = (\boldsymbol{e}_{j,v} \otimes \boldsymbol{I}_d)\,\boldsymbol{G}_1 + \boldsymbol{O}\Big(\mathring{\boldsymbol{\Gamma}}\big(\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\big), \boldsymbol{A}\Big)^{\mathrm{T}}\boldsymbol{G}_2\,. \tag{68}$$

Taking transpose of (68), right multiplying it by $\mathbf{x}_t$, and using (15), we obtain

$$\boldsymbol{\uptheta}_t = \boldsymbol{G}_1^{\mathrm{T}}\mathbf{x}_t^{(j)} + \boldsymbol{G}_2^{\mathrm{T}}\,\boldsymbol{O}\Big(\mathring{\boldsymbol{\Gamma}}\big(\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\big), \boldsymbol{A}\Big)\mathbf{x}_t\,. \tag{69}$$

Let $\hat{\mathbf{a}}_t$ and $\hat{\mathbf{b}}_t$ represent MMSE estimators of $\mathbf{x}_t^{(j)}$ and $\boldsymbol{O}\big(\mathring{\boldsymbol{\Gamma}}\big(\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\big), \boldsymbol{A}\big)\mathbf{x}_t$, respectively, using $\check{\mathbf{z}}_{0:t}$, $\check{\mathbf{r}}_{0:t}$, and $\mathring{\mathbf{r}}_{0:t}^{(j)}\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)$. By definition,

$$\mathbb{E}\Big\{\big\|\hat{\mathbf{a}}_t - \mathbf{x}_t^{(j)}\big\|^2\Big\} = \epsilon\Big(\mathbf{x}_t^{(j)};\, \check{\mathbf{z}}_{0:t}, \check{\mathbf{r}}_{0:t}, \mathring{\mathbf{r}}_{0:t}^{(j)}\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\Big)\,. \tag{70}$$

Moreover, according to the linearity of conditional expectation, $\hat{\boldsymbol{\uptheta}}_t = \boldsymbol{G}_1^{\mathrm{T}}\hat{\mathbf{a}}_t + \boldsymbol{G}_2^{\mathrm{T}}\hat{\mathbf{b}}_t$. Combining this with (69) and applying Cauchy–Schwarz inequality and the definition of a matrix's spectral norm, we obtain

$$\begin{aligned}&\epsilon\Big(\boldsymbol{\uptheta}_t;\, \check{\mathbf{z}}_{0:t}, \check{\mathbf{r}}_{0:t}, \mathring{\mathbf{r}}_{0:t}^{(j)}\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\Big)\\ &\quad = \mathbb{E}\Big\{\big\|\boldsymbol{\uptheta}_t - \hat{\boldsymbol{\uptheta}}_t\big\|^2\Big\}\\ &\quad \leqslant 2\big\|\boldsymbol{G}_1^{\mathrm{T}}\big\|^2\,\mathbb{E}\Big\{\big\|\hat{\mathbf{a}}_t - \mathbf{x}_t^{(j)}\big\|^2\Big\}\\ &\qquad + 2\big\|\boldsymbol{G}_2^{\mathrm{T}}\big\|^2\,\mathbb{E}\Big\{\Big\|\hat{\mathbf{b}}_t - \boldsymbol{O}\Big(\mathring{\boldsymbol{\Gamma}}\big(\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\big), \boldsymbol{A}\Big)\mathbf{x}_t\Big\|^2\Big\}.\end{aligned} \tag{71}$$

Combining (70) with (60) gives

$$\sup_{t\geqslant 0}\, \mathbb{E}\Big\{\big\|\hat{\mathbf{a}}_t - \mathbf{x}_t^{(j)}\big\|^2\Big\} < \infty\,. \tag{72}$$

Moreover, applying Lemma 1 with $\mathcal{V}_0 = \mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)$, we conclude that an estimator $\hat{\boldsymbol{\upbeta}}_t$ of $\boldsymbol{O}\big(\mathring{\boldsymbol{\Gamma}}\big(\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\big), \boldsymbol{A}\big)\mathbf{x}_t$ can be constructed at each time t using $\check{\mathbf{z}}_{0:t}$ so that the MSE of $\hat{\boldsymbol{\upbeta}}_t$ is bounded over time. Combining this with the fact that the MSE of $\hat{\mathbf{b}}_t$ is smaller than that of $\hat{\boldsymbol{\upbeta}}_t$ since $\hat{\mathbf{b}}_t$ is the MMSE estimator using $\check{\mathbf{z}}_{0:t}$, $\check{\mathbf{r}}_{0:t}$, and $\mathring{\mathbf{r}}_{0:t}^{(j)}\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)$, we obtain

$$\sup_{t\geqslant 0}\, \mathbb{E}\Big\{\Big\|\hat{\mathbf{b}}_t - \boldsymbol{O}\Big(\mathring{\boldsymbol{\Gamma}}\big(\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\big), \boldsymbol{A}\Big)\mathbf{x}_t\Big\|^2\Big\} < \infty\,. \tag{73}$$

Combining (71)–(73) gives (66).

Next, an inequality between the MSE for estimating $\boldsymbol{\uptheta}_t$ and the conditional entropy power of $\boldsymbol{\uptheta}_t$ is established. To this end, define $\boldsymbol{\uppsi}_t$ as

$$\boldsymbol{\uppsi}_t := \Big[\boldsymbol{\upvarphi}_0^{\mathrm{T}} \quad \big(\boldsymbol{T}_\varphi^{\mathrm{T}}\boldsymbol{\upzeta}_1\big)^{\mathrm{T}} \quad \big(\boldsymbol{T}_\varphi^{\mathrm{T}}\boldsymbol{\upzeta}_2\big)^{\mathrm{T}} \cdots \big(\boldsymbol{T}_\varphi^{\mathrm{T}}\boldsymbol{\upzeta}_t\big)^{\mathrm{T}} \\ \check{\mathbf{z}}_{0:t}^{\mathrm{T}} \quad \check{\mathbf{r}}_{0:t}^{\mathrm{T}} \quad \mathring{\mathbf{r}}_{0:t}^{(j)}\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)^{\mathrm{T}}\Big]^{\mathrm{T}}. \tag{74}$$

Recall from (36) that $N\big(\boldsymbol{\theta}_t \,\big|\, \boldsymbol{\psi}_t\big)$ represents the conditional entropy power of $\boldsymbol{\theta}_t$ given $\boldsymbol{\psi}_t$. Applying maximum differential entropy lemma gives

$$\begin{aligned}\frac{1}{2\pi e} N\big(\boldsymbol{\theta}_t \,\big|\, \boldsymbol{\psi}_t\big) &\leqslant \Big|\det\Big(\mathbb{V}\big\{\mathbb{E}\{\boldsymbol{\theta}_t t \,|\, \boldsymbol{\psi}_t\} - \boldsymbol{\theta}_t\big\}\Big)\Big|^{1/d_\theta} \\ &\leqslant \operatorname{tr}\Big\{\mathbb{V}\big\{\mathbb{E}\{\boldsymbol{\theta}_t t \,|\, \boldsymbol{\psi}_t\} - \boldsymbol{\theta}_t\big\}\Big\} \\ &= \epsilon\big(\boldsymbol{\theta}_t ; \boldsymbol{\psi}_t\big) \leqslant \epsilon\big(\boldsymbol{\theta}_t ; \breve{\mathbf{z}}_{0:t}, \breve{\mathbf{r}}_{0:t}, \mathring{\mathbf{r}}^{(j)}_{0:t}(\mathcal{N}^{(j)}_{\mathrm{s}})\big)\end{aligned} \tag{75}$$

where d_θ represents the number of entries in $\boldsymbol{\theta}_t$. The second inequality in (75) is obtained by applying the relationship between the geometric and arithmetic means of all the eigenvalues of $\mathbb{V}\big\{\mathbb{E}\{\boldsymbol{\theta}_t t \,|\, \boldsymbol{\psi}_t\} - \boldsymbol{\theta}_t\big\}$. The equality in (75) is obtained using the definition (58). The last inequality in (75) is obtained by using the property of $\epsilon(\cdot;\cdot)$ and by noticing that $\boldsymbol{\psi}_t$ defined in (74) contains all the entries of $\breve{\mathbf{z}}_{0:t}$, $\breve{\mathbf{r}}_{0:t}$, and $\mathring{\mathbf{r}}^{(j)}_{0:t}(\mathcal{N}^{(j)}_{\mathrm{s}})$. Combining (75) with (66) shows that $\sup_{t\geqslant 0} N\big(\boldsymbol{\theta}_t \,\big|\, \boldsymbol{\psi}_t\big) < \infty$.

Next, it is shown that $\sup_{t\geqslant 0} N\big(\boldsymbol{\theta}_t \,\big|\, \boldsymbol{\psi}_t\big) < \infty$ requires (28) to hold. To this end, define a random vector $\check{\boldsymbol{\theta}}_t$ as

$$\check{\boldsymbol{\theta}}_t = \boldsymbol{A}_\theta \check{\boldsymbol{\theta}}_{t-1} + \boldsymbol{T}_\theta^{\mathrm{T}} \boldsymbol{\zeta}_t, \quad \check{\boldsymbol{\theta}}_0 = \boldsymbol{\theta}_0\,. \tag{76}$$

Comparing this with (64) gives

$$\boldsymbol{\theta}_t = \check{\boldsymbol{\theta}}_t + \sum_{\tau=0}^{t-1} \boldsymbol{A}_\theta^{t-1-\tau} \boldsymbol{A}_{\theta\varphi} \boldsymbol{\varphi}_\tau\,. \tag{77}$$

According to (64), random vector $\boldsymbol{\varphi}_t$ is a function of $\boldsymbol{\varphi}_0^{\mathrm{T}}, \boldsymbol{T}_\varphi^{\mathrm{T}}\boldsymbol{\zeta}_1, \boldsymbol{T}_\varphi^{\mathrm{T}}\boldsymbol{\zeta}_2, \ldots, \boldsymbol{T}_\varphi^{\mathrm{T}}\boldsymbol{\zeta}_t$. Therefore, the term $\sum_{\tau=0}^{t-1} \boldsymbol{A}_\theta^{t-1-\tau} \boldsymbol{A}_{\theta\varphi} \boldsymbol{\varphi}_\tau$ in (77) is a function of $\boldsymbol{\psi}_t$. As a result,

$$\begin{aligned}N\big(\boldsymbol{\theta}_t \,\big|\, \boldsymbol{\psi}_t\big) &= N\big(\check{\boldsymbol{\theta}}_t \,\big|\, \boldsymbol{\psi}_t\big) \\ &= N\big(\check{\boldsymbol{\theta}}_t \,\big|\, \boldsymbol{\psi}_{t-1}, \boldsymbol{T}_\varphi^{\mathrm{T}}\boldsymbol{\zeta}_t, \breve{\mathbf{z}}_t, \breve{\mathbf{r}}_t, \mathring{\mathbf{r}}^{(j)}_t(\mathcal{N}^{(j)}_{\mathrm{s}})\big)\,.\end{aligned} \tag{78}$$

Let $\breve{\mathbf{m}}_t$ be a vector consisting of all the transmitted messages along edges in $\mathcal{E}_j\big(\mathcal{N}^{(j)}_{\mathrm{s}}\big)$ (see (32)) at time t, i.e., $\breve{\mathbf{m}}_t := \big[\mathbf{m}_t^{(il)}\big]_{(i,l)\in\mathcal{E}_j(\mathcal{N}^{(j)}_{\mathrm{s}})}$. Using the second equality in (78) and the fact that conditioning reduces differential entropy, we obtain

$$\begin{aligned}N\big(\check{\boldsymbol{\theta}}_t \,\big|\, \boldsymbol{\psi}_t\big) &\geqslant N\big(\check{\boldsymbol{\theta}}_t \,\big|\, \boldsymbol{\psi}_{t-1}, \boldsymbol{T}_\varphi^{\mathrm{T}}\boldsymbol{\zeta}_t, \breve{\mathbf{z}}_t, \breve{\mathbf{r}}_t, \mathring{\mathbf{r}}^{(j)}_t(\mathcal{N}^{(j)}_{\mathrm{s}}), \breve{\mathbf{m}}_t\big) \\ &= N\big(\check{\boldsymbol{\theta}}_t \,\big|\, \boldsymbol{\psi}_{t-1}, \boldsymbol{T}_\varphi^{\mathrm{T}}\boldsymbol{\zeta}_t, \breve{\mathbf{z}}_t, \mathring{\mathbf{r}}^{(j)}_t(\mathcal{N}^{(j)}_{\mathrm{s}}), \breve{\mathbf{m}}_t\big) \\ &= N\big(\check{\boldsymbol{\theta}}_t \,\big|\, \boldsymbol{\psi}_{t-1}, \boldsymbol{T}_\varphi^{\mathrm{T}}\boldsymbol{\zeta}_t, \breve{\mathbf{z}}_t, \mathring{\mathbf{r}}^{(j)}_t(\mathcal{N}^{(j)}_{\mathrm{s}})\big)\,.\end{aligned} \tag{79}$$

The first equality in (79) is due to the conditional independence

$$\breve{\mathbf{r}}_t \perp\!\!\!\perp \check{\boldsymbol{\theta}}_t, \boldsymbol{\psi}_{t-1}, \boldsymbol{T}_\varphi^{\mathrm{T}}\boldsymbol{\zeta}_t, \breve{\mathbf{z}}_t, \mathring{\mathbf{r}}^{(j)}_t(\mathcal{N}^{(j)}_{\mathrm{s}}) \,\big|\, \breve{\mathbf{m}}_t$$

indicated by Assumption A4 of Section II. The second equality in (79) is because $\breve{\mathbf{m}}_t$ is a function of $\breve{\mathbf{z}}_{0:t}$, $\breve{\mathbf{r}}_{0:t-1}$, and $\mathring{\mathbf{r}}^{(j)}_{0:t-1}(\mathcal{N}^{(j)}_{\mathrm{s}})$ based on the model of transmitted messages presented in Section II, whereas $\breve{\mathbf{z}}_{0:t}$, $\breve{\mathbf{r}}_{0:t-1}$, and $\mathring{\mathbf{r}}^{(j)}_{0:t-1}(\mathcal{N}^{(j)}_{\mathrm{s}})$ are functions of $\boldsymbol{\psi}_{t-1}$ and $\breve{\mathbf{z}}_t$ as shown by (74). Applying properties of conditional differential entropy and mutual information gives

$$\begin{aligned}&h\big(\check{\boldsymbol{\theta}}_t \,\big|\, \boldsymbol{\psi}_{t-1}, \boldsymbol{T}_\varphi^{\mathrm{T}}\boldsymbol{\zeta}_t, \breve{\mathbf{z}}_t, \mathring{\mathbf{r}}^{(j)}_t(\mathcal{N}^{(j)}_{\mathrm{s}})\big) \\ &= h\big(\check{\boldsymbol{\theta}}_t \,\big|\, \boldsymbol{\psi}_{t-1}, \boldsymbol{T}_\varphi^{\mathrm{T}}\boldsymbol{\zeta}_t, \breve{\mathbf{z}}_t\big) - I\big(\check{\boldsymbol{\theta}}_t ; \mathring{\mathbf{r}}^{(j)}_t(\mathcal{N}^{(j)}_{\mathrm{s}}) \,\big|\, \boldsymbol{\psi}_{t-1}, \boldsymbol{T}_\varphi^{\mathrm{T}}\boldsymbol{\zeta}_t, \breve{\mathbf{z}}_t\big) \\ &\geqslant h\big(\check{\boldsymbol{\theta}}_t \,\big|\, \boldsymbol{\psi}_{t-1}, \boldsymbol{T}_\varphi^{\mathrm{T}}\boldsymbol{\zeta}_t, \breve{\mathbf{z}}_t\big) \\ &\quad - I\big(\check{\boldsymbol{\theta}}_t, \boldsymbol{\psi}_{t-1}, \boldsymbol{T}_\varphi^{\mathrm{T}}\boldsymbol{\zeta}_t, \breve{\mathbf{z}}_t \,;\, \mathring{\mathbf{r}}^{(j)}_t(\mathcal{N}^{(j)}_{\mathrm{s}})\big)\,.\end{aligned} \tag{80}$$

According to Assumption A4 in Section II,

$$\mathring{\mathbf{r}}^{(j)}_t(\mathcal{N}^{(j)}_{\mathrm{s}}) \perp\!\!\!\perp \check{\boldsymbol{\theta}}_t, \boldsymbol{\psi}_{t-1}, \boldsymbol{T}_\varphi^{\mathrm{T}}\boldsymbol{\zeta}_t, \breve{\mathbf{z}}_t \,\big|\, \mathring{\mathbf{m}}^{(j)}_t(\mathcal{N}^{(j)}_{\mathrm{s}})$$

where $\mathring{\mathbf{m}}^{(j)}_t(\mathcal{N}^{(j)}_{\mathrm{s}})$ represents the transmitted messages from all the nodes in $\mathcal{N}^{(j)}_{\mathrm{s}}$ to node j, i.e., $\mathring{\mathbf{m}}^{(j)}_t(\mathcal{N}^{(j)}_{\mathrm{s}})$ is obtained by replacing $\mathbf{r}$ in (31) with $\mathbf{m}$. Applying data processing inequality for mutual information [79, Chapter 2],

$$\begin{aligned}&I\big(\check{\boldsymbol{\theta}}_t, \boldsymbol{\psi}_{t-1}, \boldsymbol{T}_\varphi^{\mathrm{T}}\boldsymbol{\zeta}_t, \breve{\mathbf{z}}_t \,;\, \mathring{\mathbf{r}}^{(j)}_t(\mathcal{N}^{(j)}_{\mathrm{s}})\big) \\ &\quad\leqslant I\big(\mathring{\mathbf{m}}^{(j)}_t(\mathcal{N}^{(j)}_{\mathrm{s}}) ; \mathring{\mathbf{r}}^{(j)}_t(\mathcal{N}^{(j)}_{\mathrm{s}})\big) \\ &\quad= \sum_{i\in\mathcal{N}^{(j)}_{\mathrm{s}}} I\big(\mathbf{m}^{(ij)}_t ; \mathbf{r}^{(ij)}_t\big) \leqslant \sum_{i\in\mathcal{N}^{(j)}_{\mathrm{s}}} C^{(ij)}_{\mathrm{n}}\end{aligned} \tag{81}$$

where $C^{(ij)}_{\mathrm{n}}$ represents the Shannon capacity of the channel from node i to node j in the unit of nats. Combining (79)–(81) and using the definition (36) gives

$$N\big(\check{\boldsymbol{\theta}}_t \,\big|\, \boldsymbol{\psi}_t\big) \geqslant N\big(\check{\boldsymbol{\theta}}_t \,\big|\, \boldsymbol{\psi}_{t-1}, \boldsymbol{T}_\varphi^{\mathrm{T}}\boldsymbol{\zeta}_t, \breve{\mathbf{z}}_t\big) \exp\Big\{-\frac{2}{d_\theta} \sum_{i\in\mathcal{N}^{(j)}_{\mathrm{s}}} C^{(ij)}_{\mathrm{n}}\Big\} \tag{82}$$

where we recall that d_θ represents the number of entries in $\check{\boldsymbol{\theta}}_t$. Define $\breve{\mathbf{n}}_t$ as the vertical concatenation of $\mathring{\mathbf{n}}^{(j)}_t, \mathring{\mathbf{n}}^{(k_1)}_t, \mathring{\mathbf{n}}^{(k_2)}_t, \ldots, \mathring{\mathbf{n}}^{(k_n)}_t$, where $\mathring{\mathbf{n}}^{(j)}_t$ is defined in (21) for any $j \in \mathcal{V}$. Then $\breve{\mathbf{z}}_t$ can be written as

$$\breve{\mathbf{z}}_t = \mathring{\boldsymbol{\Gamma}}\big(\mathcal{V}_j(\mathcal{N}^{(j)}_{\mathrm{s}})\big)\mathbf{x}_t + \breve{\mathbf{n}}_t\,. \tag{83}$$

Since

$$\begin{aligned}\mathcal{C}\Big(\mathring{\boldsymbol{\Gamma}}\big(\mathcal{V}_j(\mathcal{N}^{(j)}_{\mathrm{s}})\big)^{\mathrm{T}}\Big) &\subseteq \mathcal{C}\Big(\boldsymbol{O}\Big(\mathring{\boldsymbol{\Gamma}}\big(\mathcal{V}_j(\mathcal{N}^{(j)}_{\mathrm{s}})\big), \boldsymbol{A}\Big)^{\mathrm{T}}\Big) \\ &= \mathcal{S}\big(\mathcal{V}_j(\mathcal{N}^{(j)}_{\mathrm{s}})\big) = \mathcal{C}(\boldsymbol{T}_\varphi)\end{aligned}$$

and $\boldsymbol{T}$ partitioned as (61) is orthonormal, it holds that $\mathring{\boldsymbol{\Gamma}}\big(\mathcal{V}_j(\mathcal{N}^{(j)}_{\mathrm{s}})\big)\boldsymbol{T}_0 = \mathring{\boldsymbol{\Gamma}}\big(\mathcal{V}_j(\mathcal{N}^{(j)}_{\mathrm{s}})\big)\boldsymbol{T}_\theta = \mathbf{0}$. Substituting this into (83) and using (65a) gives

$$\breve{\mathbf{z}}_t = \mathring{\boldsymbol{\Gamma}}\big(\mathcal{V}_j(\mathcal{N}^{(j)}_{\mathrm{s}})\big)\boldsymbol{T}_\varphi \boldsymbol{\varphi}_t + \breve{\mathbf{n}}_t\,. \tag{84}$$

Applying (64) recursively, and substituting the result into (84),

$$\begin{aligned}\breve{\mathbf{z}}_t = \mathring{\boldsymbol{\Gamma}}\big(\mathcal{V}_j(\mathcal{N}^{(j)}_{\mathrm{s}})\big)\boldsymbol{T}_\varphi \Big[\boldsymbol{A}_\varphi^t \boldsymbol{\varphi}_0 + \Big(\sum_{\tau=1}^{t-1} \boldsymbol{A}_\varphi^{t-\tau} \boldsymbol{T}_\varphi^{\mathrm{T}} \boldsymbol{\zeta}_\tau\Big) + \boldsymbol{T}_\varphi^{\mathrm{T}} \boldsymbol{\zeta}_t\Big] \\ + \breve{\mathbf{n}}_t.\end{aligned}$$

Combining this with (74) shows that $\breve{\mathbf{z}}_t$ is a function of $\boldsymbol{\psi}_{t-1}$, $\boldsymbol{T}_\varphi^{\mathrm{T}}\boldsymbol{\zeta}_t$, and $\breve{\mathbf{n}}_t$. According to Assumption A2, $\breve{\mathbf{n}}_t \perp\!\!\!\perp \boldsymbol{\psi}_{t-1}, \boldsymbol{T}_\varphi^{\mathrm{T}}\boldsymbol{\zeta}_t, \check{\boldsymbol{\theta}}_t$. Therefore, $\breve{\mathbf{z}}_t \perp\!\!\!\perp \check{\boldsymbol{\theta}}_t \,\big|\, \boldsymbol{\psi}_{t-1}, \boldsymbol{T}_\varphi^{\mathrm{T}}\boldsymbol{\zeta}_t$. As a result,

$$N\big(\check{\boldsymbol{\theta}}_t \,\big|\, \boldsymbol{\psi}_{t-1}, \boldsymbol{T}_\varphi^{\mathrm{T}}\boldsymbol{\zeta}_t, \breve{\mathbf{z}}_t\big) = N\big(\check{\boldsymbol{\theta}}_t \,\big|\, \boldsymbol{\psi}_{t-1}, \boldsymbol{T}_\varphi^{\mathrm{T}}\boldsymbol{\zeta}_t\big)\,. \tag{85}$$

According to Assumption A2 in Section II,

$$\boldsymbol{\zeta}_t \perp\!\!\!\perp \check{\boldsymbol{\theta}}_{t-1}, \boldsymbol{\psi}_{t-1}\,. \tag{86}$$

Therefore, $\boldsymbol{A}_\theta \check{\boldsymbol{\theta}}_{t-1} \perp\!\!\!\perp \boldsymbol{T}_\theta^{\mathrm{T}}\boldsymbol{\zeta}_t \,\big|\, \boldsymbol{\psi}_{t-1}, \boldsymbol{T}_\varphi^{\mathrm{T}}\boldsymbol{\zeta}_t$. Using (76) and applying conditional entropy power inequality (EPI) [76, Chapter 2] gives

$$\begin{aligned}&N\big(\check{\boldsymbol{\theta}}_t \,\big|\, \boldsymbol{\psi}_{t-1}, \boldsymbol{T}_\varphi^{\mathrm{T}}\boldsymbol{\zeta}_t\big) \\ &\quad\geqslant N\big(\boldsymbol{A}_\theta \check{\boldsymbol{\theta}}_{t-1} \,\big|\, \boldsymbol{\psi}_{t-1}, \boldsymbol{T}_\varphi^{\mathrm{T}}\boldsymbol{\zeta}_t\big) + N\big(\boldsymbol{T}_\theta^{\mathrm{T}}\boldsymbol{\zeta}_t \,\big|\, \boldsymbol{\psi}_{t-1}, \boldsymbol{T}_\varphi^{\mathrm{T}}\boldsymbol{\zeta}_t\big) \\ &\quad= \big|\det(\boldsymbol{A}_\theta)\big|^{2/d_\theta} N\big(\check{\boldsymbol{\theta}}_{t-1} \,\big|\, \boldsymbol{\psi}_{t-1}\big) + N\big(\boldsymbol{T}_\theta^{\mathrm{T}}\boldsymbol{\zeta}_t \,\big|\, \boldsymbol{\psi}_{t-1}, \boldsymbol{T}_\varphi^{\mathrm{T}}\boldsymbol{\zeta}_t\big) \\ &\quad= \big|\det(\boldsymbol{A}_\theta)\big|^{2/d_\theta} N\big(\check{\boldsymbol{\theta}}_{t-1} \,\big|\, \boldsymbol{\psi}_{t-1}\big) + N\big(\boldsymbol{T}_\theta^{\mathrm{T}}\boldsymbol{\zeta}_t \,\big|\, \boldsymbol{T}_\varphi^{\mathrm{T}}\boldsymbol{\zeta}_t\big)\end{aligned} \tag{87}$$

where the first equality is due to a property of differential entropy, and the second equality is obtained using (86). Based on Assumption A3 in Section II, it can be shown that there exists a number $\underline{h}(\boldsymbol{T}_\theta, \boldsymbol{T}_\varphi) > -\infty$ such that

$$h\big(\boldsymbol{T}_\theta^{\mathrm{T}}\boldsymbol{\zeta}_t \,\big|\, \boldsymbol{T}_\varphi^{\mathrm{T}}\boldsymbol{\zeta}_t\big) \geqslant \underline{h}(\boldsymbol{T}_\theta, \boldsymbol{T}_\varphi) \quad \forall t \geqslant 0\,. \tag{88}$$

Combining (78) and (82)–(88) gives

$$\begin{aligned} N\big(\boldsymbol{\theta}_t \,\big|\, \boldsymbol{\psi}_t\big) \geqslant \Big(&\frac{|\det(\boldsymbol{A}_\theta)|}{\exp\{\sum_{i\in\mathcal{N}_{\mathrm{s}}^{(j)}} C_{\mathrm{n}}^{(ij)}\}}\Big)^{2/d_\theta} N\big(\boldsymbol{\theta}_{t-1} \,\big|\, \boldsymbol{\psi}_{t-1}\big) \\ &+ \exp\Big\{\frac{2}{d_\theta}\Big(\underline{h}(\boldsymbol{T}_\theta, \boldsymbol{T}_\varphi) - \sum_{i\in\mathcal{N}_{\mathrm{s}}^{(j)}} C_{\mathrm{n}}^{(ij)}\Big)\Big\}\,. \end{aligned} \tag{89}$$

According to (89), $\sup_{t\geqslant 0} N\big(\boldsymbol{\theta}_t \,\big|\, \boldsymbol{\psi}_t\big) < \infty$ holds only if

$$\begin{aligned} \sum_{i\in\mathcal{N}_{\mathrm{s}}^{(j)}} C_{\mathrm{n}}^{(ij)} > \ln\big|\det(\boldsymbol{A}_\theta)\big| &= \ln\big|\det\big(\boldsymbol{T}_\theta^{\mathrm{T}}\boldsymbol{A}\boldsymbol{T}_\theta\big)\big| \\ &= \ln\big|\det\big(\boldsymbol{T}_\theta^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{T}_\theta\big)\big| \end{aligned}$$

which translates to (38) in the unit of bits. Recall that $\mathcal{C}(\boldsymbol{T}_\theta)$ is an orthogonal complement of $\mathcal{S}\big(\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\big)$ with respect to $\mathcal{X}^{(j)} + \mathcal{S}\big(\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\big)$. Since both $\mathcal{S}\big(\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\big)$ and $\mathcal{X}^{(j)}$ are $\boldsymbol{A}^{\mathrm{T}}$-invariant, applying Lemma 3 by setting $\boldsymbol{F} = \boldsymbol{A}^{\mathrm{T}}$ and $\mathcal{Y} = \mathcal{S}\big(\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\big)$ gives

$$\big|\det\big(\boldsymbol{T}_\theta^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{T}_\theta\big)\big| = \sum_{\lambda\in\Lambda((\boldsymbol{A}^{(j)})^{\mathrm{T}})} |\lambda|^{\check{r}(\lambda)} \tag{90}$$

where

$$\begin{aligned} \check{r}(\lambda) := \dim&\big(\mathcal{M}_\lambda\big((\boldsymbol{A}^{(j)})^{\mathrm{T}}\big)\big) \\ &- \dim\big(\mathcal{S}\big(\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\big) \cap \mathcal{X}^{(j)} \cap \mathcal{M}_\lambda(\boldsymbol{A}^{\mathrm{T}})\big)\,. \end{aligned}$$

Since $\boldsymbol{A}$ is block-diagonal, it can be shown that $\dim\big(\mathcal{M}_\lambda\big((\boldsymbol{A}^{(j)})^{\mathrm{T}}\big)\big) = \dim\big(\mathcal{X}^{(j)} \cap \mathcal{M}_\lambda(\boldsymbol{A}^{\mathrm{T}})\big)$, and thus $\check{r}(\lambda) = r_\lambda^{(j)}\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)$ according to (29). Substituting this and $\Lambda\big((\boldsymbol{A}^{(j)})^{\mathrm{T}}\big) = \Lambda\big(\boldsymbol{A}^{(j)}\big)$ into (90) gives (39). This shows that $\sup_{t\geqslant 0} N\big(\boldsymbol{\theta}_t \,\big|\, \boldsymbol{\psi}_t\big) < \infty$ only if (28) holds, thus completing the proof for Subcondition (ii).

Next, Subcondition (i) is proved by identifying a contradiction if (8) holds but (27) does not hold. In fact, this contradiction can be shown by replacing $\mathcal{N}_{\mathrm{s}}^{(j)}$ and $\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)$ in the proof for Subcondition (ii) with $\varnothing$ and $\mathcal{V}$, respectively. Specifically, let $\breve{\mathbf{z}}_t$ represent the observations obtained by all the nodes in $\mathcal{V}$ at time t, and let $\breve{\mathbf{r}}_t$ represent the messages received via all the edges $\mathcal{E}_{\mathrm{u}}$ in the network at time t. If $\mathcal{X}^{(j)} \not\subseteq \mathcal{S}(\mathcal{V})$, then there exists a matrix $\boldsymbol{T}_\theta$ whose columns are orthonormal and form an orthogonal complement of $\mathcal{S}\big(\mathcal{V}\big)$ with respect to $\mathcal{X}^{(j)} + \mathcal{S}\big(\mathcal{V}\big)$. Define $\boldsymbol{\theta}_t := \boldsymbol{T}_\theta^{\mathrm{T}}\mathbf{x}_t$. If $\sup_{t\geqslant 0} \varepsilon_t^{(j)} < \infty$, then $\sup_{t\geqslant 0} \epsilon\big(\boldsymbol{\theta}_t;\, \breve{\mathbf{z}}_{0:t}, \breve{\mathbf{r}}_{0:t}\big) < \infty$ can be shown to hold. Moreover, define $\boldsymbol{\psi}_t$ as the right-hand side of (74) with $\mathring{\mathbf{r}}_{0:t}^{(j)}\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)$ removed, where $\boldsymbol{T}_\varphi$ is a matrix with orthonormal columns such that $\mathcal{C}(\boldsymbol{T}_\varphi) = \mathcal{S}(\mathcal{V})$. It can be shown using maximum differential entropy lemma that $N\big(\boldsymbol{\theta}_t \,\big|\, \boldsymbol{\psi}_t\big)/(2\pi e) \leqslant \epsilon\big(\boldsymbol{\theta}_t;\, \breve{\mathbf{z}}_{0:t}, \breve{\mathbf{r}}_{0:t}\big)$, and thus $\sup_{t\geqslant 0} N\big(\boldsymbol{\theta}_t \,\big|\, \boldsymbol{\psi}_t\big) < \infty$. However, by deriving a recursive expression of $\big\{N\big(\boldsymbol{\theta}_t \,\big|\, \boldsymbol{\psi}_t\big)\big\}_{t\geqslant 0}$, we can show that $N\big(\boldsymbol{\theta}_t \,\big|\, \boldsymbol{\psi}_t\big)$ approaches infinity as $t \to \infty$, which is a contradiction. This shows that $\sup_{t\geqslant 0} \varepsilon_t^{(j)} < \infty$ only if $\mathcal{X}^{(j)} \subseteq \mathcal{S}(\mathcal{V})$, and thus Subcondition (i) is proved. □

Appendix C
Proof of Proposition 2

Proof: First, the expression of $\bar{\mathcal{R}}^{(j)}$, the closure of $\mathcal{R}^{(j)}$, is derived. Specifically, $\bar{\mathcal{R}}^{(j)}$ is obtained by replacing the grater-than signs in (42) with greater-than-or-equal-to signs. As a result, $\bar{\mathcal{R}}^{(j)}$ can be shown to be the set of vectors $\boldsymbol{c} = \big[c^{(i_1\, j)} \quad c^{(i_2\, j)} \quad \dots \quad c^{(i_{N^{(j)}}\, j)}\big]^{\mathrm{T}}$ that satisfy

$$\sum_{i\in\mathcal{N}_{\mathrm{s}}^{(j)}} c^{(ij)} \geqslant \begin{cases} \sum_{\lambda\in\Lambda(\boldsymbol{A}^{(j)})} r_\lambda^{(j)}\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big) \log|\lambda| & \\ \qquad\qquad\qquad \text{if } \mathcal{X}^{(j)} \not\subseteq \mathcal{S}\big(\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\big) & \\ 0 \qquad\qquad\qquad \text{otherwise} & \end{cases} \tag{91}$$

for all non-empty $\mathcal{N}_{\mathrm{s}}^{(j)} \subseteq \mathcal{N}^{(j)}$. Expression (91) is rewritten as follows. For any $\mathcal{N}_{\mathrm{s}}^{(j)} \subseteq \mathcal{N}^{(j)}$ such that $\mathcal{X}^{(j)} \subseteq \mathcal{S}\big(\mathcal{V}_j\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big)\big)$, it can be seen from (29) that $r_\lambda^{(j)}\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big) = 0$. Consequently, $\sum_{\lambda\in\Lambda(\boldsymbol{A}^{(j)})} r_\lambda^{(j)}\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big) \log|\lambda| = 0$. This shows that the two cases in (91) can be combined to obtain the following equivalent expression

$$\sum_{i\in\mathcal{N}_{\mathrm{s}}^{(j)}} c^{(ij)} \geqslant \sum_{\lambda\in\Lambda(\boldsymbol{A}^{(j)})} r_\lambda^{(j)}\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big) \log|\lambda| \quad \forall\mathcal{N}_{\mathrm{s}}^{(j)} \subseteq \mathcal{N}^{(j)}.$$

Consequently, $\bar{\mathcal{R}}^{(j)}$ consists of all the solutions to the following feasibility problem

$$\text{find} \quad \boldsymbol{c} = \big[c^{(i_1\, j)} \quad c^{(i_2\, j)} \quad \dots \quad c^{(i_{N^{(j)}}\, j)}\big]^{\mathrm{T}} \tag{92a}$$

$$\begin{aligned} \text{subject to} \sum_{i\in\mathcal{N}_{\mathrm{s}}^{(j)}} c^{(ij)} \geqslant \sum_{\lambda\in\Lambda(\boldsymbol{A}^{(j)})} r_\lambda^{(j)}\big(\mathcal{N}_{\mathrm{s}}^{(j)}\big) \log|\lambda| \\ \forall\mathcal{N}_{\mathrm{s}}^{(j)} \subseteq \mathcal{N}^{(j)}\,. \end{aligned} \tag{92b}$$

Note that (92b) specifies $2^{N^{(j)}} - 1$ constraints as $\mathcal{N}_{\mathrm{s}}^{(j)}$ can be an arbitrary non-empty subset of $\mathcal{N}^{(j)}$. Since all these constraints are linear, region $\bar{\mathcal{R}}^{(j)}$ is a polyhedron.

We next show that $\boldsymbol{\gamma}_{\mathcal{T}}^{(j)}$ is an extreme point of $\bar{\mathcal{R}}^{(j)}$ if $\mathcal{I}_{(\breve{\boldsymbol{A}}^{(ij)})^{\mathrm{T}}}\big(\mathcal{G}_\lambda^{(ij)}\big) = \mathcal{G}_\lambda^{(ij)}$. Indeed, it is shown in [68] that $\mathcal{G}_\lambda^{(ij)} \subseteq \mathcal{X}^{(i)} + \mathcal{X}^{(j)}$. Combining this with (44) gives

$$\mathcal{I}_{\boldsymbol{A}^{\mathrm{T}}}\big(\mathcal{G}_\lambda^{(ij)}\big) = \mathcal{I}_{(\breve{\boldsymbol{A}}^{(ij)})^{\mathrm{T}}}\big(\mathcal{G}_\lambda^{(ij)}\big) = \mathcal{G}_\lambda^{(ij)}\,.$$

Combining this with (40) gives

$$\gamma_{\mathcal{T}}^{(ij)} = \sum_{\lambda\in\Lambda(\boldsymbol{A}^{(j)})} \dim\big(\mathcal{G}_\lambda^{(ij)}\big) \log|\lambda|\,. \tag{93}$$

Recall that $\check{c}_n^{(j)}$ represents the nth child of node j for $n \in \{1, 2, \dots, N^{(j)} - 1\}$ in the ordered tree $\mathcal{T}$. Moreover, recall that $\check{c}_{N^{(j)}}^{(j)}$ represents the $N^{(j)}$th child or the parent of node j, depending on whether j is the root of $\mathcal{T}$ or not. Combining (93) with (41) gives

$$\begin{aligned} \sum_{i\in\mathcal{N}_n^{(j)}} \gamma_{\mathcal{T}}^{(ij)} = \sum_{\lambda\in\Lambda(\boldsymbol{A}^{(j)})} r_\lambda^{(j)}\big(\mathcal{N}_n^{(j)}\big) \log|\lambda| \\ \forall n \in \big\{1, 2, \dots, N^{(j)}\big\}\,. \end{aligned}$$

This shows that the vector $\boldsymbol{\gamma}_{\mathcal{T}}^{(j)}$ defined in (45) achieves equality in linear constraint (92b) with $\mathcal{N}_{\mathrm{s}}^{(j)} = \mathcal{N}_n^{(j)}$, i.e., the constraint is active at $\boldsymbol{\gamma}_{\mathcal{T}}^{(j)}$. Indeed, by setting

$n = 1, 2, \dots, N^{(j)}$, there are $N^{(j)}$ inequality constraints that are active at $\gamma_{\mathcal{T}}^{(j)}$, and these constraints are linearly independent. This shows that $\gamma_{\mathcal{T}}^{(j)}$ is a basic solution [78, Chapter 2] for the linear program given by (92). Since a basic feasible solution is an extreme point for linear programs [78, Theorem 2.3], we only need to show that $\gamma_{\mathcal{T}}^{(j)}$ is a feasible solution to (92), i.e., $\gamma_{\mathcal{T}}^{(j)}$ satisfies (92b) for all $\mathcal{N}_\mathrm{s}^{(j)} \subseteq \mathcal{N}^{(j)}$.

Next, it is proved that $\gamma_{\mathcal{T}}^{(j)}$ is a feasible solution to (92). For any $\mathcal{N}_\mathrm{s}^{(j)} \subseteq \mathcal{N}^{(j)}$, it can be shown that (see [81, Appendix B.3])

$$\mathcal{X}^{(j)} \subseteq \check{\mathcal{H}} + \mathcal{S}\big(\mathcal{V}_j\big(\mathcal{N}_\mathrm{s}^{(j)}\big)\big) \tag{94}$$

where

$$\check{\mathcal{H}} := \bigoplus_{\lambda \in \Lambda(\boldsymbol{A}^\mathrm{T})} \sum_{i \in \mathcal{N}_\mathrm{s}^{(j)}} \mathcal{G}_\lambda^{(ij)}. \tag{95}$$

According to the definition of $\mathcal{G}_\lambda^{(ij)}$ given in [68], $\sum_{i \in \mathcal{N}_\mathrm{s}^{(j)}} \mathcal{G}_\lambda^{(ij)} \subseteq \mathcal{M}_\lambda(\boldsymbol{A}^\mathrm{T})$ for all $\lambda \in \Lambda(\boldsymbol{A}^\mathrm{T})$. Applying Lemma 2 gives $\check{\mathcal{H}} \cap \mathcal{M}_\lambda(\boldsymbol{A}^\mathrm{T}) = \sum_{i \in \mathcal{N}_\mathrm{s}^{(j)}} \mathcal{G}_\lambda^{(ij)}$. Combining this with (95) gives

$$\check{\mathcal{H}} = \bigoplus_{\lambda \in \Lambda(\boldsymbol{A}^\mathrm{T})} \big(\check{\mathcal{H}} \cap \mathcal{M}_\lambda(\boldsymbol{A}^\mathrm{T})\big).$$

Moreover, since $\mathcal{S}\big(\mathcal{V}_j\big(\mathcal{N}_\mathrm{s}^{(j)}\big)\big)$ is $\boldsymbol{A}^\mathrm{T}$-invariant, applying Proposition 1 gives

$$\mathcal{S}\big(\mathcal{V}_j\big(\mathcal{N}_\mathrm{s}^{(j)}\big)\big) = \bigoplus_{\lambda \in \Lambda(\boldsymbol{A}^\mathrm{T})} \Big(\mathcal{S}\big(\mathcal{V}_j\big(\mathcal{N}_\mathrm{s}^{(j)}\big)\big) \cap \mathcal{M}_\lambda(\boldsymbol{A}^\mathrm{T})\Big).$$

Applying Corollary 1 with $\mathcal{V}_1 = \check{\mathcal{H}}$ and $\mathcal{V}_2 = \mathcal{S}\big(\mathcal{V}_j\big(\mathcal{N}_\mathrm{s}^{(j)}\big)\big)$ gives

$$\begin{aligned}&\Big(\check{\mathcal{H}} + \mathcal{S}\big(\mathcal{V}_j\big(\mathcal{N}_\mathrm{s}^{(j)}\big)\big)\Big) \cap \mathcal{M}_\lambda(\boldsymbol{A}^\mathrm{T})\\ &\quad = \Big(\sum_{i \in \mathcal{N}_\mathrm{s}^{(j)}} \mathcal{G}_\lambda^{(ij)}\Big) + \Big(\mathcal{S}\big(\mathcal{V}_j\big(\mathcal{N}_\mathrm{s}^{(j)}\big)\big) \cap \mathcal{M}_\lambda(\boldsymbol{A}^\mathrm{T})\Big).\end{aligned}$$

Combining this with (94) gives

$$\begin{aligned}&\mathcal{X}^{(j)} \cap \mathcal{M}_\lambda(\boldsymbol{A}^\mathrm{T})\\ &\quad \subseteq \Big[\Big(\sum_{i \in \mathcal{N}_\mathrm{s}^{(j)}} \mathcal{G}_\lambda^{(ij)}\Big) + \Big(\mathcal{S}\big(\mathcal{V}_j\big(\mathcal{N}_\mathrm{s}^{(j)}\big)\big) \cap \mathcal{M}_\lambda(\boldsymbol{A}^\mathrm{T})\Big)\Big]\\ &\qquad \cap \mathcal{X}^{(j)} \cap \mathcal{M}_\lambda(\boldsymbol{A}^\mathrm{T}).\end{aligned} \tag{96}$$

Noting that the left-hand side of (96) contains its right-hand side, we replace the subset sign in this equation by an equal sign. Applying Lemma 2 in [68] with $\mathcal{Y} = \mathcal{X}^{(j)} \cap \mathcal{M}_\lambda(\boldsymbol{A}^\mathrm{T})$, $\mathcal{U} = \mathcal{S}\big(\mathcal{V}_j\big(\mathcal{N}_\mathrm{s}^{(j)}\big)\big) \cap \mathcal{M}_\lambda(\boldsymbol{A}^\mathrm{T})$, $\mathcal{W} = \mathcal{X}^{(j)} \cap \mathcal{M}_\lambda(\boldsymbol{A}^\mathrm{T})$, and $\tilde{\mathcal{Y}} = \sum_{i \in \mathcal{N}_\mathrm{s}^{(j)}} \mathcal{G}_\lambda^{(ij)}$ gives

$$\begin{aligned}&\dim\Big(\sum_{i \in \mathcal{N}_\mathrm{s}^{(j)}} \mathcal{G}_\lambda^{(ij)}\Big)\\ &\quad \geqslant \dim\big(\mathcal{X}^{(j)} \cap \mathcal{M}_\lambda(\boldsymbol{A}^\mathrm{T})\big)\\ &\qquad - \dim\Big(\mathcal{S}\big(\mathcal{V}_j\big(\mathcal{N}_\mathrm{s}^{(j)}\big)\big) \cap \mathcal{X}^{(j)} \cap \mathcal{M}_\lambda(\boldsymbol{A}^\mathrm{T})\Big)\\ &\quad = r_\lambda^{(j)}\big(\mathcal{N}_\mathrm{s}^{(j)}\big)\end{aligned} \tag{97}$$

where the last equality is due to (29). According to (93),

$$\begin{aligned}\sum_{i \in \mathcal{N}_\mathrm{s}^{(j)}} \gamma_{\mathcal{T}}^{(ij)} &= \sum_{\lambda \in \Lambda(\boldsymbol{A}^{(j)})} \Big(\sum_{i \in \mathcal{N}_\mathrm{s}^{(j)}} \dim\big(\mathcal{G}_\lambda^{(ij)}\big)\Big) \log|\lambda|\\ &\geqslant \sum_{\lambda \in \Lambda(\boldsymbol{A}^{(j)})} \dim\Big(\sum_{i \in \mathcal{N}_\mathrm{s}^{(j)}} \mathcal{G}_\lambda^{(ij)}\Big) \log|\lambda|\\ &\geqslant \sum_{\lambda \in \Lambda(\boldsymbol{A}^{(j)})} r_\lambda^{(j)}\big(\mathcal{N}_\mathrm{s}^{(j)}\big) \log|\lambda|\end{aligned}$$

where (97) is used to obtain the last inequality. This shows that $\gamma_{\mathcal{T}}^{(j)}$ satisfies (92b) for an arbitrary $\mathcal{N}_\mathrm{s}^{(j)} \subseteq \mathcal{N}^{(j)}$. Therefore, $\gamma_{\mathcal{T}}^{(j)}$ is a basic feasible solution to (92), and is thus an extreme point of $\bar{\mathcal{R}}^{(j)}$. □

References

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] J. Kang, Z. Xiong, D. Niyato, S. Xie, and J. Zhang, "Incentive mechanism for reliable federated learning: A joint optimization approach to combining reputation and contract theory," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 10700–10714, Dec. 2019.

[7] 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.

[8] 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.

[9] 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.

[10] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, "A joint learning and communications framework for federated learning over wireless networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 269–283, Jan. 2021.

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

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

[13] A. H. Sayed, A. Tarighat, and N. Khajehnouri, "Network-based wireless location: Challenges faced in developing techniques for accurate wireless location information," *IEEE Signal Process. Mag.*, vol. 22, no. 4, pp. 24–40, Jul. 2005.

[14] 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.

[15] M. Chiani, A. Giorgetti, and E. Paolini, "Sensor radar for object tracking," *Proc. IEEE*, vol. 106, no. 6, pp. 1022–1041, Jun. 2018.

[16] 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.

[17] J. Thomas, J. Welde, G. Loianno, K. Daniilidis, and V. Kumar, "Autonomous flight for detection, localization, and tracking of moving targets with a small quadrotor," *IEEE Robot. Autom. Lett.*, vol. 2, no. 3, pp. 1762–1769, Jul. 2017.

[18] D. Wu, D. Chatzigeorgiou, K. Youcef-Toumi, and R. Ben-Mansour, "Node localization in robotic sensor networks for pipeline inspection," *IEEE Trans. Ind. Informat.*, vol. 12, no. 2, pp. 809–819, Apr. 2016.

[19] R. Karlsson and F. Gustafsson, “The future of automotive localization algorithms: Available, reliable, and scalable localization: Anywhere and anytime,” *IEEE Signal Process. Mag.*, vol. 34, no. 2, pp. 60–69, Mar. 2017.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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, Nov. 2019.

[28] U. A. Khan, S. Kar, and J. M. F. Moura, “DILAND: An algorithm for distributed sensor localization with noisy distance measurements,” *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1940–1947, Mar. 2010.

[29] 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.

[30] 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.

[31] 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.

[32] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, “Distributed optimization and statistical learning via the alternating direction method of multipliers,” *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan. 2011.

[33] 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.

[34] 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.

[35] 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.

[36] S. Marano, V. Matta, L. Tong, and P. Willett, “A likelihood-based multiple access for estimation in sensor networks,” *IEEE Trans. Signal Process.*, vol. 55, no. 11, pp. 5155–5166, Nov. 2007.

[37] 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.

[38] B. Teague, Z. Liu, F. Meyer, A. Conti, and M. Z. Win, “Network localization and navigation with scalable inference and efficient operation,” *IEEE Trans. Mobile Comput.*, vol. 21, no. 6, pp. 2072–2087, Jun. 2022.

[39] 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.

[40] 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.

[41] S. Kar, J. M. F. Moura, and K. Ramanan, “Distributed parameter estimation in sensor networks: Nonlinear observation models and imperfect communication,” *IEEE Trans. Inf. Theory*, vol. 58, no. 6, pp. 3575–3605, Jun. 2012.

[42] 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.

[43] 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.

[44] F. S. Cattivelli and A. H. Sayed, “Diffusion strategies for distributed Kalman filtering and smoothing,” *IEEE Trans. Autom. Control*, vol. 55, no. 9, pp. 2069–2084, Sep. 2010.

[45] C. G. Lopes and A. H. Sayed, “Diffusion least-mean squares over adaptive networks: Formulation and performance analysis,” *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 3122–3136, Jul. 2008.

[46] J. Du, S. D. Ma, Y. C. Wu, S. Kar, and J. M. F. Moura, “Convergence analysis of distributed inference with vector-valued Gaussian belief propagation,” *J. Mach. Learn. Res.*, vol. 18, no. 172, pp. 1–38, Apr. 2018.

[47] M. Cetin et al., “Distributed fusion in sensor networks,” *IEEE Signal Process. Mag.*, vol. 23, no. 4, pp. 42–55, Jul. 2006.

[48] A. T. Ihler, J. W. Fisher, R. L. Moses, and A. S. Willsky, “Nonparametric belief propagation for self-localization of sensor networks,” *IEEE J. Sel. Areas Commun.*, vol. 23, no. 4, pp. 809–819, Apr. 2005.

[49] X. Cao and T. Başar, “Decentralized online convex optimization with event-triggered communications,” *IEEE Trans. Signal Process.*, vol. 69, pp. 284–299, 2021.

[50] M. J. Khojasteh, P. Tallapragada, J. Cortés, and M. Franceschetti, “The value of timing information in event-triggered control,” *IEEE Trans. Autom. Control*, vol. 65, no. 3, pp. 925–940, Mar. 2020.

[51] D. V. Dimarogonas, E. Frazzoli, and K. H. Johansson, “Distributed event-triggered control for multi-agent systems,” *IEEE Trans. Autom. Control*, vol. 57, no. 5, pp. 1291–1297, May 2012.

[52] T. Wang, Y. Shen, A. Conti, and M. Z. Win, “Network navigation with scheduling: Error evolution,” *IEEE Trans. Inf. Theory*, vol. 63, no. 11, pp. 7509–7534, Nov. 2017.

[53] Y. Shen, W. Dai, and M. Z. Win, “Power optimization for network localization,” *IEEE/ACM Trans. Netw.*, vol. 22, no. 4, pp. 1337–1350, Aug. 2014.

[54] 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.

[55] O. A. Hanna, Y. H. Ezzeldin, T. Sadjadpour, C. Fragouli, and S. Diggavi, “On distributed quantization for classification,” *IEEE J. Sel. Areas Inf. Theory*, vol. 1, no. 1, pp. 237–249, May 2020.

[56] K. Ozkara, N. Singh, D. Data, and S. Diggavi, “QuPeD: Quantized personalization via distillation with applications to federated learning,” in *Proc. Adv. Neural Inf. Procss. Sys.*, Dec. 2021, pp. 3622–3634.

[57] N. Shlezinger, M. Chen, Y. C. Eldar, H. V. Poor, and S. Cui, “UVeQFed: Universal vector quantization for federated learning,” *IEEE Trans. Signal Process.*, vol. 69, pp. 500–514, 2021.

[58] V. S. Borkar, S. Mitter, and S. Tatikonda, “Markov control problems under communication constraints,” *Commun. Inf. Syst.*, vol. 1, no. 1, pp. 15–32, 2001.

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

[60] 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.

[61] R. G. Gallager and B. Nakiboğlu, “Variations on a theme by Schalkwijk and Kailath,” *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 6–17, Jan. 2010.

[62] 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.

[63] 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.

[64] 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.

[65] J. Pearson, J. P. Hespanha, and D. Liberzon, “Control with minimal cost-per-symbol encoding and quasi-optimality of event-based encoders,” *IEEE Trans. Autom. Control*, vol. 62, no. 5, pp. 2286–2301, May 2017.

[66] 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.

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

[68] Z. Liu, A. Conti, S. K. Mitter, and M. Z. Win, "Communication-efficient distributed learning over networks—Part I: Sufficient conditions for accuracy," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 4, pp. 1081–1101, Apr. 2023.

[69] 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.

[70] 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.

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

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

[73] 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.

[74] T. Kailath, *Linear Systems* (Prentice-Hall Information and System Sciences Series). Englewood Cliffs, NJ, USA: Prentice-Hall, 1980.

[75] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1979.

[76] A. E. Gamal and Y.-H. Kim, *Network Information Theory*, 1st ed. Cambridge, U.K.: Cambridge Univ. Press, 2011.

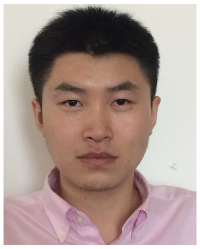
[77] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, 3rd ed. Cambridge, MA, USA: MIT Press, 2009.

[78] D. Bertsimas and J. N. Tsitsiklis, *Introduction to Linear Optimization*, 1st ed. Belmont, MA, USA: Athena Scientific, 1997.

[79] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: Wiley, 2006.

[80] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inf. Theory*, vol. IT-19, no. 4, pp. 471–480, Jul. 1973.

[81] 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.



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

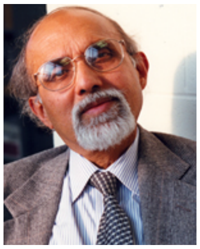
Since 2022, he has been a Post-Doctoral Associate in the wireless information and network sciences laboratory at MIT. His research interests include wireless communications, network localization, distributed inference, networked control, and quantum information science.

Dr. Liu received the first prize of the IEEE Communications Society's Student Competition in 2016 and 2019, the R&D100 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.

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, coauthored 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.