On Distributed Estimation in Hierarchical Power Constrained Wireless Sensor Networks

Mojtaba Shirazi, Azadeh Vosoughi, *Senior Member, IEEE*

Abstract—We consider distributed estimation of a random source in a hierarchical power constrained wireless sensor network. Sensors within each cluster send their measurements to a cluster head (CH). CHs optimally fuse the received signals and transmit to the fusion center (FC) over orthogonal fading channels. To enable channel estimation at the FC, CHs send pilots, prior to data transmission. We derive the mean square error (MSE) corresponding to the linear minimum mean square error (LMMSE) estimator of the source at the FC, and obtain the Bayesian Cramér-Rao bound (CRB). Our goal is to find (i) the optimal training power, (ii) the optimal power that sensors in a cluster spend to transmit their amplified measurements to their CH, and (iii) the optimal weight vector employed by each CH for its linear signal fusion, such that the MSE is minimized, subject to a network power constraint. To untangle the performance gain that optimizing each set of these variables provide, we also analyze three special cases of the original problem, where in each special case, only two sets of variables are optimized across clusters. We define three factors that allow us to quantify the effectiveness of each power allocation scheme in achieving an MSE-power tradeoff that is close to that of the Bayesian CRB. Combining the information gained from the factors and Bayesian CRB with our computational complexity analysis provides the system designer with quantitative complexity-versus-MSE improvement tradeoffs offered by different power allocation schemes.

I. INTRODUCTION

The plethora of wireless sensor network (WSN) applications, with stringent power constraints, raises challenging technical problems for system-level engineers, one of which is distributed estimation (DES) in a power constrained WSN [1]–[5]. In this work, we address DES of a random signal θ in a WSN, where sensors are deployed in a large field and make noisy measurements of θ. Due to limited communication range, however, the battery-powered sensors cannot directly communicate with the fusion center (FC). Hence, the field is divided into L geographically disjoint zones (clusters) and hierarchically into three tiers: sensors, cluster-heads (CHs) one per cluster, and the FC [6]–[9]. The implicit assumption is that the communication ranges of CHs are larger, and their energy and computational resources are higher (compared with sensors). After local signal processing, CHs transmit signals received from sensors over orthogonal fading channels to the FC, whose task is to find an estimate of θ, based on the received signals from CHs [10], [11].

There is a rich body of literature on DES and distributed detection in a power constrained WSN, where the researchers study and optimize an estimation-theoretic-based or a detection-theoretic-based performance metric, subject to power constraints. Examples in the context of distributed detection are [12]–[16]. An alternative direction is to study

the outlier contamination of the data in WSNs by outlier detection methods such as [17]–[19] caused by imperfect sensors and power deteriorations [20], [21]. We focus on power optimization and to conserve space, we elaborate only the most related ones to our current work in the following. The authors in [6]–[8] studied DES in a three-layered hierarchical power constrained WSN, assuming that the FC forms the linear minimum mean square error (LMMSE) estimate of random θ, and the objective is to minimize the MSE of this estimator. The authors in [22] considered DES in a WSN, where sensors transmit to the FC over orthogonal fading channels and the FC finds the LMMSE estimate of θ. The authors studied how partial channel state information (CSI) at the sensors affects the MSE performance and the optimal power allocation among the sensors. The authors in [23] considered DES in a hierarchical WSN, where the CHs amplify and forward their received signals over orthogonal Nakagami fading channels to the FC. Assuming the FC finds the LMMSE estimate of θ, the authors studied how partial CSI at the CHs impacts the outage probability of the MSE. None of the works in [22], [23] consider the cost of channel estimation at the FC. To enable channel estimation at the FC, each CH needs to transmit a training (pilot) symbol, prior to data symbol. In a hierarchical WSN, where there is a cap on the network transmit power, the cost of channel estimation cannot be overlooked. Note that training symbol transmission consumes the power that could have been used otherwise for data symbol transmission. Hence, *training and data transmit power should be optimized judiciously, such that the estimation accuracy of θ at the FC is maximized*.

Assuming the FC employs the LMMSE estimator of θ, we address this problem, by formulating and solving a new optimization problem that allows us to analyze the effect of channel estimation on the MSE performance and transmit power allocation. The optimization problem is novel, since considering training transmit power introduces a new dimension to the network performance analysis and power allocation optimization. In this regard, the most relevant works are [24], [25], where the authors considered channel estimation for DES in a WSN with one FC only. Our work is different from [24], [25], since in the hierarchical WSN, our problem formulation considers power distribution among different clusters for sensor-CH data transmission as well as power allocation among different CHs for CH-FC data and training transmissions. Moreover, we obtain the optimal linear fusion rules at CHs as the by-product of solving the network power

allocation problem[1].

Contribution: We derive the MSE corresponding to the LMMSE estimator of θ at the FC, denoted as D, and establish lower bounds on D, including the Bayesian Cramér-Rao bound (CRB). We then formulate a new constrained optimization problem that minimizes D, subject to network transmit power constraint P_{tot}, where the optimization variables are: i) training power for CH_l, ii) total power that sensors in cluster l spend to transmit their amplified measurements to CH_l (which we refer to as intra-cluster power), iii) power that CH_l spends to send its fused signal to the FC. We demonstrate the superior performance of our proposed power allocation scheme with respect to the following spacial case schemes: scheme (i) allots a fixed percentage of P_{tot} for training power and distributes this power equally among CHs, however, it optimally allocates intra-cluster power among clusters, and optimally allocates power among CHs for data transmission, scheme (ii) optimally allocates power among CHs for training, equally allocates intra-cluster power among clusters, and optimally allocates power among CHs for data transmission, scheme (iii) optimally allocates power among CHs for training, optimally allocates intra-cluster power among clusters, and equally allocates power among CHs for data transmission. We analytically and numerically compare the power allocation scheme obtained from solving the original problem with the special case schemes, and show their effectiveness in providing an MSE-power tradeoff that is close to that of the Bayesian CRB. Our numerical results demonstrate that power allocations among CHs for training and CH-FC data transmission are always beneficial for low-region of P_{tot}, and power allocation among clusters for sensor-CH data transmission is beneficial for low-region to moderate-region of P_{tot}.

Organization: The rest of the paper is organized as follows. Section II describes our system model and power constraints and states the problem we aim to solve (i.e., the constrained minimization of MSE D at the FC, with respect to three sets of optimization variables). Section III characterizes D and its lower bounds. We also derive the Bayesian CRB. In Section IV we solve our proposed constrained MSE minimization problem. We also briefly discuss the constrained minimization of MSE lower bounds. In Section V we solve three special cases of the original problem, where in each special case, only two (of three) sets of variables are optimized across clusters. This analysis allows us to entangle the performance gain that optimizing each set of these variables provide. Section VI compares the computational complexity of the proposed algorithms for solving the original problem as well as its three special cases. In Section VII we discuss the convergence analysis of our proposed algorithms. Section VIII presents our numerical and simulation results. Section IX concludes the work and outlines our future research directions.

Notations: Matrices are denoted by bold uppercase letters, vectors by bold lowercase letters, and scalars by normal letters. $\mathbb{E}$ denotes the mathematical expectation operator, $[.]^T$ represents the matrix-vector transpose operation, and $|\mathcal{A}|$ is the cardinality of set $\mathcal{A}$. The real and imaginary parts of a complex random variable x are represented by $x_r = \mathcal{R}e\{x\}$ and $x_i = \mathcal{I}m\{x\}$. The probability distribution function (pdf) of x, denoted as $f(x)$, is defined as the joint pdf of x_r and x_i, i.e., we have $f(x) = f(x_r, x_i)$ [27].

TABLE I: Notations and their corresponding definitions.

Notation	Vector and Matrix Definitions
$\boldsymbol{x}_l, \boldsymbol{t}_l$	$\boldsymbol{x}_l = [x_{l,1}, ..., x_{l,K_l}]^T$, $\boldsymbol{t}_l = [t_{l,1}, ..., t_{l,K_l}]^T$
$\boldsymbol{n}_l, \boldsymbol{q}_l$	$\boldsymbol{n}_l = [n_{l,1}, ..., n_{l,K_l}]^T$, $\boldsymbol{q}_l = [q_{l,1}, ..., q_{l,K_l}]^T$
$\sqrt{\boldsymbol{A}_l}$	$\sqrt{\boldsymbol{A}_l} = \text{diag}(\sqrt{\alpha_{l,1}}, ..., \sqrt{\alpha_{l,K_l}})$
$\boldsymbol{x}, \boldsymbol{t}$	$\boldsymbol{x} = [\boldsymbol{x}_1^T, ..., \boldsymbol{x}_L^T]^T$, $\boldsymbol{t} = [\boldsymbol{t}_1^T, ..., \boldsymbol{t}_L^T]^T$
$\boldsymbol{y}, \boldsymbol{z}$	$\boldsymbol{y} = [y_1, ..., y_L]^T$, $\boldsymbol{z} = [z_1, ..., z_L]^T$
$\boldsymbol{n}, \boldsymbol{q}$	$\boldsymbol{n} = [\boldsymbol{n}_1^T, ..., \boldsymbol{n}_L^T]^T$, $\boldsymbol{q} = [\boldsymbol{q}_1^T, ..., \boldsymbol{q}_L^T]^T$
$\boldsymbol{v}, \boldsymbol{H}$	$\boldsymbol{v} = [v_1, ..., v_L]^T$, $\boldsymbol{H} = \text{diag}([h_1, ..., h_L])$
$\boldsymbol{M}, \boldsymbol{W}$	$\boldsymbol{M} = \text{diag}(\sqrt{\boldsymbol{A}_1}, ..., \sqrt{\boldsymbol{A}_L})$, $\boldsymbol{W} = \text{diag}(\boldsymbol{w}_1^T, ..., \boldsymbol{w}_L^T)$
$\boldsymbol{\Sigma}_n, \boldsymbol{\Sigma}_{n_l}$	$\boldsymbol{\Sigma}_n = \text{diag}(\boldsymbol{\Sigma}_{n_1}, ..., \boldsymbol{\Sigma}_{n_L})$, $\boldsymbol{\Sigma}_{n_l}$ is arbitrary
$\boldsymbol{\Sigma}_q, \boldsymbol{\Sigma}_{q_l}$	$\boldsymbol{\Sigma}_q = \text{diag}(\boldsymbol{\Sigma}_{q_1}, ..., \boldsymbol{\Sigma}_{q_L})$, $\boldsymbol{\Sigma}_{q_l} = \text{diag}(\sigma_{q_{l,1}}^2, ..., \sigma_{q_{l,K_l}}^2)$
$\boldsymbol{\Sigma}_v$	$\boldsymbol{\Sigma}_v = \text{diag}([2\sigma_{v_1}^2, ..., 2\sigma_{v_L}^2])$
$\hat{\boldsymbol{H}}, \tilde{\boldsymbol{H}}$	$\hat{\boldsymbol{H}} = \text{diag}([\hat{h}_1, ..., \hat{h}_L])$, $\tilde{\boldsymbol{H}} = \text{diag}([\tilde{h}_1, ..., \tilde{h}_L])$
$\boldsymbol{\Gamma}, \boldsymbol{\Sigma}$	$\boldsymbol{\Gamma} = \text{diag}([\zeta_1, ..., \zeta_L])$, $\boldsymbol{\Sigma} = \text{diag}(\boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_L)$
$\boldsymbol{\mu}, \boldsymbol{\Lambda}_1$	$\boldsymbol{\mu} = [\boldsymbol{\mu}_1^H, ..., \boldsymbol{\mu}_L^H]^H$, $\boldsymbol{\Lambda}_1 = \text{diag}(\boldsymbol{\Lambda}_{1_1}, ..., \boldsymbol{\Lambda}_{1_L})$
$\boldsymbol{D}_l$	$\boldsymbol{D}_l = \text{diag}([\sqrt{d_{l,1}}, ..., \sqrt{d_{l,K_l}}])$

II. SYSTEM MODEL AND PROBLEM FORMULATION

A. System Model Description

We consider a DES problem in a hierarchical power constrained WSN (see Fig. 1), consisting of K spatially-distributed sensors deployed in L disjoint clusters, L cluster heads (CHs), and a FC. Each sensor makes a noisy measurement of an unknown random variable θ, that we wish to estimate at the FC. Cluster l includes K_l sensors and its associated CH, denoted as CH_l, and we have $\sum_{l=1}^{L} K_l = K$. We assume θ is zero-mean with variance σ_θ^2. Let $x_{l,k}$ denote the measurement of sensor k in cluster l. We have:

$$x_{l,k} = \theta + n_{l,k}, \quad l = 1, ..., L, \quad k = 1, ..., K_l, \qquad (1)$$

where $n_{l,k}$ denotes zero-mean additive measurement noise with variance $\sigma_{n_{l,k}}^2$. We assume that $n_{l,k}$'s are correlated across sensors, due to their proximity within cluster l. Sensors within a cluster amplify and forward their measurements to their respective CH over orthogonal AWGN channels[2], such that the received signal at CH_l from sensor k within cluster l is:

$$t_{l,k} = \sqrt{\alpha_{l,k}} x_{l,k} + q_{l,k}, \quad l = 1, ..., L, \quad k = 1, ..., K_l, \quad (2)$$

where $\alpha_{l,k} \geq 0$ is an amplifying factor (to be determined) used by sensor k, and $q_{l,k} \sim \mathcal{N}(0, \sigma_{q_{l,k}}^2)$ is the additive communication channel noise. We assume that $q_{l,k}$'s are uncorrelated across the sensors. For a compact representation, we define the column vectors $\boldsymbol{x}_l$ and $\boldsymbol{t}_l$ in Table I corresponding to cluster l and rewrite (1) and (2) as:

$$\boldsymbol{x}_l = \theta \boldsymbol{1}_l + \boldsymbol{n}_l, \quad \boldsymbol{t}_l = \sqrt{\boldsymbol{A}_l} \boldsymbol{x}_l + \boldsymbol{q}_l, \quad l = 1, ..., L. \quad (3)$$

---

[1]We note that there is a rich body of literature on clustering algorithms and energy efficient routing protocols [26]. Similar to [6]–[9], we assume that clusters and their CHs are given. Given this network structure, our goal is designing (sub-)optimal distributed signal processing such that the MSE distortion at the FC is minimized, under a network power constraint.

[2]The AWGN channel model is equivalent to the channel model with a static and known channel gain. Given the channel gain, CH_l can equalize it, which is equivalent to scaling the communication noise variance $\sigma_{q_{l,k}}^2$. The AWGN channel model for communication channels within a cluster is reasonable, since sensors are closely located and typically there are direct line of sight transmissions between sensors and their CH [9], [28]. On the other hand, we model the communication channels between CHs and the FC as randomly-varying fading channels that require channel estimation. The reason is that the transmission distances between CHs and the FC are large and hence communication becomes subject to multipath fading effect.

Fig. 1: Our system model consists of L clusters, each with a CH, and a FC that is tasked with estimating a random scalar $\boldsymbol{\theta}$.

where $\boldsymbol{1}_l$ is a column vector of K_l ones, column vectors $\boldsymbol{n}_l$, $\boldsymbol{q}_l$, with covarinace matrices $\boldsymbol{\Sigma}_{n_l}, \boldsymbol{\Sigma}_{q_l}$, matrix $\sqrt{\boldsymbol{A}_l}$ defined in Table I. We assume $\boldsymbol{n}_l$, $\boldsymbol{q}_l$, θ are uncorrelated, i.e., $\mathbb{E}\{\boldsymbol{n}_l \boldsymbol{q}_l^T\} = \boldsymbol{0}$, $\mathbb{E}\{\boldsymbol{n}_l \theta\} = \boldsymbol{0}$, $\mathbb{E}\{\boldsymbol{q}_l \theta\} = \boldsymbol{0}$, $\forall l$, and the noise vectors across different clusters are mutually uncorrelated, i.e., $\mathbb{E}\{\boldsymbol{n}_i \boldsymbol{n}_j^T\} = \boldsymbol{0}$ and $\mathbb{E}\{\boldsymbol{q}_i \boldsymbol{q}_j^T\} = \boldsymbol{0}$, $\forall i \neq j$.

Each CH linearly fuses the signals received from the sensors within its cluster. Let $y_l = \boldsymbol{w}_l^T \boldsymbol{t}_l$, where y_l is the scalar fused signal at CH$_l$ and $\boldsymbol{w}_l$ is the linear weight vector employed by CH$_l$ for linear fusion[3] (to be optimized). CHs transmit these fused signals to the FC over orthogonal Rayleigh fading channels, such that the received signal at the FC from CH$_l$ is:
$$z_l = h_l y_l + v_l, \quad l = 1, ..., L, \quad (4)$$
where $h_l \sim \mathcal{CN}\left(0, 2\sigma_{h_l}^2\right)$ is fading channel coefficient corresponding to the link between CH$_l$ and the FC and $v_l \sim \mathcal{CN}\left(0, 2\sigma_{v_l}^2\right)$ is the additive communication channel noise. We assume v_l is uncorrelated with $\theta, \boldsymbol{n}_l, \boldsymbol{q}_l, \forall l$.

To enable estimating h_l at the FC, CH$_l$ transmits a pilot symbol [14] with power ψ_l to the FC, prior to sending its signal y_l. Without loss of generality, we assume training symbols are all ones. Assuming h_l does not change during transmission of y_l and the training symbol[4], the received signal at the FC from CH$_l$ corresponding to the training symbol is:
$$\hat{z}_l = h_l \sqrt{\psi_l} + \nu_l, \quad l = 1, ..., L, \quad (5)$$
where ν_l in (5) is independent of v_l in (4) and is identically distributed. The FC adopts the following two-stage strategy to process the received signals from the CHs and reconstruct θ: stage 1) the FC uses the received signals $\{\hat{z}_l\}_{l=1}^L$ corresponding to training symbol transmissions to estimate $\{h_l\}_{l=1}^L$ and obtain the channel estimates $\{\hat{h}_l\}_{l=1}^L$, stage 2) the FC uses these channel estimates and the received signals $\{z_l\}_{l=1}^L$ corresponding to $\{y_l\}_{l=1}^L$ transmissions and find the LMMSE estimate of θ, denoted as $\hat{\theta}$. Finding the LMMSE estimator has a lower computational complexity, compared with the optimal MMSE estimator, and it requires only the knowledge of first and second order statistics. Let $D = \mathbb{E}\{(\theta - \hat{\theta})^2\}$ denote the MSE corresponding to the LMMSE estimator $\hat{\theta}$. Our main objective is to study power allocation among different clusters, subject to a network transmit power constraint (including power for training and data transmissions), such that D is minimized. Section II-B provides a formal description of our constrained optimization problem, including the power constraints and the set of our optimization variables[5].

B. Power Constraints

We describe our power constraints. Let $P_{l,k}$ denote the average power that sensor k consumes to send its amplified measurement to CH$_l$ and $P_l = \sum_{k=1}^{K_l} P_{l,k}$ be the total power that sensors in cluster l spend to send their amplified measurements to CH$_l$. From (2) we have:
$$P_{l,k} = \alpha_{l,k} \mathbb{E}\{x_{l,k}^2\} = \alpha_{l,k}(\sigma_\theta^2 + \sigma_{n_{l,k}}^2), \quad k = 1, ..., K_l. \quad (6)$$
For tractability, similar to [9], [29] we assume P_l is equally divided between sensors within cluster l, i.e., $P_{l,k} = P_l/K_l$. Under this assumption from (6) we obtain $\alpha_{l,k} = P_l d_{l,k}$ where $d_{l,k} = \frac{1}{K_l(\sigma_\theta^2 + \sigma_{n_{l,k}}^2)}$, or equivalently in matrix form, we find $\sqrt{\boldsymbol{A}_l} = \sqrt{P_l} \boldsymbol{D}_l$, where $\boldsymbol{D}_l$ is given in Table I. Let $\mathcal{P}_l$ represent the average power that CH$_l$ spends to send its fused signal y_l to the FC. We have:
$$\mathcal{P}_l = \mathbb{E}\{y_l^2\} = \boldsymbol{w}_l^T \underbrace{\mathbb{E}\{\boldsymbol{t}_l \boldsymbol{t}_l^T\}}_{=\boldsymbol{R}_{t_l}} \boldsymbol{w}_l. \quad (7)$$
Applying (3) and noting that $\boldsymbol{x}_l, \boldsymbol{q}_l$ in (3) are zero mean and uncorrelated, it is easy to verify that:
$$\boldsymbol{R}_{t_l} = P_l \boldsymbol{\Omega}_l + \boldsymbol{\Sigma}_{q_l}, \quad (8)$$
where
$$\boldsymbol{\Omega}_l = \boldsymbol{\Delta}_l + \sigma_\theta^2 \boldsymbol{\Pi}_l, \quad \boldsymbol{\Delta}_l = \boldsymbol{D}_l \boldsymbol{\Sigma}_{n_l} \boldsymbol{D}_l, \quad \boldsymbol{\Pi}_l = \boldsymbol{\rho}_l \boldsymbol{\rho}_l^T, \quad \boldsymbol{\rho}_l = \boldsymbol{D}_l \boldsymbol{1}_l.$$
Combining (7) and (8) we obtain:
$$\mathcal{P}_l = \boldsymbol{w}_l^T (P_l \boldsymbol{\Omega}_l + \boldsymbol{\Sigma}_{q_l}) \boldsymbol{w}_l. \quad (9)$$
Let $P_{trn} = \sum_{l=1}^L \psi_l$ be the total power that CHs spend to transmit their pilot symbols to the FC for channel estimation. We assume there is a constraint on the network transmit power, such that:

---

[3]When the pdf of θ is unknown, it is reasonable to assume that CH$_l$ applies a linear fusion rule $\boldsymbol{w}_l$ and we seek the best $\boldsymbol{w}_l$. In Section IV-A1 we show that $\boldsymbol{w}_l^{opt}$ is equal to the linear operator corresponding to the LMMSE estimation of θ based on $\boldsymbol{t}_l$, multiplied by an optimized scalar χ_l. When $\theta \sim N(0, \sigma_\theta^2)$ the MMSE and LMMSE estimates of θ based on $\boldsymbol{t}_l$ coincide.

[4]We assume time-division-duplex transmission and channel reciprocity. We also assume that the channel coherence time is larger than the overall duration of pilot transmission, channel estimation, power optimization, information feedback, and data transmission.

[5]Comparing orthogonal channel model and multiple-access channel (MAC) model adopted in [7], [13], [25], the former consumes more time or bandwidth for transmission, however, it does not require symbol-level synchronization for compensating complex channel phase at transmitter. We note that the complexity of the sequence of operation in our work (pilot transmission, channel estimation, power optimization, information feedback, and data transmission) is comparable with that of those works that rely upon perfect CSI, since implementing power allocation solutions obtained based on perfect CSI [6]–[8], [11], [13] requires pilot transmission and channel estimation, prior to data transmission.

$$P_{trn} + \sum_{l=1}^{L} P_l + \mathcal{P}_l \le P_{tot}. \tag{10}$$

Substituting $\mathcal{P}_l$ in (9) into the constraint in (10) we reach:

$$P_{trn} + \sum_{l=1}^{L} \boldsymbol{w}_l^T \boldsymbol{\Sigma}_{q_l} \boldsymbol{w}_l + P_l(1 + \boldsymbol{w}_l^T \boldsymbol{\Omega}_l \boldsymbol{w}_l) \le P_{tot}. \tag{11}$$

C. *Problem Statement*

Under the network power constraint in (10), our goal is to find the optimal $P_{trn}, \{P_l, \mathcal{P}_l\}_{l=1}^{L}$ such that D is minimized. The constraint in (11) shows that finding the optimal $\{P_l, \mathcal{P}_l\}_{l=1}^{L}$ in our problem is equivalent to finding the optimal $\{P_l, \boldsymbol{w}_l\}_{l=1}^{L}$, since given $\{P_l, \boldsymbol{w}_l\}_{l=1}^{L}$ one can find $\{\mathcal{P}_l\}_{l=1}^{L}$ using (9). Therefore, our goal is to find the optimal total training power P_{trn}, the optimal total power that sensors in cluster l spend to transmit their measurements to their CH P_l, and the optimal $\boldsymbol{w}_l$ employed by CH$_l$ for its linear fusion, such that D is minimized. In other words, we are interested in solving the following constrained optimization problem:

$$\min_{P_{trn}, \{P_l, \boldsymbol{w}_l\}_{l=1}^{L}} D(P_{trn}, \{P_l, \boldsymbol{w}_l\}_{l=1}^{L}) \tag{12}$$

$$\text{s.t.} \quad P_{trn} + \sum_{l=1}^{L} \boldsymbol{w}_l^T \boldsymbol{\Sigma}_{q_l} \boldsymbol{w}_l + P_l(1 + \boldsymbol{w}_l^T \boldsymbol{\Omega}_l \boldsymbol{w}_l) \le P_{tot},$$

$$P_{trn} \in \mathbb{R}^+, P_l \in \mathbb{R}^+, \boldsymbol{w}_l \in \mathbb{R}^{K_l}, \forall l.$$

We note that $\boldsymbol{\Sigma}_{q_l}$ and $\boldsymbol{\Omega}_l$ in the network transmit power constraint do not depend on our optimization variables.

III. CHARACTERIZING D AND ITS LOWER BOUNDS

A. *Characterization of D in terms of Channel Estimates*

We characterize the objective function D in (12), in terms of our optimization variables. Before delving into the derivations of D, we introduce the following notations. Considering our signal model in Section II, we define column vectors $\boldsymbol{x}$, $\boldsymbol{t}$, $\boldsymbol{y}$, $\boldsymbol{z}$ in Table I, which are obtained from stacking the signals corresponding to all clusters. We have:

$$\boldsymbol{x} = \theta \boldsymbol{1} + \boldsymbol{n}, \quad \boldsymbol{t} = \boldsymbol{M}\boldsymbol{x} + \boldsymbol{q}, \quad \boldsymbol{y} = \boldsymbol{W}\boldsymbol{t}, \tag{13a}$$

$$\boldsymbol{z} = \boldsymbol{H}\boldsymbol{y} + \boldsymbol{v}, \tag{13b}$$

where $\boldsymbol{1}$ is a column vector of K ones, column vectors $\boldsymbol{n}$, $\boldsymbol{q}$, $\boldsymbol{v}$, and matrices $\boldsymbol{M}$, $\boldsymbol{W}$, $\boldsymbol{H}$ are defined in Table I. The noise vectors $\boldsymbol{n}$, $\boldsymbol{q}$, $\boldsymbol{v}$ are zero-mean with covariance matrices $\boldsymbol{\Sigma}_n$, $\boldsymbol{\Sigma}_q$, $\boldsymbol{\Sigma}_v$, respectively, given in Table I. We model the fading coefficient as $h_l = \hat{h}_l + \tilde{h}_l$, where $\hat{h}_l$ is the MMSE channel estimate and $\tilde{h}_l$ is the corresponding zero-mean estimation error with the variance ζ_l^2. The expressions for $\hat{h}_l$ and ζ_l^2 in terms of training power ψ_l are [30]:

$$\hat{h}_l = \frac{\sigma_{h_l}^2 \sqrt{\psi_l} \hat{z}_l}{\sigma_{v_l}^2 + \psi_l \sigma_{h_l}^2}, \quad \zeta_l^2 = \frac{2\sigma_{h_l}^2 \sigma_{v_l}^2}{\sigma_{v_l}^2 + \psi_l \sigma_{h_l}^2}. \tag{14}$$

We define matrices $\hat{\boldsymbol{H}}, \tilde{\boldsymbol{H}}$ in Table I and thus we have $\boldsymbol{H} = \hat{\boldsymbol{H}} + \tilde{\boldsymbol{H}}$. Substituting this channel model into (13b), we can rewrite the received signal $\boldsymbol{z}$ as the following:

$$\boldsymbol{z} = \underbrace{[\hat{\boldsymbol{H}}\boldsymbol{W}\boldsymbol{M}\boldsymbol{1}]\theta}_{=\boldsymbol{z}_1} + \underbrace{(\tilde{\boldsymbol{H}}\boldsymbol{W}\boldsymbol{M}\boldsymbol{1})\theta}_{=\boldsymbol{z}_2} + \underbrace{(\hat{\boldsymbol{H}} + \tilde{\boldsymbol{H}})\boldsymbol{W}(\boldsymbol{q} + \boldsymbol{M}\boldsymbol{n}) + \boldsymbol{v}}_{=\boldsymbol{z}_3}. \tag{15}$$

We proceed with characterizing D in terms of the channel estimates. From optimal linear estimation theory, we have:

$$\hat{\theta} = \boldsymbol{g}^H \boldsymbol{z}, \quad \text{where } \boldsymbol{g} = (\mathbb{E}\{\boldsymbol{z}\boldsymbol{z}^H\})^{-1} \mathbb{E}\{\theta \boldsymbol{z}\},$$

$$D = \sigma_\theta^2 - \mathbb{E}\{\theta \boldsymbol{z}\}^H (\mathbb{E}\{\boldsymbol{z}\boldsymbol{z}^H\})^{-1} \mathbb{E}\{\theta \boldsymbol{z}\}. \tag{16}$$

where $\hat{\theta}$ and D depend on the channel estimates $\{\hat{h}_l\}_{l=1}^{L}$. In the following, we find $\mathbb{E}\{\boldsymbol{z}\boldsymbol{z}^H\}$ and $\mathbb{E}\{\theta \boldsymbol{z}\}$ in (16) by examining the statistics of channel estimation error. By the orthogonality principle of LMMSE estimation [31], $\tilde{h}_l$ is orthogonal to $\hat{h}_l$, that is $\mathbb{E}\{\tilde{h}_l \hat{h}_l\} = 0$, $\forall l$, and therefore, $\mathbb{E}\{\boldsymbol{z}_1 \boldsymbol{z}_2^H\} = \boldsymbol{0}$. Using the fact that θ, $\boldsymbol{n}$, $\boldsymbol{q}$, $\boldsymbol{v}$ are mutually uncorrelated, we have $\mathbb{E}\{\boldsymbol{z}_1 \boldsymbol{z}_3^H\} = \boldsymbol{0}$, $\mathbb{E}\{\boldsymbol{z}_2 \boldsymbol{z}_3^H\} = \boldsymbol{0}$. Combined these with the fact that $\mathbb{E}\{\boldsymbol{z}\} = \boldsymbol{0}$, the covariance matrix $\boldsymbol{C}_{\boldsymbol{z}} = \mathbb{E}\{\boldsymbol{z}\boldsymbol{z}^H\}$ given $\hat{\boldsymbol{H}}$ can be expressed as:

$$\boldsymbol{C}_{\boldsymbol{z}} = \sigma_\theta^2 \hat{\boldsymbol{H}}\boldsymbol{W}\boldsymbol{M}\boldsymbol{1}\boldsymbol{1}^T\boldsymbol{M}\boldsymbol{W}^T\hat{\boldsymbol{H}}^H + \sigma_\theta^2(\boldsymbol{\Gamma}\boldsymbol{W}\boldsymbol{M}\boldsymbol{\Sigma}\boldsymbol{M}\boldsymbol{W}^T\boldsymbol{\Gamma})$$
$$+ \hat{\boldsymbol{H}}\boldsymbol{W}(\boldsymbol{\Sigma}_q + \boldsymbol{M}\boldsymbol{\Sigma}_n\boldsymbol{M})\boldsymbol{W}^T\hat{\boldsymbol{H}}^H$$
$$+ \boldsymbol{\Gamma}\boldsymbol{W}(\boldsymbol{\Sigma}_q + \boldsymbol{M}\boldsymbol{\Sigma}_n\boldsymbol{M})\boldsymbol{W}^T\boldsymbol{\Gamma} + \boldsymbol{\Sigma}_v, \tag{17}$$

where $\boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma}$ are defined in Table I and $\boldsymbol{\Sigma}_l$ is a $K_l \times K_l$ matrix of all ones. We define $\boldsymbol{\Lambda}_{1_l}$ and $\boldsymbol{\Lambda}_2$ as bellow:

$$\boldsymbol{\Lambda}_{1_l} = \sigma_\theta^2 \zeta_l^2 P_l \boldsymbol{\Pi}_l + (|\hat{h}_l|^2 + \zeta_l^2)(\boldsymbol{\Sigma}_{q_l} + P_l \boldsymbol{\Delta}_l), \tag{18}$$

$$\boldsymbol{\Lambda}_2 = |\boldsymbol{\mu}||\boldsymbol{\mu}|^T, \quad \boldsymbol{\mu}_l = \sqrt{P_l} \hat{h}_l \boldsymbol{\rho}_l, \forall l.$$

where $\boldsymbol{\mu}$ is defined in Table I. It is straightforward to simplify (17) and write it as the following:

$$\boldsymbol{C}_{\boldsymbol{z}} = \boldsymbol{W}(\boldsymbol{\Lambda}_1 + \sigma_\theta^2 \boldsymbol{\Lambda}_2)\boldsymbol{W}^T + \boldsymbol{\Sigma}_v. \tag{19}$$

where $\boldsymbol{\Lambda}_1$ is defined in Table I. To find $\mathbb{E}\{\theta \boldsymbol{z}\}$ we consider (15) and we realize that $\mathbb{E}\{\theta \boldsymbol{z}_3\} = \boldsymbol{0}$. Therefore:

$$\mathbb{E}\{\theta \boldsymbol{z}\} = \mathbb{E}\{\theta \boldsymbol{z}_1\} + \mathbb{E}\{\theta \boldsymbol{z}_2\} \overset{(a)}{=} \sigma_\theta^2 \hat{\boldsymbol{H}}\boldsymbol{W}\boldsymbol{M}\boldsymbol{1} = \sigma_\theta^2 \boldsymbol{W}\boldsymbol{\mu}, \tag{20}$$

where (a) in (20) is obtained from the fact that $\mathbb{E}\{\tilde{\boldsymbol{H}}\} = \boldsymbol{0}$. Based on (19), (20), the LMMSE estimator $\hat{\theta}$ and its corresponding MSE in (16) can be written as:

$$\hat{\theta} = \sigma_\theta^2 \boldsymbol{\mu}^H \boldsymbol{W}^T \boldsymbol{C}_{\boldsymbol{z}}^{-1} \boldsymbol{z},$$

$$D = \sigma_\theta^2 - \sigma_\theta^4 \boldsymbol{\mu}^H \boldsymbol{W}^T \boldsymbol{C}_{\boldsymbol{z}}^{-1} \boldsymbol{W}\boldsymbol{\mu}. \tag{21}$$

in which $\boldsymbol{\mu}$ and $\boldsymbol{C}_{\boldsymbol{z}}$ depend on the channel estimates. Substituting (19) in (21) and applying the Woodbury identity[6] yields:

$$D = (\sigma_\theta^{-2} + \boldsymbol{\mu}^H \boldsymbol{W}^T (\boldsymbol{W}\boldsymbol{\Lambda}_1\boldsymbol{W}^T + \boldsymbol{\Sigma}_v)^{-1} \boldsymbol{W}\boldsymbol{\mu})^{-1}$$

$$= (\sigma_\theta^{-2} + \sum_{l=1}^{L} \frac{P_l|\hat{h}_l|^2 \boldsymbol{w}_l^T \boldsymbol{\Pi}_l \boldsymbol{w}_l}{\sigma_{v_l}^2 + \boldsymbol{w}_l^T \boldsymbol{\Lambda}_{1_l} \boldsymbol{w}_l})^{-1}, \tag{22}$$

Examining D in (22) we notice that $\boldsymbol{\Pi}_l$ does not depend on our optimization variables. However, $\boldsymbol{\Lambda}_{1_l}$ depends on P_l and ψ_l (through the channel estimate $|\hat{h}_l|^2$ and the channel estimation error variance ζ_l^2). Clearly, D depends on $\boldsymbol{w}_l$.

B. *Three Lower Bounds on D*

We provide three lower bounds on D, denoted as D_1, D_2, D_3. To obtain D_1 we consider the scenario when $\{h_l\}_{l=1}^{L}$ are available at the FC (perfect CSI). This implies $\hat{h}_l = h_l$ and $\zeta_l^2 = 0, \forall l$, in (22), and the MSE becomes:

$$D_1 = (\sigma_\theta^{-2} + \sum_{l=1}^{L} \frac{P_l|h_l|^2 \boldsymbol{w}_l^T \boldsymbol{\Pi}_l \boldsymbol{w}_l}{\sigma_{v_l}^2 + |h_l|^2 \boldsymbol{w}_l^T (\boldsymbol{\Sigma}_{q_l} + P_l \boldsymbol{\Delta}_l)\boldsymbol{w}_l})^{-1}. \tag{23}$$

---

[6]For matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}$ the Woodbury identity states that $(\boldsymbol{A} + \boldsymbol{B}\boldsymbol{C}\boldsymbol{D})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{C}^{-1} + \boldsymbol{D}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\boldsymbol{D}\boldsymbol{A}^{-1}$ [32].

To obtain D_2 we consider the scenario when in addition to perfect CSI, sensors' noisy measurement vector $\boldsymbol{x}_l$ is available at CH_l (i.e., error-free channels between sensors and their CHs). Therefore, $\boldsymbol{A}_l = \boldsymbol{I}_l$, where $\boldsymbol{I}_l$ denotes the identity matrix, and $\boldsymbol{\Sigma}_{q_l} = \boldsymbol{0}, \forall l$. In this scenario (23) simplifies to:

$$D_2 = (\sigma_\theta^{-2} + \sum_{l=1}^{L} \frac{|h_l|^2 \boldsymbol{w}_l^T \boldsymbol{\Sigma}_l \boldsymbol{w}_l}{\sigma_{v_l}^2 + |h_l|^2 \boldsymbol{w}_l^T \boldsymbol{\Sigma}_{n_l} \boldsymbol{w}_l})^{-1}. \quad (24)$$

To obtain D_3 we consider the scenario when $\boldsymbol{x}_l$ is available at CH_l and y_l is available at the FC. This is equivalent to having all measurements $\{\boldsymbol{x}_l\}_{l=1}^{L}$ available at the FC (i.e., error-free channels between sensors and their CHs, and between CHs and the FC). Therefore, the MSE becomes:

$$D_3 = (\sigma_\theta^{-2} + \sum_{l=1}^{L} \boldsymbol{1}_l^T \boldsymbol{\Sigma}_{n_l}^{-1} \boldsymbol{1}_l)^{-1}. \quad (25)$$

Clearly, we have $D_3 < D_2 < D_1 < D$.

C. Bayesian CRB

Let G denote the Bayesian Fisher information corresponding to estimating θ, given $\boldsymbol{z}$ and the vector of channel estimates $\hat{\boldsymbol{h}} = [\hat{h}_1, ..., \hat{h}_L]$ at the FC. The inverse of G is the Bayesian CRB and it sets an estimation-theoretic lower bound on the MSE of any Bayesian estimation of θ, given $\boldsymbol{z}, \hat{\boldsymbol{h}}$ [33]–[35]. Using the definition in [33]–[35] in our problem $G = \mathbb{E}\{(\frac{\partial \ln f(\boldsymbol{z}, \hat{\boldsymbol{h}}, \theta)}{\partial \theta})^2\}$, where $f(\boldsymbol{z}, \hat{\boldsymbol{h}}, \theta)$ denotes the joint pdf of $\boldsymbol{z}, \hat{\boldsymbol{h}}, \theta$ and and the expectation is taken over $f(\boldsymbol{z}, \hat{\boldsymbol{h}}, \theta)$.

Lemma 1. The Bayesian Fisher information corresponding to estimating θ, given $\boldsymbol{z}, \hat{\boldsymbol{h}}$ is:

$$G = \mathbb{E}\{G_1(\theta)\} + \mathbb{E}\{G_2(\theta)\}, \quad (26)$$

where $G_1(\theta) = -\frac{\partial^2 \ln f(\theta)}{\partial \theta^2}$ and $G_2(\theta)$ is given below. For $\theta \sim N(0, \sigma_\theta^2)$ we have $\mathbb{E}\{G_1(\theta)\} = \sigma_\theta^{-2}$. Both expectations in (26) are taken over $f(\theta)$, which represents the pdf of θ.

$$G_2(\theta) = \sum_{l=1}^{L} \int_{\hat{h}_l} \int_{z_l} \frac{f(\hat{h}_l)}{f(z_l|\hat{h}_l, \theta)} (\frac{\partial f(z_l|\hat{h}_l, \theta)}{\partial \theta})^2 dz_l d\hat{h}_l, \quad (27)$$

where $f(z_l|\hat{h}_l, \theta)$ and its derivative with respect to θ are:

$$f(z_l|\hat{h}_l, \theta) = a_1 e^{-a_2\theta^2} \sum_{m=0}^{\infty} \sum_{n=0}^{m} \sum_{p=0}^{m-n} c_{m,n,p}(\theta)$$
$$\times \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} s_{m,n,p,b}(\theta) \exp(-\frac{|z_l - b|^2}{2\sigma_{v_l}^2}) db, \quad (28)$$

$$\frac{\partial f(z_l|\hat{h}_l, \theta)}{\partial \theta} = a_1 e^{-a_2\theta^2} \sum_{m=0}^{\infty} \sum_{n=0}^{m} \sum_{p=0}^{m-n} [(\frac{m-n+p}{\theta} - 2a_2\theta)$$
$$\times c_{m,n,p}(\theta) \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} s_{m,n,p,b}(\theta) \exp(-\frac{|z_l - b|^2}{2\sigma_{v_l}^2}) db], \quad (29)$$

and the parameters $a_1, a_2, a_3, c_{m,n,p}(\theta), s_{m,n,p,b}(\theta), \bar{\phi}$ are:

$$a_1 = \frac{\exp(-|\hat{h}_l|^2/\zeta_l^2)}{\pi^2 \zeta_l^2 \bar{\sigma}_l^2 \sigma_{v_l}^2}, a_2 = \frac{a_3^2}{\bar{\sigma}_l^2}, a_3 = \boldsymbol{w}_l^T \sqrt{\boldsymbol{A}_l} \boldsymbol{1}_l, \quad (30)$$

$$c_{m,n,p}(\theta) = \frac{|\hat{h}_l|^{m+n-p} |a_3\theta|^{m-n+p}}{m!n!p!(m-n-p)!\zeta_l^{2m+n-p} \bar{\sigma}_l^{2m-n+p}},$$

$$s_{m,n,p,b}(\theta) = |b|^m K_{n-p}(\frac{2|b|}{\bar{\sigma}_l \zeta_l})(2\cos(\bar{\phi} - \frac{\pi}{2}(1 - \operatorname{sgn}(a_3\theta))))^{m-n-p},$$

$$\bar{\phi} = \angle b - \angle \hat{h}_l, \quad \bar{\sigma}_l^2 = \boldsymbol{w}_l^T (\sqrt{\boldsymbol{A}_l} \boldsymbol{\Sigma}_{n_l} \sqrt{\boldsymbol{A}_l} + \boldsymbol{\Sigma}_{q_l}) \boldsymbol{w}_l.$$

Proof. See Appendix A. □

IV. Solving the Constrained Minimization of D

We consider the constrained optimization problem in (12), where D is provided in (22). We define:

$$\mathcal{J}_l(P_{trn}, P_l, \boldsymbol{w}_l) = \frac{P_l |\hat{h}_l|^2 \boldsymbol{w}_l^T \boldsymbol{\Pi}_l \boldsymbol{w}_l}{\sigma_{v_l}^2 + \boldsymbol{w}_l^T \boldsymbol{\Lambda}_{1_l} \boldsymbol{w}_l}, \quad (31)$$
$$C_l(P_l, \boldsymbol{w}_l) = \boldsymbol{w}_l^T \boldsymbol{\Sigma}_{q_l} \boldsymbol{w}_l + P_l(1 + \boldsymbol{w}_l^T \boldsymbol{\Omega}_l \boldsymbol{w}_l).$$

Using the two definitions in (31) we can replace the problem in (12) with its equivalent, problem (P1), that has a simpler presentation. In particular, we can write $D^{-1} = \sigma_\theta^{-2} + \sum_{l=1}^{L} \mathcal{J}_l(P_{trn}, P_l, \boldsymbol{w}_l)$. Hence, problem (P1) becomes:

$$(P1) \quad \max_{P_{trn}, \{P_l, \boldsymbol{w}_l\}_{l=1}^{L}} \quad \sum_{l=1}^{L} \mathcal{J}_l(P_{trn}, P_l, \boldsymbol{w}_l)$$

$$\text{s.t.} \quad P_{trn} + \sum_{l=1}^{L} C_l(P_l, \boldsymbol{w}_l) \le P_{tot}, P_{trn}, P_l \in \mathbb{R}^+, \boldsymbol{w}_l \in \mathbb{R}^{K_l}, \forall l.$$

It is easy to show that the solution of (P1) holds with active constraint $P_{trn} + \sum_{l=1}^{L} C_l(P_l, \boldsymbol{w}_l) = P_{tot}$. We further note that due to the cap on the network transmit power, only a subset of the clusters may become active at each observation period. We refer to this active subset as $\mathcal{A} = \{l : P_l > 0, l = 1, \ldots, L\}$, where $|\mathcal{A}| \le L$. Regarding the objective function $\mathcal{J}_l$ in (P1) we note that it depends on $\hat{h}_l$ (through $|\hat{h}_l|^2$ in the numerator and $\boldsymbol{\Lambda}_{1_l}$ in the denominator of (31)). Regarding the optimization variables in (P1) we notice that, since pilot transmission proceeds data transmission, P_{trn} cannot depend on the channel estimates $\{\hat{h}_l\}_{l=1}^{L}$ and can only depend on the statistical information of communication channels and the observation model. Examining (P1), we note however, that solving it for P_{trn} provides an answer that depends on $\hat{h}_l$ (which is unrealizable). On the other hand, the variables $P_l, \boldsymbol{w}_l$ should be chosen according to the available CSI $\hat{h}_l$. Based on these observations, we propose to consider two problems (P$_A$) and (P$_B$) stemming from (P1). problem (P$_A$) finds the optimal $\{P_l, \boldsymbol{w}_l\}_{l=1}^{L}$ that minimizes D, given P_{trn}. Let $\sigma \in (0, 1)$ such that $P_{trn} = (1 - \sigma)P_{tot}$. Given P_{trn} (and thus σ), we define $\mathcal{F}_l(P_l, \boldsymbol{w}_l) = \mathcal{J}_l(P_{trn}, P_l, \boldsymbol{w}_l)$. Problem (P$_A$) becomes:

$$(P_A) \quad \text{given } P_{trn}, \quad \max_{\{P_l, \boldsymbol{w}_l\}_{l=1}^{L}} \quad \sum_{l=1}^{L} \mathcal{F}_l(P_l, \boldsymbol{w}_l)$$

$$\text{s.t.} \quad \sum_{l=1}^{L} C_l(P_l, \boldsymbol{w}_l) \le \sigma P_{tot}, \ P_l \in \mathbb{R}^+, \boldsymbol{w}_l \in \mathbb{R}^{K_l}, \forall l.$$

Section IV-A is devoted to solving (P$_A$). Problem (P$_B$) finds the optimal P_{trn} that, instead of minimizing D, it minimizes a modified objective function $\mathbb{E}\{D\}$, where an average is taken

over the channel estimates. In Section IV-B we address (P$_B$) and find P_{trn} as well as training power distribution $\{\psi_l\}_{l=1}^L$ among the CHs such that $\sum_{l=1}^L \psi_l = P_{trn}$.

A. Finding Optimal $\{P_l, \boldsymbol{w}_l\}_{l=1}^L$ Given Total Training Power

We start with (P$_A$). By taking the second derivative of $\sum_{l=1}^L \mathcal{F}_l(P_l, \boldsymbol{w}_l)$ w.r.t $\{P_l, \boldsymbol{w}_l\}$, it is straightforward to show that (P$_A$) is not jointly concave over the optimization variables. Alternatively, we propose a solution approach that converges to a stationary point of (P$_A$). Problem (P$_A$) contains the constraint $\sum_{l=1}^L C_l(P_l, \boldsymbol{w}_l) \leq \sigma P_{tot}$, which is referred to as coupling or complicating constraint in the literature [36]. By introducing additional auxiliary variables $\{\mathcal{V}_l\}_{l=1}^L$, problem (P$_A$) becomes:

$$(P2) \quad \text{given } P_{trn}, \max_{\{\mathcal{V}_l, P_l, \boldsymbol{w}_l\}_{l=1}^L} \sum_{l=1}^L \mathcal{F}_l(P_l, \boldsymbol{w}_l)$$

$$\text{s.t. } C_l(P_l, \boldsymbol{w}_l) \leq \mathcal{V}_l, \sum_{l=1}^L \mathcal{V}_l \leq \sigma P_{tot}, \mathcal{V}_l, P_l \in \mathbb{R}^+, \boldsymbol{w}_l \in \mathbb{R}^{K_l}, \forall l.$$

Note that the auxiliary variable $\mathcal{V}_l$ represents the total amount of power allocated to cluster l (for sensors within cluster l to transmit their observations to CH$_l$ and for CH$_l$ to transmit y_l to the FC). According to the *primal decomposition* [36], problem (P2) can be decomposed as the following:

$$(SP2\text{-}1) \quad \text{given } P_{trn}, \mathcal{V}_l, \max_{P_l, \boldsymbol{w}_l} \mathcal{F}_l(P_l, \boldsymbol{w}_l)$$

$$\text{s.t. } C_l(P_l, \boldsymbol{w}_l) \leq \mathcal{V}_l, P_l \in \mathbb{R}^+, \boldsymbol{w}_l \in \mathbb{R}^{K_l},$$

$$(SP2\text{-}2) \quad \text{given } P_{trn}, \{P_l, \boldsymbol{w}_l\}_{l=1}^L, \max_{\{\mathcal{V}_l\}_{l=1}^L} \sum_{l=1}^L \mathcal{F}_l^{opt}$$

$$\text{s.t. } \sum_{l=1}^L \mathcal{V}_l \leq \sigma P_{tot}, \mathcal{V}_l \in \mathbb{R}^+, \forall l,$$

where $\mathcal{F}_l^{opt}$ denotes the maximum of $\mathcal{F}_l(P_l, \boldsymbol{w}_l)$, which depends on $\mathcal{V}_l$. The solution can be reached by iteratively solving sub-problems (SP2-1) and (SP2-2). In the following, we provide the detailed solutions for (SP2-1) and (SP2-2).

1) Solving Optimization Problem (SP2-1): We start with a brief overview of this section. Let $\boldsymbol{w}_l^{opt}, P_l^{opt}$ denote the solution of (SP2-1). We will show how to compute $\boldsymbol{w}_l^{opt}$ in terms of P_l using (42) and how to compute P_l^{opt} in terms of $\boldsymbol{w}_l$ using (47). Having two equations (42), (47), we substitute $\boldsymbol{w}_l$ from (42) into (47) to reach (48), which is a function of P_l^{opt} only. Employing a numerical line search method we obtain P_l^{opt} from (48). Having P_l^{opt}, we find $\boldsymbol{w}_l^{opt}$ using (42). The detailed explanations follow.

Examining $\mathcal{F}_l(P_l, \boldsymbol{w}_l)$ and $C_l(P_l, \boldsymbol{w}_l)$ expressions given in (31), it is evident that scaling up equally $P_l, \boldsymbol{w}_l$ increases both $\mathcal{F}_l(P_l, \boldsymbol{w}_l)$ and $C_l(P_l, \boldsymbol{w}_l)$. Therefore, (SP2-1) is equivalent to its converse formulation, where $C_l(P_l, \boldsymbol{w}_l)$ is minimized subject to a constraint on $\mathcal{F}_l(P_l, \boldsymbol{w}_l)$:

$$(CSP2\text{-}1) \quad \text{given } P_{trn}, \mathcal{U}_l, \min_{P_l, \boldsymbol{w}_l} C_l(P_l, \boldsymbol{w}_l)$$

$$\text{s.t. } \mathcal{F}_l(P_l, \boldsymbol{w}_l) \geq \mathcal{U}_l, P_l \in \mathbb{R}^+, \boldsymbol{w}_l \in \mathbb{R}^{K_l}.$$

Let C_l^{opt} be the minimum of $C_l(P_l, \boldsymbol{w}_l)$, which depends on $\mathcal{U}_l$. To solve (CSP2-1) we simplify its constraint by substituting

$\boldsymbol{\Lambda}_{1_l}$ from (18) into $\mathcal{F}_l(P_l, \boldsymbol{w}_l)$ in (31). Let $\boldsymbol{B}_l = \sigma_\theta^2 \zeta_l^2 \boldsymbol{\Pi}_l + (|\hat{h}_l|^2 + \zeta_l^2) \boldsymbol{\Delta}_l$. The constraint in (CSP2-1) becomes:

$$P_l \boldsymbol{w}_l^T (|\hat{h}_l|^2 \boldsymbol{\Pi}_l - \mathcal{U}_l \boldsymbol{B}_l) \boldsymbol{w}_l - (|\hat{h}_l|^2 + \zeta_l^2) \mathcal{U}_l \boldsymbol{w}_l^T \boldsymbol{\Sigma}_{q_l} \boldsymbol{w}_l - \sigma_{v_l}^2 \mathcal{U}_l \geq 0. \tag{32}$$

Consider (CSP2-1) where its constraint is now replaced with the inequality in (32). To solve (CSP2-1) we use the Lagrange multiplier method. Let $\mathcal{L}(\gamma, \eta, P_l, \boldsymbol{w}_l)$ be the Lagrangian for this problem and γ and η be the lagrange multipliers for the constraint in (32) and the constraint $P_l \geq 0$, respectively. Equation (33) shows $\mathcal{L}(\gamma, \eta, P_l, \boldsymbol{w}_l)$. The corresponding Karush-Kuhn-Tucker (KKT) optimality conditions are [37, pp. 243-244]:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{w}_l} = [\boldsymbol{R}_{t_l} + \gamma((|\hat{h}_l|^2 + \zeta_l^2)\mathcal{U}_l \boldsymbol{\Sigma}_{q_l} - P_l(|\hat{h}_l|^2 \boldsymbol{\Pi}_l - \mathcal{U}_l \boldsymbol{B}_l))]\boldsymbol{w}_l = \boldsymbol{0}; \tag{34a}$$

$$\gamma(P_l \boldsymbol{w}_l^T(|\hat{h}_l|^2 \boldsymbol{\Pi}_l - \mathcal{U}_l \boldsymbol{B}_l)\boldsymbol{w}_l - (|\hat{h}_l|^2 + \zeta_l^2)\mathcal{U}_l \boldsymbol{w}_l^T \boldsymbol{\Sigma}_{q_l} \boldsymbol{w}_l - \sigma_{v_l}^2 \mathcal{U}_l) = 0; \tag{34b}$$

$$\frac{\partial \mathcal{L}}{\partial P_l} = 1 + \boldsymbol{w}_l^T \boldsymbol{\Omega}_l \boldsymbol{w}_l - \gamma \boldsymbol{w}_l^T (|\hat{h}_l|^2 \boldsymbol{\Pi}_l - \mathcal{U}_l \boldsymbol{B}_l)\boldsymbol{w}_l - \eta = 0; \tag{34c}$$

$$\eta P_l = 0, \tag{34d}$$

where $\boldsymbol{R}_{t_l}$, defined in (8), depends on P_l. Similar to the solution of (P1), one can show that the solutions of (SP2-1) and (CSP2-1) must satisfy the equality constraints $C_l(P_l, \boldsymbol{w}_l) = \mathcal{V}_l$ and $\mathcal{F}_l(P_l, \boldsymbol{w}_l) = \mathcal{U}_l$ (or equivalently (34b)), respectively. Thus we find:

$$\boldsymbol{w}_l^T \boldsymbol{R}_{t_l} \boldsymbol{w}_l = \mathcal{V}_l - P_l, \tag{35a}$$

$$\boldsymbol{w}_l^T [P_l(|\hat{h}_l|^2 \boldsymbol{\Pi}_l - \mathcal{U}_l \boldsymbol{B}_l) - (|\hat{h}_l|^2 + \zeta_l^2)\mathcal{U}_l \boldsymbol{\Sigma}_{q_l}] \boldsymbol{w}_l = \sigma_{v_l}^2 \mathcal{U}_l. \tag{35b}$$

Combining (35a) and (35b) we reach:

$$\boldsymbol{w}_l^T [\boldsymbol{R}_{t_l} + \frac{\mathcal{V}_l - P_l}{\sigma_{v_l}^2 \mathcal{U}_l}((|\hat{h}_l|^2 + \zeta_l^2)\mathcal{U}_l \boldsymbol{\Sigma}_{q_l} - P_l(|\hat{h}_l|^2 \boldsymbol{\Pi}_l - \mathcal{U}_l \boldsymbol{B}_l))]\boldsymbol{w}_l = \boldsymbol{0}. \tag{36}$$

From (34a) and (36) we find the lagrange multiplier γ:

$$\gamma = \frac{\mathcal{V}_l - P_l}{\sigma_{v_l}^2 \mathcal{U}_l}. \tag{37}$$

• Computing $\boldsymbol{w}_l^{opt}$ given P_l: Substituting (37) into (34a) and conducting some mathematical manipulations result in:

$$\mathcal{U}_l[\underbrace{\frac{\sigma_{v_l}^2 \boldsymbol{R}_{t_l}}{\mathcal{V}_l - P_l} + (|\hat{h}_l|^2 + \zeta_l^2)\boldsymbol{\Sigma}_{q_l} + P_l \boldsymbol{B}_l}_{=\boldsymbol{\mathcal{B}}_1}]\boldsymbol{w}_l = |\boldsymbol{\mu}_l||\boldsymbol{\mu}_l|^T \boldsymbol{w}_l, \tag{38}$$

where $\boldsymbol{\mu}_l$ is defined in (18). Since $\boldsymbol{R}_{t_l} \succ \boldsymbol{0}, \boldsymbol{\Sigma}_{q_l} \succ \boldsymbol{0}, \boldsymbol{B}_l \succ \boldsymbol{0}$, the matrix $\boldsymbol{\mathcal{B}}_1$ is positive definite and full rank and hence invertible. Multiplying both sides of (38) with $\boldsymbol{\mathcal{B}}_1^{-1}$, we find:

$$\mathcal{U}_l \boldsymbol{w}_l = \boldsymbol{\mathcal{B}}_1^{-1} |\boldsymbol{\mu}_l||\boldsymbol{\mu}_l|^T \boldsymbol{w}_l. \tag{39}$$

Also, multiplying both sides of (38) with $\frac{1}{\mathcal{U}_l} \boldsymbol{R}_{t_l}^{-1}$ we reach:

$$\frac{\sigma_{v_l}^2}{\mathcal{V}_l - P_l} \boldsymbol{w}_l = \boldsymbol{R}_{t_l}^{-1}[\underbrace{\frac{|\boldsymbol{\mu}_l||\boldsymbol{\mu}_l|^T}{\mathcal{U}_l} - (|\hat{h}_l|^2 + \zeta_l^2)\boldsymbol{\Sigma}_{q_l} - P_l \boldsymbol{B}_l}_{=\boldsymbol{\mathcal{B}}_2}]\boldsymbol{w}_l. \tag{40}$$

Inspecting (39) and (40), and aiming at finding vector $\boldsymbol{w}_l$, we realize that (39) and (40) are ordinary eigenvalue problems. Since the solutions to (SP2-1) and (CSP2-1) satisfy the equality constraints $C_l(P_l, \boldsymbol{w}_l) = \mathcal{V}_l$ and $\mathcal{F}_l(P_l, \boldsymbol{w}_l) = \mathcal{U}_l$,

respectively, from (39) and (40) we find:

$$\mathcal{F}_l^{opt}=\lambda_{max}(\boldsymbol{\mathcal{B}}_1^{-1}|\boldsymbol{\mu}_l||\boldsymbol{\mu}_l|^T), \ C_l^{opt}=\frac{\sigma_{v_l}^2}{\lambda_{max}(\boldsymbol{\mathcal{B}}_2)}+P_l. \quad (41)$$

Let $\boldsymbol{s}_l^{opt}$ be the eigenvector corresponding to $\lambda_{max}(\boldsymbol{\mathcal{B}}_1^{-1}|\boldsymbol{\mu}_l||\boldsymbol{\mu}_l|^T)$. We note that $\mathcal{F}_l^{opt}$ is achieved when $\boldsymbol{w}_l$ is an appropriately scaled version of $\boldsymbol{s}_l^{opt}$, i.e., $\boldsymbol{w}_l^{opt}=r_l\boldsymbol{s}_l^{opt}$, where scalar r_l such that (35a) is satisfied. Also recall $\boldsymbol{\Pi}_l=\boldsymbol{\rho}_l\boldsymbol{\rho}_l^T$ is rank-1. Thus $\mathcal{F}_l^{opt}$ is the only non-zero eigenvalue of $\boldsymbol{\mathcal{B}}_1^{-1}|\boldsymbol{\mu}_l||\boldsymbol{\mu}_l|^T$ and $\boldsymbol{s}_l^{opt}$ is the corresponding eigenvector. Proposition 1 gives expressions for $\boldsymbol{w}_l^{opt}$ and $\mathcal{F}_l^{opt}$ in terms of P_l.

Proposition 1. Considering problem (SP2-1), the optimal fusion vector $\boldsymbol{w}_l^{opt}$ and the maximum value of the objective function $\mathcal{F}_l^{opt}$ in terms of P_l are:

$$\boldsymbol{w}_l^{opt}=\sqrt{\frac{\mathcal{V}_l-P_l}{\tau_l}}\boldsymbol{R}_{t_l}^{-1}\boldsymbol{\rho}_l, \quad \mathcal{F}_l^{opt}=\frac{|\hat{h}_l|^2\beta_l P_l\tau_l}{\sigma_{v_l}^2(1+\frac{\beta_l}{\mathcal{V}_l-P_l})}, \quad (42)$$

where $\tau_l=\boldsymbol{\rho}_l^T\boldsymbol{R}_{t_l}^{-1}\boldsymbol{\rho}_l$, $\beta_l=\frac{\sigma_{v_l}^2}{|\hat{h}_l|^2(1-\sigma_\theta^2 P_l\tau_l)+\zeta_l^2}$.

Proof. See Appendix B. $\qquad\square$

For our system model $\boldsymbol{R}_{t_l\theta}=\mathbb{E}\{\theta t_l\}=\sigma_\theta^2\sqrt{P_l}\boldsymbol{\rho}_l$. Hence, we can rewrite $\boldsymbol{w}_l^{opt}$ in (42) as:

$$\boldsymbol{w}_l^{opt}=\underbrace{\sigma_\theta^{-2}\sqrt{\frac{\mathcal{V}_l-P_l}{P_l\tau_l}}}_{=\chi_l}(\boldsymbol{R}_{t_l}^{-1}\boldsymbol{R}_{t_l\theta}). \quad (43)$$

Since $\boldsymbol{R}_{t_l}^{-1}\boldsymbol{R}_{t_l\theta}$ is the linear operator corresponding to the LMMSE estimator, (43) implies that the optimal linear fusion rule at CH$_l$ is equal to the linear operator corresponding to the LMMSE estimation of θ based on $\boldsymbol{t}_l$, multiplied by the amplification factor χ_l.

• Computing P_l^{opt} given $\boldsymbol{w}_l$: Note that (34d) results in $\eta=0$ for active clusters with $P_l>0$. Letting $\eta=0$ in (34c) and solving for γ we find:

$$\gamma=\frac{1+\boldsymbol{w}_l^T\boldsymbol{\Omega}_l\boldsymbol{w}_l}{\boldsymbol{w}_l^T(|\hat{h}_l|^2\boldsymbol{\Pi}_l-\mathcal{U}_l\boldsymbol{B}_l)\boldsymbol{w}_l}. \quad (44)$$

Equating (44) with (37) and solving for $\mathcal{U}_l$ we get:

$$\mathcal{U}_l=\frac{(\mathcal{V}_l-P_l)|\hat{h}_l|^2\boldsymbol{w}_l^T\boldsymbol{\Pi}_l\boldsymbol{w}_l}{\sigma_{v_l}^2(1+\boldsymbol{w}_l^T\boldsymbol{\Omega}_l\boldsymbol{w}_l)+(\mathcal{V}_l-P_l)\boldsymbol{w}_l^T\boldsymbol{B}_l\boldsymbol{w}_l}. \quad (45)$$

On the other hand, solving (35b) for $\mathcal{U}_l$ results in:

$$\mathcal{U}_l=\frac{P_l|\hat{h}_l|^2\boldsymbol{w}_l^T\boldsymbol{\Pi}_l\boldsymbol{w}_l}{\sigma_{v_l}^2+(|\hat{h}_l|^2+\zeta_l^2)\boldsymbol{w}_l^T\boldsymbol{\Sigma}_{q_l}\boldsymbol{w}_l+P_l\boldsymbol{w}_l^T\boldsymbol{B}_l\boldsymbol{w}_l}. \quad (46)$$

Combining (45) and (46), we obtain P_l^{opt} in terms of $\boldsymbol{w}_l$ as the following:

$$P_l^{opt}=\frac{\mathcal{V}_l(\sigma_{v_l}^2+(|\hat{h}_l|^2+\zeta_l^2)\boldsymbol{w}_l^T\boldsymbol{\Sigma}_{q_l}\boldsymbol{w}_l)}{\sigma_{v_l}^2(2+\boldsymbol{w}_l^T\boldsymbol{\Omega}_l\boldsymbol{w}_l)+(|\hat{h}_l|^2+\zeta_l^2)\boldsymbol{w}_l^T\boldsymbol{\Sigma}_{q_l}\boldsymbol{w}_l}. \quad (47)$$

At this point, we have obtained two equations: (42) provides $\boldsymbol{w}_l^{opt}$ in terms of P_l, and (47) provides P_l^{opt} in terms of $\boldsymbol{w}_l$. Substituting $\boldsymbol{w}_l^{opt}$ from (42) in (47) yields in:

$$\sigma_{v_l}^2(\mathcal{V}_l-2P_l^{opt})\tau_l+(|\hat{h}_l|^2+\zeta_l^2)(\mathcal{V}_l-P_l^{opt})^2\boldsymbol{\rho}_l^T\boldsymbol{R}_{t_l}^{-1}\boldsymbol{\Sigma}_{q_l}\boldsymbol{R}_{t_l}^{-1}\boldsymbol{\rho}_l$$
$$-P_l^{opt}\sigma_{v_l}^2(\mathcal{V}_l-P_l^{opt})\boldsymbol{\rho}_l^T\boldsymbol{R}_{t_l}^{-1}\boldsymbol{\Omega}_l\boldsymbol{R}_{t_l}^{-1}\boldsymbol{\rho}_l=0. \quad (48)$$

Note that $\tau_l,\boldsymbol{R}_{t_l}$ in (48) depend on P_l^{opt}, and thus, a closed-form solution for P_l^{opt} remains elusive. One can employ a line search method (e.g., the Golden section method [38, p. 216]) to solve (48) in the interval $(0,\mathcal{V}_l)$. Having P_l^{opt} we find $\boldsymbol{w}_l^{opt}$ using (42).

2) Solving Optimization Problem (SP2-2): By substituting $\mathcal{F}_l^{opt}$ from (42) in the objective function, problem (SP2-2) becomes:

$$\text{given } P_{trn},\{P_l,\boldsymbol{w}_l\}_{l=1}^L, \ \max_{\{\mathcal{V}_l\}_{l=1}^L} \ \sum_{l=1}^L\frac{|\hat{h}_l|^2\beta_l P_l\tau_l}{\sigma_{v_l}^2(1+\frac{\beta_l}{\mathcal{V}_l-P_l})}$$

$$\text{s.t.} \ \sum_{l=1}^L\mathcal{V}_l\leq\sigma P_{tot},\mathcal{V}_l\in\mathbb{R}^+,\forall l. \quad (49)$$

The maximization problem in (49) is concave and its solution can be found via solving the KKT conditions. In particular, we find (see Appendix C for derivations):

$$\mathcal{V}_l^{opt}=\left[\beta_l(\frac{|\hat{h}_l|}{\sigma_{v_l}}\sqrt{\frac{P_l\tau_l}{\lambda}}-1)\right]^++P_l, \quad (50a)$$

$$\lambda=(\frac{\sum_{l\in\mathcal{A}}\frac{|\hat{h}_l|\beta_l\sqrt{P_l\tau_l}}{\sigma_{v_l}}}{\sigma P_{tot}-\sum_{l\in\mathcal{A}}P_l+\sum_{l\in\mathcal{A}}\beta_l})^2. \quad (50b)$$

Note that the first term of the right side of the equality in (50a) is $\mathcal{P}_l$ introduced in Section II-B. Given $\lambda,|\hat{h}_l|,\sigma_{v_l}$ in (50a) and the easy-to-prove fact that $\tau_l+P_l\frac{\partial\tau_l}{\partial P_l}>0$, it is straightforward to show that $\frac{\partial\mathcal{P}_l}{\partial P_l}>0$ for active clusters, i.e., increasing P_l increases $\mathcal{P}_l$. Having the solutions to problems (SP2-1) and (SP2-2), Algorithm 1 summarizes our proposed solution to problem (P$_A$). Essentially, this algorithm iteratively solves (SP2-1) and (SP2-2) in a block-coordinate ascent manner until the convergence is reached. In Section VII, we argue that the algorithm output converges to a stationary point of (P$_A$).

B. Finding Optimal Total Training Power and its Distribution Among CHs

In this section, we focus on (P$_B$) and find P_{trn} as well as training power distribution $\{\psi_l\}_{l=1}^L$ among the CHs such that $\sum_{l=1}^L\psi_l=P_{trn}$. As we mentioned earlier, to find P_{trn} we consider a modified objective function, i.e., instead of $\sum_{l=1}^L\mathcal{J}_l$ in (P1) we consider $\sum_{l=1}^L\mathbb{E}\{\mathcal{J}_l\}$, where the expectation is taken over the channel estimates $|\hat{h}_l|^2$. Since solving this problem analytically is still intractable, we use the Jensen's inequality for concave functions [37, pp. 77-78], to establish a lower bound on $\mathbb{E}\{\mathcal{J}_l(P_{trn},P_l,\boldsymbol{w}_l)\}$:

$$\mathbb{E}\{\mathcal{J}_l(P_{trn},P_l,\boldsymbol{w}_l)\}\leq\mathcal{G}_l(P_{trn},P_l,\boldsymbol{w}_l),$$

where $\mathcal{G}_l(P_{trn},P_l,\boldsymbol{w}_l)$ is obtained from $\mathcal{J}_l(P_{trn},P_l,\boldsymbol{w}_l)$, after replacing $|\hat{h}_l|^2$ with $\mathbb{E}\{|\hat{h}_l|^2\}$. To find $\mathbb{E}\{|\hat{h}_l|^2\}$ needed for $\mathcal{G}_l(P_{trn},P_l,\boldsymbol{w}_l)$ we revisit the error corresponding to the LMMSE channel estimation in (14). Note that $\hat{h}_l$ is a zero-mean complex Gaussian. Let $2\sigma_{\hat{h}_l}^2$ denote the variance of $\hat{h}_l$. For the model $h_l=\hat{h}_l+\tilde{h}_l$, we invoke the orthogonality principle from the linear estimation theory [30], that states

$$\mathcal{L}(\gamma,\eta,P_l,\boldsymbol{w}_l)=\boldsymbol{w}_l^T\boldsymbol{\Sigma}_{q_l}\boldsymbol{w}_l+P_l(1+\boldsymbol{w}_l^T\boldsymbol{\Omega}_l\boldsymbol{w}_l)+\gamma((|\hat{h}_l|^2+\zeta_l^2)\mathcal{U}_l\boldsymbol{w}_l^T\boldsymbol{\Sigma}_{q_l}\boldsymbol{w}_l+\sigma_{v_l}^2\mathcal{U}_l-P_l\boldsymbol{w}_l^T(|\hat{h}_l|^2\boldsymbol{\Pi}_l-\mathcal{U}_l\boldsymbol{B}_l)\boldsymbol{w}_l)-\eta P_l, \quad (33)$$

Algorithm 1: proposed solution of (P_A)

Input: $P_{tot}, P_{trn}, \{\hat{h}_l\}_{l=1}^L, \epsilon$, and system parameters defined in Section II

Output: optimal optimization variables $\{P_l^{opt}, \boldsymbol{w}_l^{opt}\}_{l=1}^L$

- -

Let i indicate the iteration index, $\mathcal{V}_l^{(i)}, P_l^{(i)}, \boldsymbol{w}_l^{(i)}, \mathcal{A}^{(i)}$ denote $\mathcal{V}_l, P_l, \boldsymbol{w}_l, \mathcal{A}$ values and $\mathcal{F}^{(i)} = \sum_{l \in \mathcal{A}^{(i)}} \mathcal{F}_l(P_l^{(i)}, \boldsymbol{w}_l^{(i)})$ at iteration i.
- Given the channel estimates, sort the clusters as described in Appendix C.
- Initialization: $i=1$, $\mathcal{A}^{(0)} = \{1, ..., L\}$, randomly choose $\{P_l^{(0)}, \boldsymbol{w}_l^{(0)}\}_{l=1}^L$ such that $0 < P_l^{(0)} < \mathcal{V}_l^{(0)} = \frac{P_{tot} - P_{trn}}{L}$ and (P2) holds with active constraints, and compute $\mathcal{F}^{(0)}$.
- Iterate between solving (SP2-1) and (SP2-2) until convergence. At iteration i do below:
  1: Obtain $P_l^{(i)} \in (0, \mathcal{V}_l^{(i-1)})$ via solving (48), substitute $P_l^{(i)}$ into (42) to obtain $\boldsymbol{w}_l^{(i)}$, compute $\mathcal{F}^{(i)}$.
  2: If $|\frac{\mathcal{F}^{(i)} - \mathcal{F}^{(i-1)}}{\mathcal{F}^{(i-1)}}| \leq \epsilon$, terminate the iteration and return the optimal solution $\{P_l^{opt} = P_l^{(i)}, \boldsymbol{w}_l^{opt} = \boldsymbol{w}_l^{(i)}\}_{\forall l \in \mathcal{A}^{(i)}}$ and $\{P_l^{opt} = 0, \boldsymbol{w}_l^{opt} = \boldsymbol{0}\}_{\forall l \notin \mathcal{A}^{(i)}}$.
  3: Increase i, update $\mathcal{A}^{(i)}$, and find $\{\mathcal{V}_l^{(i+1)}\}_{\forall l \in \mathcal{A}^{(i)}}$ using (50a), (50b).
- Continue the iteration until the stopping criteria in step 2 is met.

$var(\hat{h}_l) = var(h_l) - var(\tilde{h}_l) = 2\sigma_{h_l}^2 - \zeta_l^2$, where ζ_l^2 in (14) depends on ψ_l. Since $\hat{h}_l$ is zero-mean, we have $\mathbb{E}\{|\hat{h}_l|^2\} = var(\hat{h}_l)$. Thus, $\mathcal{G}_l(P_{trn}, P_l, \boldsymbol{w}_l) = \frac{(2\sigma_{h_l}^2 - \zeta_l^2)P_l \boldsymbol{w}_l^T \boldsymbol{\Pi}_l \boldsymbol{w}_l}{\sigma_{v_l}^2 + \boldsymbol{w}_l^T \boldsymbol{\Lambda}_{1_l} \boldsymbol{w}_l}$, where $\boldsymbol{\Lambda}_{1_l} = \sigma_\theta^2 \zeta_l^2 P_l \boldsymbol{\Pi}_l + 2\sigma_{h_l}^2 (\boldsymbol{\Sigma}_{q_l} + P_l \boldsymbol{\Delta}_l)$. Notice that $\mathcal{G}_l$ depends on the optimization variable P_{trn} through ζ_l^2 in the numerator and $\boldsymbol{\Lambda}_{1_l}$ in the denominator. We reconsider (P1) in which $\mathcal{J}_l$ is now replaced with $\mathcal{G}_l$:

$(P_{B'}) \quad \max_{P_{trn}, \{P_l, \boldsymbol{w}_l\}_{l=1}^L} \quad \sum_{l=1}^L \mathcal{G}_l(P_{trn}, P_l, \boldsymbol{w}_l)$

s.t. $\quad P_{trn} + \sum_{l=1}^L C_l(P_l, \boldsymbol{w}_l) \leq P_{tot}, P_{trn}, P_l \in \mathbb{R}^+, \boldsymbol{w}_l \in \mathbb{R}^{K_l}, \forall l.$

Examining $(P_{B'})$, we realize that solving it for P_{trn} provides an answer that depends on $P_l, \boldsymbol{w}_l$ (which is undesirable). To circumvent this problem we propose a method to find P_{trn} based on the following observation. We observe that, although $(P_{B'})$ is a non-concave maximization problem, given $\{P_l, \boldsymbol{w}_l\}_{l=1}^L$ and letting $\sigma = 1 - \frac{P_{trn}}{P_{tot}}$, the problem $(P_{B'})$ under these conditions becomes strictly concave with respect to the variable σ over the interval $(0, 1)$, and hence, the objective function has a unique global maximum in this interval. Let σ^{opt} denote the solution to this problem, which can be efficiently found using numerical line search methods

(e.g., the Golden section[7] method [38, p. 216]). Since this problem is concave over $(0, 1)$, the convergence of Golden section method to σ^{opt} is guaranteed.

Based on the above observation, we propose the method described in Algorithm 2 to solve $(P_{B'})$ and find P_{trn}^{opt}. The proposed method is basically Golden section method, where in each iteration we apply Algorithm 1 to find $\{P_l, \boldsymbol{w}_l\}_{l=1}^L$, only for the purpose of successively narrowing the search interval for σ. The output of Algorithm 2 converges to σ^{opt} and thus $P_{trn}^{opt} = (1 - \sigma^{opt})P_{tot}$.

Algorithm 2: proposed solution of $(P_{B'})$

Input: P_{tot}, ϵ, system parameters defined in Section II

Output: optimal optimization variable σ^{opt}

- -

Apply the iterative Golden section method to find $\sigma^{opt} \in (0, 1)$
- Initialization: $i=0, \sigma_b^{(0)} = 0, \sigma_e^{(0)} = 1$.
- At iteration i of Golden section method, do below:
  1: Compute two evaluating points $\alpha_b^{(i)}$ and $\alpha_e^{(i)}$ using the starting and the ending points of the search interval $(\sigma_b^{(i)}, \sigma_e^{(i)})$.
  2: For each evaluating point, use Algorithm 1 to obtain $\{P_l, \boldsymbol{w}_l\}_{l=1}^L$ and compute the objective function $\sum_{l=1}^L \mathcal{G}_l$. Suppose $\mathcal{G}_b^{(i)}, \mathcal{G}_e^{(i)}$ denote the values of $\sum_{l=1}^L \mathcal{G}_l$ when it is evaluated at $\alpha_b^{(i)}$ and $\alpha_e^{(i)}$, respectively.
- Depending on the values of $\mathcal{G}_b^{(i)}, \mathcal{G}_e^{(i)}$ update the search interval $(\sigma_b^{(i)}, \sigma_e^{(i)})$.
- Continue the iteration until $\sigma_e^{(i)} - \sigma_b^{(i)} \leq \epsilon$.

Given P_{trn}^{opt}, we find $\{\psi_l\}_{l=1}^L$, via minimizing the MSE of the LMMSE channel estimates for all clusters:

$$\text{given } P_{trn}^{opt} \quad \min_{\{\psi_l\}_{l=1}^L} \quad \sum_{l=1}^L \zeta_l^2 \tag{51}$$

$$\text{s.t.} \quad \sum_{l=1}^L \psi_l \leq P_{trn}^{opt}, \psi_l \in \mathbb{R}^+, \forall l.$$

The above is a convex minimization problem. Solving the associated KKT conditions, we obtain:

---

[7]Let x^{opt} denote the maximum value that a concave function $f(x)$ attains over a search interval $x \in (x_b, x_e)$. This numerical method finds x^{opt} via successively narrowing the range of the search interval. Let i be the iteration index, $(\mathcal{I}_b^{(i)}, \mathcal{I}_e^{(i)})$ be the starting and ending points of the search interval at iteration i, $\alpha_b^{(i)} = 0.382(\mathcal{I}_e^{(i)} - \mathcal{I}_b^{(i)}) + \mathcal{I}_b^{(i)}$ and $\alpha_e^{(i)} = 0.618(\mathcal{I}_e^{(i)} - \mathcal{I}_b^{(i)}) + \mathcal{I}_b^{(i)}$ be the evaluating points. Also let $f_b^{(i)}, f_e^{(i)}$, denote the values of the function $f(x)$ when it is evaluated at the evaluating points $\alpha_b^{(i)}, \alpha_e^{(i)}$, respectively. For initialization, we let $i=0, \mathcal{I}_b^{(0)} = x_b, \mathcal{I}_e^{(0)} = x_e$. At iteration i, we compute $f_b^{(i)}$ and $f_e^{(i)}$ and then update the search interval to find x^{opt} as the following: if $f_b^{(i)} > f_e^{(i)}$, then $\mathcal{I}_b^{(i+1)} = \mathcal{I}_b^{(i)}, \mathcal{I}_e^{(i+1)} = \alpha_e^{(i)}$, if $f_b^{(i)} = f_e^{(i)}$, then $\mathcal{I}_b^{(i+1)} = \alpha_b^{(i)}, \mathcal{I}_e^{(i+1)} = \alpha_e^{(i)}$, and if $f_b^{(i)} < f_e^{(i)}$, then $\mathcal{I}_b^{(i+1)} = \alpha_b^{(i)}, \mathcal{I}_e^{(i+1)} = \mathcal{I}_e^{(i)}$. As the stopping criterion, we check whether the length of the search interval exceeds a pre-determined threshold ϵ. If the stopping criterion is met at iteration j, the algorithm returns the optimal solution $x^{opt} = \mathcal{I}_b^{(j)}$. Otherwise, we update the search interval and continue the iterations until the stopping criterion is met.

Algorithm 1: the output of this algorithm is $\{P_l^{opt}, \boldsymbol{w}_l^{opt}\}_{l=1}^L$

Fig. 2: This block diagram is the pictorial narrative of our approach to solve the original constrained optimization problem (P1).

$$\psi_l = \Big[\frac{\sigma_{v_l}^2}{\sigma_{h_l}^2}\Big(\frac{\sigma_{h_l}^2}{\kappa\sigma_{v_l}} - 1\Big)\Big]^+, \quad \kappa = \frac{\sum_{l=1}^L \sigma_{v_l}}{P_{trn}^{opt} + \sum_{l=1}^L \frac{\sigma_{v_l}^2}{\sigma_{h_l}^2}}. \quad (52)$$

The solution in (52) is based on the assumption that all CHs participate in pilot transmission and P_{trn}^{opt} satisfies the inequality $P_{trn}^{opt} \geq \frac{\sigma_{v_L}}{\sigma_{h_L}^2}\sum_{l=1}^L \sigma_{v_l} - \sum_{l=1}^L \frac{\sigma_{v_l}^2}{\sigma_{h_l}^2} = \Upsilon$. However, when $P_{trn}^{opt} < \Upsilon$, the solutions in (52) imply that $\psi_l = 0$ for some clusters. In this case, we propose to choose $\psi_l = a\frac{\sigma_{h_l}}{\sigma_{v_l}}$, in which a is a common factor. Imposing the constraint $\sum_{l=1}^L \psi_l = P_{trn}^{opt}$ results in:

$$\psi_l = \frac{\sigma_{h_l}^2 P_{trn}^{opt}}{\sigma_{v_l}\sum_{l=1}^L \frac{\sigma_{h_l}^2}{\sigma_{v_l}}}, \quad l = 1, ..., L, \text{ when } P_{trn}^{opt} < \Upsilon. \quad (53)$$

Fig. 2 shows a block diagram that summarizes our approach to solve the original constrained optimization problem (P1). Overall, the sequence of algorithm implementations and network operation follow. The FC implements Algorithm 2 to obtain P_{trn}^{opt}, and consequently to find $\{\psi_l\}_{l=1}^L$ given in (52). The FC feeds back this information to CHs (all the obtained $\{P_l, \boldsymbol{w}_l\}_{l=1}^L$ values during the execution of Algorithm 2 are discarded at this point). CHs send their pilot symbols to the FC and the FC estimates the channels $\{\hat{h}_l\}_{l=1}^L$. Now, given $P_{trn}^{opt}, \{\hat{h}_l\}_{l=1}^L$, the FC implements Algorithm 1, finds $\{P_l^{opt}, \boldsymbol{w}_l^{opt}\}_{l=1}^L$, feeds back[8] this new information to CHs, and feeds back $P_{l,k} = \frac{P_l^{opt}}{K_l}$ to sensors. Sensors send their amplified measurements to their CHs. CHs send their fused signals to the FC. Finally, the FC estimates θ.

C. Minimizing Lower Bounds on MSE D

This section discusses constrained minimization of the lower bounds D_1, D_2 we derived in Section III-B. The lower bound D_1 depends on $\{P_l, \boldsymbol{w}_l\}_{l=1}^L$ and hence its constrained minimization becomes:

---

[8]Similar to [24], [25] we assume that the FC energy resource is much larger than those of the sensors/CHs. Therefore, the overhead required for feeding back the necessary information from the FC to the sensors/CHs is neglected.

$$\text{(P3)} \quad \max_{\{P_l, \boldsymbol{w}_l\}_{l=1}^L} \quad \sum_{l=1}^L \frac{P_l|h_l|^2 \boldsymbol{w}_l^T \boldsymbol{\Pi}_l \boldsymbol{w}_l}{\sigma_{v_l}^2 + |h_l|^2 \boldsymbol{w}_l^T(\boldsymbol{\Sigma}_{q_l} + P_l\boldsymbol{\Delta}_l)\boldsymbol{w}_l}$$

$$\text{s.t.} \quad \sum_{l=1}^L \boldsymbol{w}_l^T\boldsymbol{\Sigma}_{q_l}\boldsymbol{w}_l + P_l(1+\boldsymbol{w}_l^T\boldsymbol{\Omega}_l\boldsymbol{w}_l) \leq P_{tot}, P_l \in \mathbb{R}^+, \boldsymbol{w}_l \in \mathbb{R}^{K_l}, \forall l.$$

This is similar to (P2), with the difference that $P_{trn} = 0$, and hence in (50a) and (50b) expressions we let $\zeta_l^2 = 0, |\hat{h}_l|^2 = |h_l|^2, \sigma = 1$. Algorithm 1 can be followed to find the solution to (P3), using $\boldsymbol{w}_l^{opt}$ in (42). The lower bound D_2 depends on $\{\boldsymbol{w}_l\}_{l=1}^L$ and hence its constrained minimization becomes:

$$\text{(P4)} \quad \max_{\{\boldsymbol{w}_l\}_{l=1}^L} \quad \sum_{l=1}^L \frac{|h_l|^2 \boldsymbol{w}_l^T \boldsymbol{\Sigma}_l \boldsymbol{w}_l}{\sigma_{v_l}^2 + |h_l|^2 \boldsymbol{w}_l^T \boldsymbol{\Sigma}_{n_l} \boldsymbol{w}_l}$$

$$\text{s.t.} \quad \sum_{l=1}^L \boldsymbol{w}_l^T(\sigma_\theta^2\boldsymbol{\Sigma}_l + \boldsymbol{\Sigma}_{n_l})\boldsymbol{w}_l \leq P_{tot}, \boldsymbol{w}_l \in \mathbb{R}^{K_l}, \forall l.$$

This is similar to (P2), with the differences that $P_{trn} = 0$ and $P_l = 0, \forall l$. Following similar steps we took in Section IV-A to solve (P2), we find that (50a) and (50b) become:

$$\mathcal{V}_l^{opt} = \Big[\beta_l''\Big(\frac{|h_l|}{\sigma_{v_l}}\sqrt{\frac{\tau_l'}{\lambda}} - 1\Big)\Big]^+, \quad \beta_l'' = \frac{\sigma_{v_l}^2}{|h_l|^2(1-\sigma_\theta^2\tau_l')},$$

$$\lambda = \Big(\frac{\sum_{l\in\mathcal{A}}\frac{|h_l|\beta_l''\sqrt{\tau_l'}}{\sigma_{v_l}}}{P_{tot}+\sum_{l\in\mathcal{A}}\beta_l''}\Big)^2, \quad \tau_l' = \mathbf{1}_l^T(\sigma_\theta^2\boldsymbol{\Sigma}_l + \boldsymbol{\Sigma}_{n_l})^{-1}\mathbf{1}_l.$$

The optimal weight vector $\boldsymbol{w}_l^{opt}$ corresponding to the solution of (P4) is computed as $\boldsymbol{w}_l^{opt} = \sqrt{\frac{\mathcal{V}_l}{\tau_l'}}(\sigma_\theta^2\boldsymbol{\Sigma}_l + \boldsymbol{\Sigma}_{n_l})^{-1}\mathbf{1}_l$.

V. SOLVING THE SPECIAL CASES OF THE ORIGINAL PROBLEM

The original problem (P1) aims at constrained minimization of D, with respect to three sets of optimization variables: P_{trn} total training power, P_l power allocated to sensors in cluster l to send their measurements to CH_l, and $\mathcal{P}_l$ power allocated to CH_l to transmit its signal to the FC. To untangle the performance gain that optimizing each set of these optimization variables provides, we consider the following three special cases of (P1). In problem (P1-SC1) assuming

P_{trn} is given and $\psi_l = P_{trn}/L$, we optimize $\{P_l, \mathcal{P}_l\}_{l=1}^L$. In problem (P1-SC2) assuming $P_l = P, \forall l$, we optimize $P_{trn}, P, \{\mathcal{P}_l\}_{l=1}^L$. In problem (P1-SC3) assuming $\mathcal{P}_l = \mathcal{P}, \forall l$, we optimize $P_{trn}, \mathcal{P}, \{P_l\}_{l=1}^L$. Note that problem (P1-SC1) is the same as problem (P$_A$) addressed in Section IV-A. In the following we address problems (P1-SC2) and (P1-SC3).

A. Solving Special Case (P1-SC2)*: When Intra-Cluster Powers of all Clusters are Equal*

Problem (P1-SC2) becomes:

$$\text{(P1-SC2)} \quad \max_{P_{trn}, P, \{\boldsymbol{w}_l\}_{l=1}^L} \quad \sum_{l=1}^L \frac{P|\hat{h}_l|^2 \boldsymbol{w}_l^T \boldsymbol{\Pi}_l \boldsymbol{w}_l}{\sigma_{v_l}^2 + \boldsymbol{w}_l^T \boldsymbol{\Lambda}_{1_l} \boldsymbol{w}_l}$$

$$\text{s.t. } P_{trn} + \sum_{l=1}^L \boldsymbol{w}_l^T \boldsymbol{\Sigma}_{q_l} \boldsymbol{w}_l + P(1 + \boldsymbol{w}_l^T \boldsymbol{\Omega}_l \boldsymbol{w}_l) \leq P_{tot}, P_{trn}, P \in \mathbb{R}^+,$$

where $\boldsymbol{\Lambda}_{1_l} = \sigma_\theta^2 \zeta_l^2 P \boldsymbol{\Pi}_l + (|\hat{h}_l|^2 + \zeta_l^2)(\boldsymbol{\Sigma}_{q_l} + P\boldsymbol{\Delta}_l)$. To address (P1-SC2) we consider the following two sub-problems: (a) finding $P^*, \{\boldsymbol{w}_l^*\}_{l=1}^L$ given P_{trn}, (b) finding P_{trn}^* as well as $\{\psi_l^*\}_{l=1}^L$ such that $\sum_{l=1}^L \psi_l^* = P_{trn}^*$. Sub-problem (a) is a special case of (P$_A$) in which, for finding P^*, we use Golden section method, and sub-problem (b) is similar to (P$_B$). Recall that $P_{trn} = (1 - \sigma) P_{tot}$ and thus $\sum_{l=1}^L (P + \mathcal{P}_l) = \sigma P_{tot}$. We let $\sigma_c \in (0, 1)$ such that $P = (1 - \sigma_c) \sigma P_{tot}$. It is simple to show that sub-problems (a) and (b) are both concave and hence P^* and P_{trn}^* are unique. Next, we summarize our proposed solutions for solving sub-problems (a) and (b) in Algorithms 3-a and 3-b, respectively.

Description of Algorithm 3-a: Let $P^* = (1 - \sigma_c^*) \sigma P_{tot}$ denote the optimal P. We apply Golden section method to find $\sigma_c^* \in (0, 1)$ and thus P^* that maximizes the objective function in (P1-SC2), denoted as $\mathcal{F}(\sigma_c)$. At iteration i, for each evaluating point we first compute the optimal $\mathcal{V}_l^{(i)}$, denoted as $\{\bar{\mathcal{V}}_l^{(i)}\}_{l=1}^L$ using (50a), and substitute $\bar{\mathcal{V}}_l^{(i)}$ into (42) to obtain $\{\bar{\boldsymbol{w}}_l^{(i)}\}_{l=1}^L$. Next we compute $\mathcal{F}_b^{(i)}$ and $\mathcal{F}_e^{(i)}$. The stopping criterion is similar to Algorithm 2. Algorithm 3-a returns the optimal $\sigma_c^*, \{\boldsymbol{w}_l^*\}_{l=1}^L$.

Description of Algorithm 3-b: We address sub-problem (b) similar to problem (P$_B$) in Section IV-B. More specifically, we consider problem (P$_{B'}$), where P_l is substituted by P, and apply a modified version of Algorithm 2 to solve it. In particular, at iteration i of Algorithm 2, we use Algorithm 3-a to obtain the optimal variables $\bar{P}^{(i)}, \{\bar{\boldsymbol{w}}_l^{(i)}\}_{l=1}^L$, and then compute $\mathcal{R}_b^{(i)}$ and $\mathcal{R}_e^{(i)}$. The rest is similar to Algorithm 2. Algorithm 3-b returns the optimal $P_{trn}^*, \{\psi_l^*\}_{l=1}^L$.

B. Solving Special Case (P1-SC3)*: When Powers of all CHs for Their Data Transmission to the FC are Equal*

To incorporate the constraint $\mathcal{P}_l = \mathcal{P}$ in the cost function of problem (P1-SC3), from Section IV-A1 we recall that $\boldsymbol{w}_l^{opt} = \chi_l (\boldsymbol{R}_{t_l}^{-1} \sigma_\theta^2 \sqrt{P_l} \boldsymbol{\rho}_l)$. Therefore from $\mathcal{P}_l = \boldsymbol{w}_l^T \boldsymbol{R}_{t_l} \boldsymbol{w}_l$ in (9) and $\mathcal{P}_l = \mathcal{P}$, we conclude $\chi_l^2 = \mathcal{P}/\sigma_\theta^4 P_l \tau_l$. Substituting for $\boldsymbol{w}_l$ in (P1), problem (P1-SC3) becomes:

$$\text{(P1-SC3)} \max_{P_{trn}, \mathcal{P}, \{P_l\}_{l=1}^L} \sum_{l=1}^L \frac{P_l |\hat{h}_l|^2 \tau_l}{\frac{\sigma_{v_l}^2}{\mathcal{P}} + \zeta_l^2 + \frac{|\hat{h}_l|^2}{\tau_l} \boldsymbol{\rho}_l^T \boldsymbol{R}_{t_l}^{-1} \boldsymbol{\Sigma}_{q_l} \boldsymbol{R}_{t_l}^{-1} \boldsymbol{\rho}_l}$$

$$\text{s.t. } P_{trn} + \sum_{l=1}^L (P_l + \mathcal{P}) \leq P_{tot}, P_{trn}, \mathcal{P} \in \mathbb{R}^+, P_l \in \mathbb{R}^+, \forall l.$$

To address (P1-SC3) we consider the following two sub-problems: (a) finding $\mathcal{P}^*, \{P_l^*\}_{l=1}^L$ given P_{trn}, (b) finding P_{trn}^* as well as $\{\psi_l^*\}_{l=1}^L$ such that $\sum_{l=1}^L \psi_l^* = P_{trn}^*$. Sub-problem (a) is a special case of (P$_A$) in which, for finding $\mathcal{P}^*$, we use Golden section method, and sub-problem (b) is similar to (P$_B$). We let $\sigma_d \in (0, 1)$ such that $\mathcal{P} = (1 - \sigma_d) \sigma P_{tot}$. It is easy to show that finding $\mathcal{P}^*, P_{trn}^*$ in sub-problems (a) and (b), respectively, are concave problems, and hence $\mathcal{P}^*$ and P_{trn}^* are unique. In Appendix D, we prove that finding $\{P_l^*\}_{l=1}^L$ in sub-problem (a) is jointly concave over P_l's and therefore its solution is unique. In the absence of a closed form expression we use gradient-ascent algorithm to find the solution. Algorithms 4-a and 4-b summarize how we solve sub-problems (a) and (b), respectively.

Description of Algorithm 4-a: Let $\mathcal{P}^* = (1 - \sigma_d^*) \sigma P_{tot}$ denote the optimal $\mathcal{P}$. We apply Golden section method to find $\sigma_d^* \in (0, 1)$ and thus $\mathcal{P}^*$ that maximizes the objective function in (P1-SC3), denoted as $\mathcal{F}(\sigma_d)$. At iteration i, for each evaluating point we compute the optimal $P_l^{(i)}$, denoted as $\{\bar{P}_l^{(i)}\}_{l=1}^L$ using gradient-ascent algorithm, and substitute them in (P1-SC3) to compute $\mathcal{F}_b^{(i)}$ and $\mathcal{F}_e^{(i)}$. The stopping criterion is similar to Algorithm 2. Algorithm 4-a returns the optimal $\sigma_d^*, \{P_l^*\}_{l=1}^L$.

Description of Algorithm 4-b: We address sub-problem (b) similar to problem (P$_B$) in Section IV-B. Specifically, we consider problem (P$_{B'}$), where $\mathcal{P}_l$ is substituted by $\mathcal{P}$ and apply a modified version of Algorithm 2 to solve it. In particular, at iteration i of Algorithm 2, we use Algorithm 4-a to obtain the optimal variables $\bar{\mathcal{P}}^{(i)}, \{\bar{P}_l^{(i)}\}_{l=1}^L$, and then compute $\mathcal{R}_b^{(i)}$ and $\mathcal{R}_e^{(i)}$. The rest is similar to Algorithm 2. Algorithm 4-b returns the optimal $P_{trn}^*, \{\psi_l^*\}_{l=1}^L$.

VI. COMPLEXITY OF ALGORITHMS

We discuss the computational complexity of Golden section method as well as Algorithms 1, 2, 3-a, 3-b, 4-a, 4-b, which allows us to compare the computational complexity of solving (P1) versus those of (P1-SC1), (P1-SC2), (P1-SC3).

• Golden section method: This method includes a one-dimensional search to find the optimal point. If no matrix inversion is required, its complexity order for convergence to an ϵ-accurate solution is $\bar{\epsilon}$, where $\bar{\epsilon} = \log(1/\epsilon)$ [38, p. 217]. We use this method for solving (48). In each iteration, to compute the left side of (48) we employ the matrix inversion algorithm in [39] to calculate $\boldsymbol{R}_{t_l}^{-1}$ with complexity order of $\mathcal{O}(K_l^{2.37})$. Therefore, the overall complexity order of finding $P_l^{opt} \in (0, \mathcal{V}_l)$ becomes $\mathcal{O}(\bar{\epsilon} K_l^{2.37})$.

• Algorithm 1 for solving (P$_A$): We switch between solving (SP2-1) and (SP2-2) until the stopping criteria is met. In each iteration, we need to (i) find $\{P_l\}_{l=1}^L$ using Golden section method, with the overall complexity order of $\mathcal{O}(\bar{\epsilon}\bar{K})$, where $\bar{K} = \sum_{l=1}^L K_l^{2.37}$, and (ii) calculate $\{\mathcal{V}_l\}_{l=1}^L$ using (50), which needs τ_l, β_l that are found in (i) and hence, the complexity order of finding $\{\mathcal{V}_l\}_{l=1}^L$ is $\mathcal{O}(L)$. The overall complexity order of Algorithm 1 becomes $\mathcal{O}(\bar{\epsilon}(L + \bar{\epsilon}\bar{K}))$.

• Algorithm 2 for solving (P$_{B'}$): In each iteration, for each evaluating point we use Algorithm 1 to obtain $\{P_l, \boldsymbol{w}_l\}_{l=1}^L$. Therefore, the overall complexity order of Algorithm 2 becomes $\mathcal{O}(\bar{\epsilon}^2(L + \bar{\epsilon}\bar{K}))$.

- Algorithm 3-a for solving sub-problem (a) of (P1-SC2): In each iteration, for each evaluating point computing τ_l in (50), (42) involves the matrix inversion $\boldsymbol{R}_{t_l}^{-1}$, and thus, the complexity order of finding $\{\mathcal{V}_l\}_{l=1}^{L}$ and then $\{\boldsymbol{w}_l\}_{l=1}^{L}$ is $\mathcal{O}(\bar{K})$. Therefore, the overall complexity order of Algorithm 3-a is $\mathcal{O}(\bar{\epsilon}\bar{K})$.

- Algorithm 3-b for solving sub-problem (b) of (P1-SC2): In each iteration, for each evaluating point we use Algorithm 3-a to obtain $P, \{\boldsymbol{w}_l\}_{l=1}^{L}$. Therefore, the overall complexity order of Algorithm 3-b is $\mathcal{O}(\bar{\epsilon}^2 \bar{K})$.

- Algorithm 4-a for solving sub-problem (a) of (P1-SC3): Note that the complexity order of the gradient-ascent algorithm to maximize a general non-smooth convex function $f(x)$ and converge to an ϵ-accurate solution is $\mathcal{O}(1/\epsilon)$, if no matrix inversion is required for finding $f(x)$ and its gradient $\nabla f(x)$ [38, p. 232]. In each iteration of Algorithm 4-a, for each evaluating point, since computing the objective function in (P1-SC3) and its derivative with respect to P_l involves the matrix inversion $\boldsymbol{R}_{t_l}^{-1}$, the complexity order of finding $\{P_l\}_{l=1}^{L}$ using the gradient-ascent algorithm is $\mathcal{O}(\bar{K}/\epsilon)$. Therefore, the overall complexity order of Algorithm 4-a becomes $\mathcal{O}(\bar{\epsilon}\bar{K}/\epsilon)$.

- Algorithm 4-b for solving sub-problem (b) of (P1-SC3): In each iteration, for each evaluating point we use Algorithm 4-a to obtain $\mathcal{P}, \{P_l\}_{l=1}^{L}$. Therefore, the overall complexity order of Algorithm 4-b is $\mathcal{O}(\bar{\epsilon}^2 \bar{K}/\epsilon)$.

To solve (P1) we need to solve (P$_A$), (P$_{B'}$). Therefore, the complexity order of solving (P1) is $e_0 = \mathcal{O}(\bar{\epsilon}(1+\bar{\epsilon})(L+\bar{\epsilon}\bar{K}))$. To solve (P1-SC1) we need to solve (P$_A$). Therefore, the complexity order of solving (P1-SC1) is $e_1 = \mathcal{O}(\bar{\epsilon}(L+\bar{\epsilon}\bar{K}))$. To solve (P1-SC2) we need to solve sub-problems (a) and (b) of (P1-SC2). Therefore, the complexity order of solving (P1-SC2) is $e_2 = \mathcal{O}(\bar{\epsilon}(1+\bar{\epsilon})\bar{K})$. To solve (P1-SC3) we need to solve sub-problems (a) and (b) of (P1-SC3). Therefore, the complexity order of solving (P1-SC3) is $e_3 = \mathcal{O}(\bar{\epsilon}(1+\bar{\epsilon})\bar{K}/\epsilon)$ It is clear that $e_1 < e_2 < e_0 < e_3$.

VII. CONVERGENCE ANALYSIS

We discuss the convergence analysis of Algorithms 1 and 2 which solve problems (P$_A$) and (P$_{B'}$), respectively.

- **Convergence of Algorithm 1**: Problems (P$_A$) and (P2) are equivalent. In (P2), the cost function is non-concave and the constraint is a closed convex set w.r.t. the optimization variables $\{\mathcal{V}_l, P_l, \boldsymbol{w}_l\}_{l=1}^{L}$. Algorithm 1 is indeed a block-coordinate ascent type algorithm with two blocks. The first block solves (SP2-1) for all clusters to obtain $\{P_l, \boldsymbol{w}_l\}_{l=1}^{L}$. (SP2-1) is a non-concave maximization problem for which we have a numerical solution for P_l using Golden Section method and a closed-form solution for $\boldsymbol{w}_l$. Since (SP2-1) is a non-concave maximization problem, we cannot claim that our proposed solution for $P_l, \boldsymbol{w}_l$ is globally optimal and unique. The second block solves (SP2-2) to obtain $\{\mathcal{V}_l\}_{l=1}^{L}$. (SP2-2) is a concave maximization problem for which we have a closed-form solution for $\mathcal{V}_l$. Since (SP2-2) is a concave maximization problem, its solution is globally optimal and unique.

The authors in [40] proved that in a block-coordinate descent algorithm with only two blocks, which solves the unconstrained minimization problem

$$\min_{(\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}} f(\boldsymbol{x}_1, \boldsymbol{x}_2), \qquad (54)$$

given we have the global minimizer $\boldsymbol{x}_1^{(k+1)} = \underset{\boldsymbol{x}_1}{\arg\min}\ f(\boldsymbol{x}_1, \boldsymbol{x}_2^{(k)}), \forall k$, the algorithm converges to a stationary point if we can find a point $\boldsymbol{x}_2^{(k+1)}$ such that $f(\boldsymbol{x}_1^{(k+1)}, \boldsymbol{x}_2^{(k+1)}) \leq f(\boldsymbol{x}_1^{(k+1)}, \boldsymbol{x}_2^{(k)})$ and $\nabla_2 f(\boldsymbol{x}_1^{(k+1)}, \boldsymbol{x}_2^{(k+1)}) = 0, \forall k$. The authors also proved the convergence when f in (54) is minimized subject to a convex constraint set (see Corollary 1 and Section 4 in [41]). Equipped with this result from [40], [41], we return to our own problem. Let $f = -\sum_{l=1}^{L} \mathcal{F}_l$, $\boldsymbol{x}_1 = \{\mathcal{V}_l\}_{l=1}^{L}$, $\boldsymbol{x}_2 = \{P_l, \boldsymbol{w}_l\}_{l=1}^{L}$ in problem (P2). When solving problem (P2) using the block-coordinate method with two blocks, we note that (SP2-2) has a globally optimal solution and thus $\boldsymbol{x}_1^{(k+1)} = \arg\min_{\boldsymbol{x}_1} f(\boldsymbol{x}_1, \boldsymbol{x}_2^{(k)})$ is completely known. Also, our proposed solution for (SP2-1) satisfies the condition $\nabla_2\ f(\boldsymbol{x}_1^{(k+1)}, \boldsymbol{x}_2^{(k+1)}) = 0$ (because it is the solution of KKT conditions for (SP2-1)). Furthermore, our extensive simulations indicate that the condition $f(\boldsymbol{x}_1^{(k+1)}, \boldsymbol{x}_2^{(k+1)}) \leq f(\boldsymbol{x}_1^{(k+1)}, \boldsymbol{x}_2^{(k)})$ is always satisfied $\forall k$. Hence, we conclude that the output of the block-coordinate ascent method between (SP2-1) and (SP2-2) converges to a stationary point.

Regarding the convergence speed of the block-coordinate descent method, few works have obtained a convergence rate under special conditions on f in (54). However, for the general case of non-convex f, even under convex constraints, no convergence rate is established in the literature. Our extensive simulations indicate that the average number of iterations needed for Algorithm 1 to converge to an ϵ-accurate solution for $\{P_l^{opt}, \boldsymbol{w}_l^{opt}\}_{l=1}^{L}$ is 30.

- **Convergence of Algorithm 2**: In this algorithm, we employ Golden section method to obtain P_{trn}^{opt}, where in each iteration we apply Algorithm 1 to find $\{P_l, \boldsymbol{w}_l\}_{l=1}^{L}$, only for the purpose of successively narrowing the search interval of Golden section method. Consider solving the following non-convex minimization problem under convex constraints:

$$\min_{x, \boldsymbol{y}}\ f(x, \boldsymbol{y})$$
$$\text{s.t.}\quad x \in \Omega_1 \subseteq \mathbb{R},\ \boldsymbol{y} \in \Omega_2 \subseteq \mathbb{R}^{n_2}, \qquad (55)$$

where Golden section method is used to obtain x^{opt}. If $\boldsymbol{y}^{(k)} = \arg\min_{\boldsymbol{y}} f(x^{(k)}, \boldsymbol{y})$ given $x^{(k)},\ k = 0, 1, ...$ is known instantly, Golden section method converges linearly, and the rate of convergence is approximately 0.62 [38, p. 217]. Equipped with this result from [38], we return to our own problem. Let $f = -\sum_{l=1}^{L} \mathcal{G}_l$, $x = P_{trn}$, $\boldsymbol{y} = \{P_l, \boldsymbol{w}_l\}_{l=1}^{L}$ in problem (P$_{B'}$). We obtain x^{opt} using Golden section method, where in each iteration Algorithm 1 is applied to obtain $\boldsymbol{y}^{opt}$. Note that in Section IV-B, we proved that f is strictly convex w.r.t. x and thus, the convergence of Algorithm 2 to x^{opt} is guaranteed. Since Algorithm 1 is an iterative algorithm with an unknown convergence rate, the exact convergence rate of Algorithm 2 is unknown. We only know that convergence rate of Algorithm 2 is less than 0.62. Our extensive simulations indicate that the average number of iterations needed for Algorithm 2 to converge to an ϵ-accurate solution for x^{opt} is 15.

Fig. 3: D, D_1, D_2, and D_3 versus P_{tot} (dB).

Fig. 4: g_t, g_c, g_d versus P_{tot} (dB) for the first set of system parameters.

Fig. 5: g_t, g_c, g_d versus P_{tot} (dB) for the second set of system parameters.

VIII. NUMERICAL AND SIMULATION RESULTS

In this section, we corroborate our analytical results with numerical simulations, compare the effectiveness of different proposed power optimization schemes in acheiveing an MSE distorion-power tradeoff which is close to the Bayesian CRB, and investigate how the allocated power across clusters vary as signal-to-noise ratio (SNR) changes.

A. Comparing D and its Lower Bounds

Suppose θ is zero-mean with $\sigma_\theta^2 = 1$ and $L = 10$ clusters. To enforce the heterogeneity in the network, we randomly choose $\sigma_{h_l}, \sigma_{v_l}, \sigma_{n_{l,k}}, \sigma_{q_{l,k}} \in (0,1)$, and $K_l \in \{1, 2, ..., 10\}, l = 1, ..., L, k = 1, ..., K_l$. To capture the effect of randomness in flat fading channel coefficients and communication noise, the numerical results are computed based on 10^6 Monte-Carlo trials, where in each trial, one realization of $|h_l|, \nu_l, \forall l$ are generated. We also assume $\epsilon = 10^{-3}$. In Section III-B we derived three lower bounds on D, of which we optimized D_1, D_2 in problems (P3), (P4), respectively. Fig. 3 plots optimized D, optimized D_1, optimized D_2 versus P_{tot}. Note that $D_3 = 0.0043$ is constant. Clearly, $D_3 < D_2 < D_1 < D < \sigma_\theta^2$. Also, D_2, D_1, D decrease as P_{tot} increases.

B. Comparing Different Power Allocation Schemes

We compare the effectiveness of power optimization schemes, obtained from solving (P1) and its special cases (P1-SC1), (P1-SC2), (P1-SC3), in decreasing the MSE of the LMMSE estimator. We also compare the optimized MSE with the Bayesian CRB G^{-1} derived in Section III-C. Let D_t, D_c, D_d denote the MSE corresponding to the optimal solutions of (P1-SC1), (P1-SC2), (P1-SC3), respectively. We know $D_3 < G^{-1} < D < D_t, D_c, D_d < \sigma_\theta^2$. To quantify the efficacy of different power allocation (w.r.t three sets of optimization variables P_{trn}, P_l's, $\mathcal{P}_l$'s) in closing the MSE performance gap $\sigma_\theta^2 - G^{-1}$, we define three factors as the following:

$$g_t = \frac{D_t - D}{\sigma_\theta^2 - G^{-1}}, \; g_c = \frac{D_c - D}{\sigma_\theta^2 - G^{-1}}, \; g_d = \frac{D_d - D}{\sigma_\theta^2 - G^{-1}}, \quad (56)$$

where $0 \leq g_t, g_c, g_d \leq 1$. A larger factor g means that the particular power allocation is more effective in reducing the MSE performance gap (closing the MSE performance gap). Fig. 4 and Fig. 5 plot g_t, g_c, g_d versus P_{tot} for two sets of noise variances (in Fig. 5 $\sigma_{h_l}, \sigma_{q_{l,k}}$ are chosen from a smaller interval $(0, 0.5)$). For g_t we plot three curves corresponding to $P_{trn} = 5\%, 25\%, 60\% P_{tot}$. Fig. 4 shows $g_c > g_t(P_{trn} = 5\% P_{tot}) > g_t(P_{trn} = 60\% P_{tot}) > g_d > g_t(P_{trn} = 25\% P_{tot})$. Whereas Fig. 5 shows $g_t(P_{trn} = 5\% P_{tot}) > g_c > g_t(P_{trn} = $

$25\% P_{tot}) > g_d > g_t(P_{trn} = 60\% P_{tot})$. Evidently, a more accurate channel estimation does not necessarily lead into a smaller D_t. Two takeaway messages are: (1) $g_t, g_c, g_d > 0$, i.e., the solution obtained from solving (P1) always leads into an MSE improvement, (2) the actual values of g_t, g_c, g_d depend on the system parameters and P_{tot}. Note that in Fig. 4 at $P_{tot} = 0$dB, $g_t = 0.15$ (for $P_{trn} = 25\% P_{tot}$), $g_d = 0.17, g_c = 0.48$, meaning that power allocation among CHs for training and $\mathcal{P}_l$, and among clusters for obtaining P_l reduce the MSE performance gap to $15\%, 17\%, 48\%$, respectively. Combining the information given by g_t, g_c, g_d, G with the computational complexity analysis in Section VI provides the system designer with quantitative complexity-versus-MSE improvement tradeoffs offered by different power optimization schemes.

C. Behavior of Power Allocation Across Clusters

We study the effect of heterogeneous clusters on the behavior of our proposed power allocation scheme to solve (P1) as P_{tot} increases. Consider a network consisting $L = 3$ clusters with $K_l = 6, \sigma_{n_{l,k}} = \sigma_{n_l}, \sigma_{q_{l,k}} = \sigma_{q_l}, \forall l, k$. We define $\gamma_l^o = \frac{\sigma_\theta^2}{\sigma_{n_l}^2}$ as observation SNR of sensors within cluster l, $\gamma_l^c = \frac{1}{\sigma_{q_l}^2}$ as channel-to-noise ratio (CNR) corresponding to sensors-CH$_l$ links, and $\gamma_l^d = \frac{\sigma_{h_l}^2}{\sigma_{v_l}^2}$ as CNR corresponding to CH$_l$-FC link. Let ψ_l (dB) $= 10\log_{10}(\psi_l), P_l$ (dB) $= 10\log_{10}(P_l), \mathcal{P}_l$ (dB) $= 10\log_{10}(\mathcal{P}_l), \mathcal{V}_l$ (dB) $= 10\log_{10}(\mathcal{V}_l)$, where $\mathcal{V}_l = P_l + \mathcal{P}_l$ represents the allocated power to cluster l, excluding its training power ψ_l. In the following we consider three scenarios: (i) when observation SNR γ_l^o and CNR γ_l^c are equal and CNR γ_l^d are different across clusters, (ii) when observation SNR γ_l^o and CNR γ_l^d are equal and CNR γ_l^c are different across clusters, (iii) when CNRs γ_l^c and γ_l^d are equal and observation SNR γ_l^o are different across clusters.

Figs. 6a, 6b, 6c, 6d, respectively, depict ψ_l (dB), $\mathcal{V}_l$ (dB), P_l (dB), $\mathcal{P}_l$ (dB), $\forall l$, versus P_{tot} for $\gamma_l^o = 5$ dB, $\gamma_l^c = 5$ dB, $\forall l$ and $\gamma_1^d = 14$ dB, $\gamma_2^d = 8$ dB, $\gamma_3^d = 2$ dB. Regarding Fig. 6 we

Fig. 6: $\{\gamma_l^o = 5 \text{ dB}, \gamma_l^c = 5 \text{ dB}\}_{l=1}^3$ and $\gamma_1^d > \gamma_2^d > \gamma_3^d$.

Fig. 8: $\{\gamma_l^c = 5 \text{ dB}, \gamma_l^d = 5 \text{ dB}\}_{l=1}^3$ and $\gamma_1^o > \gamma_2^o > \gamma_3^o$.

Fig. 7: $\{\gamma_l^o = 5 \text{ dB}, \gamma_l^d = 5 \text{ dB}\}_{l=1}^3$ and $\gamma_1^c > \gamma_2^c > \gamma_3^c$.

make the following observations: 1) all powers increase as P_{tot} increases, 2) when P_{tot} is small, only cluster 1 is active, and as P_{tot} increases, clusters 2 and 3 become active in a sequential order, 3) in all regions of P_{tot}, a cluster with a larger γ_l^d is allotted a larger ψ_l (water filling), 4) in low-region to moderate-region of P_{tot}, a cluster with a larger γ_l^d is allocated a larger $\mathcal{V}_l$ (water filling), and in high-region of P_{tot}, $\mathcal{V}_l$ of all clusters converge (uniform power allocation), 5) in all regions of P_{tot}, a cluster with a larger γ_l^d is assigned a larger P_l (water filling), 6) in low-region of P_{tot}, a cluster with a larger γ_l^d is allocated a larger $\mathcal{P}_l$ (water filling), and in high-region of P_{tot}, a cluster with a larger γ_l^d is allotted a smaller $\mathcal{P}_l$ (inverse of water filling). The behavior of P_l and $\mathcal{P}_l$ in high-region of P_{tot} can be explained by examining the behavior of $\mathcal{V}_l$. Note that, although CNRs $\gamma_1^d, \gamma_2^d, \gamma_3^d$ are different, the differences are compensated as P_{tot} increases and $\mathcal{V}_l$ of all clusters converge. This fact implies the behaviors of P_l and $\mathcal{P}_l$ in high-region of P_{tot} are opposite, i.e., water filling and inverse of water filling power allocation for P_l and $\mathcal{P}_l$, respectively.

Figs. 7a, 7b, 7c, 7d, respectively, depict ψ_l (dB), $\mathcal{V}_l$ (dB), P_l (dB), $\mathcal{P}_l$ (dB), $\forall l$, versus P_{tot} for $\gamma_l^o = 5$ dB, $\gamma_l^d = 5$ dB, $\forall l$ and $\gamma_1^c = 14$ dB, $\gamma_2^c = 8$ dB, $\gamma_3^c = 2$ dB. The following observations can be made for Fig. 7: comments 1) and 2) for Fig. 6 also hold for Fig. 7, 3) in all regions of P_{tot}, ψ_l of all clusters are equal (uniform power allocation) since γ_l^d's are equal, 4) behavior of $\mathcal{V}_l$ in Fig. 7b is the same as that of Fig. 6b, 5) in low-region of P_{tot}, a cluster with a larger γ_l^c is allocated a larger P_l (water filling), and in high-region of P_{tot}, a cluster

with a larger γ_l^c is allocated a smaller P_l (inverese of water filling), 6) in all regions of P_{tot}, a cluster with a larger γ_l^c is allocated a larger $\mathcal{P}_l$ (water filling). Note that, although CNRs $\gamma_1^c, \gamma_2^c, \gamma_3^c$ are different, the differences are compensated as P_{tot} increases and $\mathcal{V}_l$ of all clusters converge. This fact implies the behaviors of P_l and $\mathcal{P}_l$ in high-region of P_{tot} are opposite, i.e., inverse of water filling and water filling power allocation for P_l and $\mathcal{P}_l$, respectively.

Figs. 8a, 8b, 8c, 8d, respectively, depict ψ_l (dB), $\mathcal{V}_l$ (dB), P_l (dB), $\mathcal{P}_l$ (dB), $\forall l$, versus P_{tot} for $\gamma_l^c = 5$ dB, $\gamma_l^d = 5$ dB, $\forall l$ and $\gamma_1^o = 14$ dB, $\gamma_2^o = 8$ dB, $\gamma_3^o = 2$ dB. The following observations can be made for Fig. 8: comments 1) and 2) for Figs. 6 and 7 also hold for Fig. 8, 3) in all regions of P_{tot}, ψ_l of all clusters are equal (uniform power allocation) since γ_l^d's are equal, 4) in all regions of P_{tot} a cluster with a larger γ_l^o is allocated a larger $\mathcal{V}_l$, a larger P_l, and a larger $\mathcal{P}_l$ (water filling). The behaviors of $\mathcal{V}_l$, P_l, $\mathcal{P}_l$ in high-region of P_{tot} are different from the two previous scenarios (CNRs across clusters were different), in which $\mathcal{V}_l$ of all clusters converge as P_{tot} increases. Here the difference in observation SNR across clusters cannot be compensated as P_{tot} increases. Hence, $\mathcal{V}_l$ of clusters are different, such that a cluster with a larger (smaller) γ_l^o is allocated a larger (smaller) $\mathcal{V}_l$.

One may wonder given our proposed power allocation scheme, how the powers allocated to a CH and a sensor would be different. To answer this question, we let $P_{\text{CH}_l} = \mathcal{P}_l + \psi_l$ denote the sum of power that CH_l consumes for transmitting its fused signal y_l as well as its training symbol to the FC. Fig. 9 plots P_{CH_l} and $P_{l,k}$ versus P_{tot}, using the same setup parameters of Fig. 8. We observe that for all clusters $P_{\text{CH}_l} >> P_{l,k}, k = 1, ..., K_l$, i.e., the power allocated to each sensor is much smaller than the power allocated to each CH.

IX. Conclusions

We studied distributed estimation of a random source in a hierarchical power constrained WSN, where CHs linearly fuse the received signals from sensors within their clusters, and transmit over orthogonal fading channels to the FC. Prior to data transmission, CHs send pilot symbols to the FC to enable channel estimation at the FC. We derived the MSE D corresponding to the LMMSE estimator of the source at the FC, and established lower bounds on D, including the Bayesian CRB. We addressed constrained minimization of D

Fig. 9: $P_{\mathrm{CH}_l} = \mathcal{P}_l + \psi_l$ and $P_{l,k} = \frac{P_l}{K_l}$ versus P_{tot}(dB) when $\{\gamma_l^c = 5 \text{ dB}, \gamma_l^d = 5 \text{ dB}\}_{l=1}^3$ and $\gamma_1^o > \gamma_2^o > \gamma_3^o$.

under the constraint on P_{tot}, where the optimization variables are: i) training power P_{trn} and $\{\psi_l\}_{l=1}^L$, ii) sensor-CH data transmission powers $\{P_l\}_{l=1}^L$, iii) CH-FC data transmission powers $\{\mathcal{P}_l\}_{l=1}^L$. We demonstrated the superior performance of our proposed power allocation scheme, comparing with schemes obtained from solving special case problems where subsets of these variables are optimized. Our simulations revealed that 1) when CNR corresponding to CH$_l$-FC link varies across clusters, ψ_l, P_l allocation follow water filling fashion in all regions of P_{tot}, $\mathcal{P}_l$ follows (inverse of) water filling fashion in (high-region) low-region of P_{tot}, 2) when CNR corresponding to sensors-CH$_l$ links varies across clusters, P_l allocation follows (inverse of) water filling fashion in (high-region) low-region of P_{tot}, $\mathcal{P}_l$ allocation follows water filling fashion in all regions of P_{tot}, 3) when observation SNR varies across clusters, both $P_l, \mathcal{P}_l$ allocation follow water filling fashion in all regions of P_{tot}, and they diverge from uniform power allocation scheme as P_{tot} increases. Leveraging on this work, we discuss three future research directions as follows. First direction is considering a coherent multiple access channel model (instead of orthogonal channels) for intra-cluster communication, where sensors within a cluster transmit their amplified measurements to their CH simultaneously. Second direction is exploring distributed estimation of a random vector source with correlated components. Similar to our work, all sensors can make noisy measurements of a common vector source, or sensors of different clusters can make partial observations of the vector source. Third direction is studying a system where the FC is equipped with multiple antennas (MIMO system model).

APPENDIX

A. Derivation of Bayesian CRB

Using the Bayes' rule $f(\boldsymbol{z}, \hat{\boldsymbol{h}}, \theta) = f(\boldsymbol{z}, \hat{\boldsymbol{h}}|\theta) f(\theta)$, we can decompose G into two terms [33]:

$$G = \underbrace{\mathbb{E}\{-\frac{\partial^2 \ln f(\theta)}{\partial \theta^2}\}}_{=G_1(\theta)} + \underbrace{\mathbb{E}\{-\mathbb{E}\{\frac{\partial^2 \ln f(\boldsymbol{z}, \hat{\boldsymbol{h}}|\theta)}{\partial \theta^2}\}\}}_{=G_2(\theta)}, \quad (57)$$

in which the outer expectations are taken over the pdf of θ, denoted as $f(\theta)$. Note that $\mathbb{E}\{G_1(\theta)\}$ depends on $f(\theta)$ [35]. For instance, if θ is Gaussian with variance σ_θ^2, we obtain $\mathbb{E}\{G_1(\theta)\} = \sigma_\theta^{-2}$. Since $\hat{\boldsymbol{h}}$ and θ are independent, the Bayes' rule says $f(\boldsymbol{z}, \hat{\boldsymbol{h}}|\theta) = f(\boldsymbol{z}|\hat{\boldsymbol{h}}, \theta) f(\hat{\boldsymbol{h}})$, and we can rewrite $G_2(\theta) = -\mathbb{E}\{\mathbb{E}\{\frac{\partial^2 \ln f(\boldsymbol{z}|\hat{\boldsymbol{h}}, \theta)}{\partial \theta^2}|\hat{\boldsymbol{h}}\}\}$, where the outer and inner expectations are taken over the pdfs $f(\hat{\boldsymbol{h}})$ and $f(\boldsymbol{z}|\hat{\boldsymbol{h}}, \theta)$, respectively. We note that $G_2(\theta)$ depends on the

parameters of the observation model at the sensors as well as the physical layer parameters corresponding to sensors-CHs and CHs-FC links. One can show that z_l's conditioned on $\hat{\boldsymbol{h}}, \theta$ are independent, i.e., $f(\boldsymbol{z}|\hat{\boldsymbol{h}}, \theta) = \prod_{l=1}^L f(z_l|\hat{h}_l, \theta)$. Moreover, since channel estimation is performed independently for each cluster, we have $f(\hat{\boldsymbol{h}}) = \prod_{l=1}^L f(\hat{h}_l)$. Hence $G_2(\theta)$ becomes:

$$G_2(\theta) = -\int_{\hat{\boldsymbol{h}}} \int_{\boldsymbol{z}} \{\sum_{l=1}^L [\frac{\partial^2 f(z_l|\hat{h}_l, \theta)}{\partial \theta^2} - \frac{1}{f(z_l|\hat{h}_l, \theta)}(\frac{\partial f(z_l|\hat{h}_l, \theta)}{\partial \theta})^2]$$
$$\times f(\hat{h}_l)\} \prod_{\substack{i=1 \\ i\neq l}}^L f(z_i|\hat{h}_i, \theta) f(\hat{h}_i) d\boldsymbol{z} d\hat{\boldsymbol{h}}.$$

Using the following two facts:

$$\int_{\hat{h}_1} \cdots \int_{\hat{h}_{l-1}} \int_{\hat{h}_{l+1}} \cdots \int_{\hat{h}_L} \int_{z_1} \cdots \int_{z_{l-1}} \int_{z_{l+1}} \cdots \int_{z_L} \prod_{\substack{i=1 \\ i\neq l}}^L f(z_i|\hat{h}_i, \theta) f(\hat{h}_i) \times$$
$$dz_1 \ldots dz_{l-1} dz_{l+1} \ldots dz_L d\hat{h}_1 \ldots d\hat{h}_{l-1} d\hat{h}_{l+1} \ldots d\hat{h}_L = 1,$$

$$\sum_{l=1}^L \int_{z_l} \frac{\partial^2 f(z_l|\hat{h}_l, \theta)}{\partial \theta^2} dz_l = \sum_{l=1}^L \frac{\partial^2}{\partial \theta^2}(\underbrace{\int_{z_l} f(z_l|\hat{h}_l, \theta)}_{=1}) = 0,$$

we find that $G_2(\theta)$ reduces to (27). Examining (27) we realize that we need to find two terms in order to fully characterize $G_2(\theta)$: the conditional pdf $f(z_l|\hat{h}_l, \theta)$, and its first derivative with respect to θ, $\partial f(z_l|\hat{h}_l, \theta)/\partial \theta$. In the following, we derive these two terms. Using (15) we can write the received signal at the FC from CH$_l$ as:

$$z_l = \underbrace{(\hat{h}_l + \tilde{h}_l)}_{=u_{1_l}} \underbrace{\boldsymbol{w}_l^T (\sqrt{\boldsymbol{A}_l}(\theta \boldsymbol{1}_l + \boldsymbol{n}_l) + \boldsymbol{q}_l)}_{=u_{2_l}} + v_l. \quad (58)$$

in which u_{1_l}, u_{2_l}, v_l are mutually independent conditioned on $\hat{h}_l, \theta$. Let $\bar{z}_l = u_{1_l} u_{2_l}$. Hence, $z_l = \bar{z}_l + v_l$. Next, we find the conditional pdf of $\bar{z}_l$, conditioned on $\hat{h}_l, \theta$. Considering (5), we note that h_l, ν_l are zero-mean independent complex Gaussian, and hence from (14) we find that $\hat{h}_l$ is also a zero-mean complex Gaussian. Since $h_l = \hat{h}_l + \tilde{h}_l$, we have $\tilde{h}_l \sim \mathcal{CN}(0, \zeta_l^2)$. Also, $u_{1_l} \sim \mathcal{CN}(\hat{h}_l, \zeta_l^2)$ and $u_{2_l} \sim \mathcal{N}(\bar{\mu}_l, \bar{\sigma}_l^2)$ in (58), where $\bar{\mu}_l = \theta \boldsymbol{w}_l^T \sqrt{\boldsymbol{A}_l} \boldsymbol{1}_l$, $\bar{\sigma}_l^2 = \boldsymbol{w}_l^T (\sqrt{\boldsymbol{A}_l} \boldsymbol{\Sigma}_{n_l} \sqrt{\boldsymbol{A}_l} + \boldsymbol{\Sigma}_{q_l}) \boldsymbol{w}_l$. To find the conditional pdf of $\bar{z}_l$ we use the following lemma from [42].

Lemma 2. If $X \sim \mathcal{CN}(\mu_x e^{j\phi_x}, \sigma_x^2)$ and $Y \sim \mathcal{CN}(\mu_y e^{j\phi_y}, \sigma_y^2)$ are independent complex Gaussian random variables, the pdf of $Z = XY$ (which is equal to the joint pdf of its real and imaginary parts) is:

$$f(Z) = f(z_r, z_i) = \frac{2}{\pi \sigma_x^2 \sigma_y^2} e^{-(k_x^2 + k_y^2)} \tag{59}$$

$$\times \sum_{m=0}^{\infty} \sum_{n=0}^{m} \sum_{p=0}^{m-n} \frac{(2\cos(\angle Z - \phi_x - \phi_y))^{m-n-p}}{m!n!p!(m-n-p)!}$$

$$\times (\frac{|Z|k_x k_y}{\sigma_x \sigma_y})^m (\frac{k_x}{k_y})^{n-p} K_{n-p}(\frac{2|Z|}{\sigma_x \sigma_y}),$$

where $k_x = \mu_x/\sigma_x, k_y = \mu_y/\sigma_y, |Z| = \sqrt{z_r^2 + z_i^2}, \angle Z = \arctan(z_i/z_r)$, and $K_r(x)$ is the modified Bessel function of the second kind with order r and argument x.

Therefore, we can write the conditional joint pdf $f(\bar{z}_{l_r}, \bar{z}_{l_i}|\hat{h}_l, \theta)$ using (59). Recall $v_l \sim \mathcal{CN}(0, 2\sigma_{v_l}^2)$. Hence $f(v_l) = f(v_{l_r}, v_{l_i}) = \frac{1}{(2\pi\sigma_{v_l}^2)} \exp(-\frac{v_{l_r}^2 + v_{l_i}^2}{2\sigma_{v_l}^2})$. Since $\bar{z}_l$ and v_l are independent, the conditional joint pdf $f(z_{l_r}, z_{l_i}|\hat{h}_l, \theta)$ is computed as $f(z_{l_r}, z_{l_i}|\hat{h}_l, \theta) = f(\bar{z}_{l_r}, \bar{z}_{l_i}|\hat{h}_l, \theta) * f(v_{l_r}, v_{l_i})$, in which $*$ is the operator for two-dimensional convolution. Substituting for $f(\bar{z}_{l_r}, \bar{z}_{l_i}|\hat{h}_l, \theta), f(v_{l_r}, v_{l_i})$ from above and defining $b = |b|e^{j\angle b}$, after some mathematical manipulations, we reach $f(z_l|\hat{h}_l, \theta)$ and $\frac{\partial f(z_l|\hat{h}_l, \theta)}{\partial \theta}$ in (28) and (29), respectively, whose parameters are defined in (30). Substituting (28) and (29) in (27), we compute $G_2(\theta)$.

B. Proof of Proposition 1: Finding $\boldsymbol{w}_l^{opt}, \mathcal{F}_l^{opt}$ in terms of P_l

According to (39), the only non-zero eigenvalue of $\boldsymbol{\mathcal{B}}_1$ and its corresponding eigenvector are:

$$\mathcal{F}_l^{opt} = |\boldsymbol{\mu}_l|^T \boldsymbol{\mathcal{B}}_1^{-1} |\boldsymbol{\mu}_l|, \quad \boldsymbol{s}_l^{opt} = \boldsymbol{\mathcal{B}}_1^{-1} |\boldsymbol{\mu}_l|. \tag{60}$$

Define $\delta_l = \mathcal{V}_l - P_l, \xi_l = \frac{\sigma_\theta^2}{|\hat{h}_l|^2}(\frac{\sigma_{v_l}^2}{\delta_l} + \zeta_l^2), \boldsymbol{\Sigma}_{\mu_l} = |\boldsymbol{\mu}_l||\boldsymbol{\mu}_l|^T, \phi_l = |\hat{h}_l|^2 + \zeta_l^2 + \frac{\sigma_{v_l}^2}{\delta_l}, \boldsymbol{\Sigma}_{P_l} = \boldsymbol{\Sigma}_{q_l} + P_l \boldsymbol{\Delta}_l, \boldsymbol{\Sigma}_{\phi_l} = \phi_l \boldsymbol{\Sigma}_{P_l}$. By substituting $\boldsymbol{\Delta}_l, \boldsymbol{\Pi}_l$ into $\boldsymbol{\Omega}_l$ and $\boldsymbol{B}_l$, and $\boldsymbol{\Omega}_l$ into $\boldsymbol{R}_{t_l}, \boldsymbol{\mathcal{B}}_1$ in (38) becomes $\boldsymbol{\mathcal{B}}_1 = \boldsymbol{\Sigma}_{\phi_l} + \xi_l \boldsymbol{\Sigma}_{\mu_l}$. Using the Binomial inversion Lemma [32] we compute $\boldsymbol{s}_l^{opt}$ in (60):

$$\boldsymbol{s}_l^{opt} = \frac{\boldsymbol{\Sigma}_{\phi_l}^{-1}|\boldsymbol{\mu}_l|}{1 + \xi_l |\boldsymbol{\mu}_l|^T \boldsymbol{\Sigma}_{\phi_l}^{-1}|\boldsymbol{\mu}_l|}. \tag{61}$$

From (61), we obtain $\boldsymbol{w}_l^{opt}$:

$$\boldsymbol{w}_l^{opt} = \sqrt{\frac{\delta_l}{|\boldsymbol{\mu}_l|^T \boldsymbol{\Sigma}_{\phi_l}^{-1} \boldsymbol{R}_{t_l} \boldsymbol{\Sigma}_{\phi_l}^{-1}|\boldsymbol{\mu}_l|}} \boldsymbol{\Sigma}_{\phi_l}^{-1}|\boldsymbol{\mu}_l| \tag{62}$$

$$\stackrel{(a)}{=} \sqrt{\frac{\delta_l}{\boldsymbol{\rho}_l^T \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\rho}_l (1 + \sigma_\theta^2 P_l \boldsymbol{\rho}_l^T \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\rho}_l)}} \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\rho}_l \stackrel{(b)}{=} \sqrt{\frac{\delta_l}{\tau_l}} \boldsymbol{R}_{t_l}^{-1} \boldsymbol{\rho}_l,$$

where τ_l is defined in Proposition 1. To obtain (a) in (62), we use the fact that $|\boldsymbol{\mu}_l|^T \boldsymbol{\Sigma}_{\phi_l}^{-1} \boldsymbol{R}_{t_l} \boldsymbol{\Sigma}_{\phi_l}^{-1}|\boldsymbol{\mu}_l| = \frac{\epsilon_l}{\phi_l^2}(1 + \frac{\sigma_\theta^2}{|\hat{h}_l|^2}\epsilon_l)$, where $\epsilon_l = |\boldsymbol{\mu}_l|^T \boldsymbol{\Sigma}_{P_l}^{-1}|\boldsymbol{\mu}_l|$. To obtain (b) in (62), we use $\boldsymbol{R}_{t_l}^{-1} \boldsymbol{\rho}_l = \frac{\boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\rho}_l}{1 + \sigma_\theta^2 P_l \boldsymbol{\rho}_l^T \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\rho}_l}$, which is established using the Binomial inversion lemma. We have $\mathcal{F}_l^{opt} = |\boldsymbol{\mu}_l|^T \boldsymbol{s}_l^{opt}$. Substituting $\boldsymbol{s}_l^{opt}$ from (61) in (60) and using the fact that

$1 - \sigma_\theta^2 P_l \tau_l = \frac{1}{1 + \sigma_\theta^2 P_l \boldsymbol{\rho}_l^T \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\rho}_l}$ we reach:

$$\mathcal{F}_l^{opt} = \frac{|\boldsymbol{\mu}_l|^T \boldsymbol{\Sigma}_{\phi_l}^{-1}|\boldsymbol{\mu}_l|}{1 + \xi_l |\boldsymbol{\mu}_l|^T \boldsymbol{\Sigma}_{\phi_l}^{-1}|\boldsymbol{\mu}_l|} \tag{63}$$

$$= \frac{|\hat{h}_l|^2 P_l \tau_l}{|\hat{h}_l|^2 (1 - \sigma_\theta^2 P_l \tau_l) + \zeta_l^2 + \frac{\sigma_{v_l}^2}{\delta_l}} = \frac{|\hat{h}_l|^2 \beta_l P_l \tau_l}{\sigma_{v_l}^2 (1 + \frac{\beta_l}{\delta_l})},$$

C. Solution of the Problem in (49)

Define $\delta_l = \mathcal{V}_l - P_l$ and let T denote the objective function in (49). We have $\frac{\partial T}{\partial \delta_l} = \frac{|\hat{h}_l|^2 \beta_l^2 P_l \tau_l}{\sigma_{v_l}^2 (\beta_l + \delta_l)^2} > 0$, implying that the solution to (49) must satisfy the equality constraint $\sum_{l=1}^{L} \delta_l + P_l = \sigma P_{tot}$. Also, $\frac{\partial^2 T}{\partial \delta_i \partial \delta_j} = 0, \forall i \neq j$, and $\frac{\partial^2 T}{\partial \delta_l^2} = \frac{-2|\hat{h}_l|^2 \beta_l^2 P_l \tau_l}{\sigma_{v_l}^2 (\beta_l + \delta_l)^3} < 0, \forall l$. Thus the Hessian of T with respect to δ_l's is diagonal and negative definite, proving that T is jointly concave over δ_l's. Since the constraint is linear in δ_l, the problem in (49) is concave. The Lagrangian function $\mathcal{L}$ associated with (49) is:

$$\mathcal{L}(\lambda, \{\eta_l, \delta_l\}_{l=1}^L) = \sum_{l=1}^{L} \frac{|\hat{h}_l|^2 \beta_l P_l \tau_l}{\sigma_{v_l}^2 (1 + \frac{\beta_l}{\delta_l})} - \delta_l (\lambda - \eta_l) + \lambda(\sigma P_{tot} - \sum_{l=1}^{L} P_l),$$

where λ, η_l's are the Lagrange multipliers. The KKT optimality conditions are:

$$\frac{|\hat{h}_l|^2 \beta_l^2 P_l \tau_l}{\sigma_{v_l}^2 (\beta_l + \delta_l)^2} - \lambda + \eta_l = 0, \; \forall l, \tag{64a}$$

$$\lambda \left(\sum_{l=1}^{L} \delta_l + P_l - \sigma P_{tot} \right) = 0, \; \lambda \geq 0, \tag{64b}$$

$$\eta_l \delta_l = 0, \; \eta_l \geq 0, \; \delta_l \geq 0, \; \forall l. \tag{64c}$$

The condition (64c) implies $\eta_l = 0$ for active clusters with $\delta_l > 0$. From (64a) we infer:

$$\delta_l^{opt} = \left[\beta_l(\frac{|\hat{h}_l|}{\sigma_{v_l}} \sqrt{\frac{P_l \tau_l}{\lambda}} - 1) \right]^+, \tag{65}$$

in which $[x]^+ = max\{x, 0\}$. Having δ_l^{opt}, we find $\mathcal{V}_l^{opt} = \delta_l^{opt} + P_l$ given in (50a). Substituting (65) in the active constraint condition $\sum_{l=1}^{L} \delta_l + P_l = \sigma P_{tot}$, the Lagrange multiplier λ becomes equal to the expression given in (50b), in which $\mathcal{A}$ is the set of active clusters. To uniquely determine $\mathcal{A}$, we carry out the following procedure. Let $L_\mathcal{A} = |\mathcal{A}|$ where $L_\mathcal{A} \leq L$. Suppose the clusters are indexed in the descending order of $\frac{|\hat{h}_1|^2 P_1 \tau_1}{\sigma_{v_1}^2} \geq \frac{|\hat{h}_2|^2 P_2 \tau_2}{\sigma_{v_2}^2} \geq ... \geq \frac{|\hat{h}_L|^2 P_L \tau_L}{\sigma_{v_L}^2}$. Choosing an $L_\mathcal{A}$ value we find λ and compute $\delta_l^{opt} = \beta_l(\frac{|\hat{h}_l|}{\sigma_{v_l}} \sqrt{\frac{P_l \tau_l}{\lambda}} - 1), \forall l$. If $\delta_l^{opt} > 0, \; l = 1, ..., L_\mathcal{A}$ and $\delta_l^{opt} \leq 0, \; l = L_\mathcal{A}+1, ..., L$, then we have identified the set of active clusters $\mathcal{A}$ with their corresponding $P_l, l \in \mathcal{A}$. Otherwise, we repeat this process for another $L_\mathcal{A}$ value. It is proved that the solution always exists and is unique [43].

D. Proof of Concavity of sub-problem (a) of (P1-SC3) over P_l's

We rewrite the cost function of sub-problem (a), denoted as $\mathcal{F}$, as:

$$\mathcal{F} = \frac{1}{\sigma_\theta^2} \sum_{l=1}^{L} \overbrace{\frac{1}{b_l}(1 - \frac{s_l}{s_l + P_l m_l})}^{\mathcal{F}_l}, \qquad (66)$$

where $b_l = \frac{1}{|\hat{h}_l|^2}(\frac{\sigma_{v_l}^2}{\mathcal{P}} + \zeta_l^2), s_l = \frac{1+b_l}{\sigma_\theta^2 b_l}, m_l = \boldsymbol{\rho}_l^T \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\rho}_l, \boldsymbol{\Sigma}_{P_l} = \boldsymbol{\Sigma}_{q_l} + P_l \boldsymbol{\Delta}_l$. We have $b_l, s_l, m_l > 0$ and $\boldsymbol{\Sigma}_{P_l} \succ \mathbf{0}$. Also, $\frac{\partial m_l}{\partial P_l} = -\boldsymbol{\rho}_l^T \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\Delta}_l \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\rho}_l < 0, \frac{\partial^2 m_l}{\partial P_l^2} = 2\boldsymbol{\rho}_l^T \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\Delta}_l \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\Delta}_l \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\rho}_l > 0$. One can obtain $\frac{\partial \mathcal{F}_l}{\partial P_l} = \frac{s_l(m_l + P_l \frac{\partial m_l}{\partial P_l})}{b_l(s_l + P_l m_l)^2}$ and prove that $m_l + P_l \frac{\partial m_l}{\partial P_l} > 0$ which infers $\frac{\partial \mathcal{F}_l}{\partial P_l} > 0$, i.e.,

$$\boldsymbol{\Sigma}_{q_l} \succ \mathbf{0} \Rightarrow \boldsymbol{\Sigma}_{P_l} \succ P_l \boldsymbol{\Delta}_l \Rightarrow \boldsymbol{\Sigma}_{P_l}^{-1} \succ P_l \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\Delta}_l \boldsymbol{\Sigma}_{P_l}^{-1} \Rightarrow$$

$$\boldsymbol{\rho}_l^T \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\rho}_l - P_l \boldsymbol{\rho}_l^T \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\Delta}_l \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\rho}_l > 0 \Rightarrow m_l + P_l \frac{\partial m_l}{\partial P_l} > 0.$$

$\mathcal{F}$ in (66) is an increasing function of P_l, and thus, the solution of sub-problem (a) of (P1-SC3) must satisfy the equality constraint $P_{trn} + \sum_{l=1}^{L}\{P_l + \mathcal{P}\} = P_{tot}$. Furthermore

$$\frac{\partial^2 \mathcal{F}_l}{\partial P_l^2} = \frac{s_l[(s_l + P_l m_l)(2\frac{\partial m_l}{\partial P_l} + P_l \frac{\partial^2 m_l}{\partial P_l^2}) - 2(m_l + P_l \frac{\partial m_l}{\partial P_l})^2]}{b_l(s_l + P_l m_l)^3}.$$

The denominator of the right-hand side is positive. The numerator of the right-hand side can be simplified as num $= I_1 + I_2 + I_3$, where

$$I_1 = s_l \boldsymbol{\rho}_l^T (P_l \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\Delta}_l \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\Delta}_l \boldsymbol{\Sigma}_{P_l}^{-1} - \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\Delta}_l \boldsymbol{\Sigma}_{P_l}^{-1}) \boldsymbol{\rho}_l,$$

$$I_2 = \boldsymbol{\rho}_l^T (P_l \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\Pi}_l \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\Delta}_l \boldsymbol{\Sigma}_{P_l}^{-1} - \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\Pi}_l \boldsymbol{\Sigma}_{P_l}^{-1}) \boldsymbol{\rho}_l,$$

$$I_3 = P_l^2 \boldsymbol{\rho}_l^T (\boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\Pi}_l \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\Delta}_l \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\Delta}_l \boldsymbol{\Sigma}_{P_l}^{-1}$$
$$- \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\Delta}_l \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\Pi}_l \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\Delta}_l \boldsymbol{\Sigma}_{P_l}^{-1}) \boldsymbol{\rho}_l.$$

One can prove that $I_1 < 0, I_2 < 0, I_3 = 0$. Hence, num < 0 and $\frac{\partial^2 \mathcal{F}_l}{\partial P_l^2} < 0$. The following sequences of inequalities are easy to verify:

$$\boldsymbol{\Sigma}_{q_l} \succ \mathbf{0} \Rightarrow \boldsymbol{\Sigma}_{P_l} \succ P_l \boldsymbol{\Delta}_l \Rightarrow \boldsymbol{I} \succ P_l \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\Delta}_l \Rightarrow$$

$$\boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\Delta}_l \boldsymbol{\Sigma}_{P_l}^{-1} \succ P_l \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\Delta}_l P_l \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\Delta}_l \boldsymbol{\Sigma}_{P_l}^{-1} \Rightarrow \boxed{I_1 < 0},$$

$$\boldsymbol{I} \succ P_l \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\Delta}_l \Rightarrow (\boldsymbol{\rho}_l^T \boldsymbol{\rho}_l)^2 > P_l(\boldsymbol{\rho}_l^T \boldsymbol{\rho}_l)(\boldsymbol{\rho}_l^T \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\Delta}_l \boldsymbol{\rho}_l) \Rightarrow$$

$$\boldsymbol{\Pi}_l \succ P_l \boldsymbol{\Pi}_l \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\Delta}_l \Rightarrow \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\Pi}_l \boldsymbol{\Sigma}_{P_l}^{-1} \succ P_l \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\Pi}_l \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\Delta}_l \boldsymbol{\Sigma}_{P_l}^{-1}$$

$$\Rightarrow \boxed{I_2 < 0},$$

$$\boldsymbol{\rho}_l^T \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\Delta}_l \boldsymbol{\rho}_l \overset{(a)}{=} \boldsymbol{\rho}_l^T \boldsymbol{\Delta}_l \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\rho}_l \Rightarrow (\boldsymbol{\rho}_l^T \boldsymbol{\rho}_l) \boldsymbol{\rho}_l^T \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\Delta}_l \boldsymbol{\rho}_l =$$

$$\boldsymbol{\rho}_l^T \boldsymbol{\Delta}_l \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\rho}_l (\boldsymbol{\rho}_l^T \boldsymbol{\rho}_l) \Rightarrow \boldsymbol{\Pi}_l \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\Delta}_l = \boldsymbol{\Delta}_l \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\Pi}_l \Rightarrow$$

$$\boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\Pi}_l \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\Delta}_l \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\Delta}_l \boldsymbol{\Sigma}_{P_l}^{-1} = \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\Delta}_l \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\Pi}_l \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\Delta}_l \boldsymbol{\Sigma}_{P_l}^{-1}$$

$$\Rightarrow \boxed{I_3 = 0},$$

where (a) comes by the fact that $\boldsymbol{\rho}_l^T \boldsymbol{\Sigma}_{P_l}^{-1} \boldsymbol{\Delta}_l \boldsymbol{\rho}_l$ is scalar. The Hessian of $\mathcal{F}$ with respect to P_l's is diagonal and negative definite, which proves that $\mathcal{F}$ is jointly concave over P_l's. Moreover, the constraint is linear in P_l, and therefore finding P_l's in sub-problem (a) of (P1-SC3) is jointly concave over P_l's and has a unique solution.

REFERENCES

[1] M. Shirazi and A. Vosoughi, "On bayesian fisher information maximization for distributed vector estimation," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 5, no. 4, pp. 628–645, Dec 2019.

[2] M. Shirazi and A. Vosoughi, "Bayesian Cramer-Rao bound for distributed vector estimation with linear observation model," in *IEEE International Symposium on Personal, Indoor, and Mobile Radio Communication*, 2014.

[3] ——, "Bayesian Cramer-Rao bound for distributed estimation of correlated data with non-linear observation model," in *Asilomar Conference on Signals, Systems and Computers*, 2014.

[4] A. Sani and A. Vosoughi, "Distributed vector estimation for power- and bandwidth-constrained wireless sensor networks," *IEEE Transactions on Signal Processing*, vol. 64, no. 15, pp. 3879–3894, Aug 2016.

[5] A. Sani and A. Vosoughi, "On distributed linear estimation with observation model uncertainties," *IEEE Transactions on Signal Processing*, vol. 66, no. 12, pp. 3212–3227, June 2018.

[6] J. Fang and H. Li, "Power constrained distributed estimation with cluster-based sensor collaboration," *IEEE Transactions on Wireless Communications*, vol. 8, no. 7, pp. 3822–3832, July 2009.

[7] C. A. Lin and C. H. Wu, "Linear coherent distributed estimation with cluster-based sensor networks," *IET Signal Processing*, vol. 6, no. 7, pp. 626–632, Sep. 2012.

[8] M. H. Chaudhary and L. Vandendorpe, "Performance of power-constrained estimation in hierarchical wireless sensor networks," *IEEE Transactions on Signal Processing*, vol. 61, no. 3, pp. 724–739, 2013.

[9] S. A. Aldalahmeh, S. O. Al-Jazzar, D. McLernon, S. A. R. Zaidi, and M. Ghogho, "Fusion rules for distributed detection in clustered wireless sensor networks with imperfect channels," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 5, no. 3, pp. 585–597, Sep. 2019.

[10] M. Gastpar, B. Rimoldi, and M. Vetterli, "To code, or not to code: lossy source-channel communication revisited," *IEEE Transactions on Information Theory*, vol. 49, no. 5, pp. 1147–1158, May 2003.

[11] S. Cui, J. Xiao, A. J. Goldsmith, Z. Luo, and H. V. Poor, "Estimation diversity and energy efficiency in distributed sensing," *IEEE Transactions on Signal Processing*, vol. 55, no. 9, pp. 4683–4695, Sep. 2007.

[12] P. Salvo Rossi, D. Ciuonzo, K. Kansanen, and T. Ekman, "Performance analysis of energy detection for mimo decision fusion in wireless sensor networks over arbitrary fading channels," *IEEE Transactions on Wireless Communications*, vol. 15, no. 11, pp. 7794–7806, Nov 2016.

[13] F. Jiang, J. Chen, A. L. Swindlehurst, and J. A. Lpez-Salcedo, "Massive mimo for wireless sensing with a coherent multiple access channel," *IEEE Transactions on Signal Processing*, vol. 63, no. 12, pp. 3005–3017, June 2015.

[14] H. R. Ahmadi and A. Vosoughi, "Optimal training and data power allocation in distributed detection with inhomogeneous sensors," *IEEE Signal Processing Letters*, vol. 20, no. 4, pp. 339–342, April 2013.

[15] ——, "Impact of wireless channel uncertainty upon distributed detection systems," *IEEE Transactions on Wireless Communications*, vol. 12, no. 6, pp. 2566–2577, June 2013.

[16] H. R. Ahmadi, N. Maleki, and A. Vosoughi, "On power allocation for distributed detection with correlated observations and linear fusion," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 9, pp. 8396–8410, Sep. 2018.

[17] M. Sedghi, G. Atia, and M. Georgiopoulos, "Low-dimensional decomposition of manifolds in presence of outliers," in *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, Oct 2019, pp. 1–6.

[18] ——, "Robust manifold learning via conformity pursuit," *IEEE Signal Processing Letters*, vol. 26, no. 3, pp. 425–429, March 2019.

[19] M. Hosseini, A. S. Maida, M. Hosseini, and G. Raju, "Inception-inspired lstm for next-frame video prediction," 2019.

[20] Y. Zhang, N. Meratnia, and P. Havinga, "Outlier detection techniques for wireless sensor networks: A survey," *IEEE Communications Surveys Tutorials*, vol. 12, no. 2, pp. 159–170, Second 2010.

[21] M. Hosseini, M. A. Salehi, and R. Gottumukkala, "Enabling interactive video streaming for public safety monitoring through batch scheduling," in *2017 IEEE 19th International Conference on High Performance Computing and Communications; IEEE 15th International Conference on Smart City; IEEE 3rd International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, 2017, pp. 474–481.

[22] M. K. Banavar, C. Tepedelenlioglu, and A. Spanias, "Estimation over fading channels with limited feedback using distributed sensing," *IEEE Transactions on Signal Processing*, vol. 58, no. 1, pp. 414–425, 2010.

[23] C.-H. Wang and S. Dey, "Distortion outage minimization in nakagami fading using limited feedback," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, no. 1, pp. 92–107, Oct 2011.

[24] H. Senol and C. Tepedelenlioglu, "Performance of distributed estimation over unknown parallel fading channels," *IEEE Transactions on Signal Processing*, vol. 56, no. 12, pp. 6057–6068, Dec 2008.

[25] C.-H. Wu and C.-A. Lin, "Linear coherent distributed estimation over unknown channels," *Signal Proc.*, vol. 91, no. 4, pp. 1000 – 1011, 2011.

[26] N. A. Pantazis, S. A. Nikolidakis, and D. D. Vergados, "Energy-efficient routing protocols in wireless sensor networks: A survey," *IEEE Communications Surveys Tutorials*, vol. 15, no. 2, pp. 551–591, Second 2013.

[27] J. G. Proakis and M. Salehi, *Digital Communications*, 5th ed. McGraw-Hill, New York, 2007, pp. 63–64.

[28] J. H. Kotecha, V. Ramachandran, and A. M. Sayeed, "Distributed multitarget classification in wireless sensor networks," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 4, pp. 703–713, April 2005.

[29] C.-H. Wang and S. Dey, "Distortion outage minimization in nakagami fading using limited feedback," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, no. 1, pp. 92–107, Oct 2011.

[30] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice-Hall PTR, 1993, p. 382.

[31] Y. Jia and A. Vosoughi, "Transmission resource allocation for training based amplify-and-forward relay systems," *IEEE Transactions on Wireless Communications*, vol. 10, no. 2, pp. 450–455, February 2011.

[32] C. D. Meyer, *Matrix Analysis and Applied Linear Algebra*. Society for Industrial and Applied Mathematics (SIAM), 2001, p. 124.

[33] H. L. Van Trees and K. L. Bell, *Bayesian Bounds for Parameter Estimation and Nonlinear Filtering/Tracking*. Wiley, 2007, p. 5.

[34] A. Vosoughi and A. Scaglione, "Everything you always wanted to know about training: guidelines derived using the affine precoding framework and the CRB," *IEEE Transactions on Signal Proc.*, vol. 54, no. 3, pp. 940–954, March 2006.

[35] ——, "On the effect of receiver estimation error upon channel mutual information," *IEEE Transactions on Signal Processing*, vol. 54, no. 2, pp. 459–472, Feb 2006.

[36] D. P. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 8, pp. 1439–1451, Aug 2006.

[37] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[38] D. G. Luenberger and Y. Ye, *Linear and Nonlinear Programming*, 4th ed. International Series in Operations, Research and Management Science, Springer, 2015.

[39] A. M. Davie and A. J. Stothers, "Improved bound for complexity of matrix multiplication," *Proceedings of the Royal Society of Edinburgh, Section: A Mathematics*, vol. 143, pp. 351369, April 2013.

[40] L. Grippof and M. Sciandrone, "Globally convergent block-coordinate techniques for unconstrained optimization," *Optimization Methods and Software*, vol. 10, no. 4, pp. 587–637, 1999. [Online]. Available: https://doi.org/10.1080/10556789908805730

[41] L. Grippo and M. Sciandrone, "On the convergence of the block nonlinear gaussseidel method under convex constraints," *Operations Research Letters*, vol. 26, no. 3, pp. 127 – 136, 2000. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167637799000747

[42] N. O'Donoughue and J. M. F. Moura, "On the product of independent complex Gaussians," *IEEE Transactions on Signal Processing*, vol. 60, no. 3, pp. 1050–1063, March 2012.

[43] J.-J. Xiao, S. Cui, Z.-Q. Luo, and A. Goldsmith, "Power scheduling of universal decentralized estimation in sensor networks," *IEEE Transactions on Signal Processing*, vol. 54, no. 2, pp. 413–422, Feb 2006.