AirComp Based Multi-View Sensing for Edge Inference in OFDM System

X

Abstract

I. System Model

A. Network Model

We consider a system with one edge server equipped with R receive antennas and K edge devices, each of which has only one transmit antenna. The edge server aims at aggregating features sensed by all the devices and performing the inference task. Due to different circumstances of the devices (e.g., placement location, observing direction, etc.), each of them can only sense a corrupted version of the ground-true data. Without loss of generality, we model the sensed data of device k as

$$\mathbf{x}_k = \mathbf{x} + \mathbf{d}_k,\ \ k \in \{1, ..., K\}, \tag{1}$$

where $\mathbf{x}$ is a M-dimensional feature vector and $\mathbf{d}_k$ is the sensing distortion following the Gaussian distribution with mean zero and covariance $\mathbf{D}_k$, i.e.,

$$\mathbf{d}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{D}_k), \tag{2}$$

where the covariance matrix $\mathbf{D}_k$ is diagonal, given as

$$\mathbf{D}_k = \mathrm{diag}\{\delta_{k,1}^2, ..., \delta_{k,M}^2\}. \tag{3}$$

This system works in a time division manner and the time duration of transmission is divided into multiple time slots. Moreover, in each time slot, the technique of *over-the-air computation* (AirComp) is adopted to aggregate $\mathbf{x}$. We assume the $\mathbf{x}_k$ defined in (1) is transmitted in a single time slot. If constrained by communication resources, the entire sensed data cannot be transmitted at the same time, these data will be divided into packets and transmitted in multiple time slots.

B. Feature Distribution

The ground-true data is assumed to follow the distribution of a *mixture of Gaussians* with L classes. To reduce communication overhead, the method of *principal component analysis* (PCA) is then applied to project the ground-true data into the eigenspace. In terms of notation, the feature vector $\mathbf{x}$ defined in (1), has already been processed by PCA and projected into the eigenspace. Hence, different feature elements in $\mathbf{x}$ are independent. The distribution of $\mathbf{x}$ is

$$f(\mathbf{x}) = \frac{1}{L}\sum_{\ell=1}^{L}\mathcal{N}(\boldsymbol{\mu}_\ell, \boldsymbol{\Sigma}), \tag{4}$$

where $\mathcal{N}(\boldsymbol{\mu}_\ell, \boldsymbol{\Sigma})$ denotes the Gaussian probability density function of class ℓ with mean $\boldsymbol{\mu}_\ell$ and covariance $\boldsymbol{\Sigma}$. We assume the class label ℓ of $\mathbf{x}$ is uniformly distributed. $\boldsymbol{\mu}_\ell$ is given as

$$\boldsymbol{\mu}_\ell = [\mu_\ell(1), ..., \mu_\ell(M)]^T,\ \ \ell \in \{1, ..., L\}. \tag{5}$$

The covariance matrix $\boldsymbol{\Sigma}$ is diagonalized by PCA, given by

$$\boldsymbol{\Sigma} = \mathrm{diag}\{\sigma_1^2, ..., \sigma_M^2\}. \tag{6}$$

In this paper, we will not focus on the training phase. Yet it is worth noting that the PCA is first performed over the training dataset at the edge server during the training phase, so the mean vector $\boldsymbol{\mu}_\ell$ and the covariance matrix $\boldsymbol{\Sigma}$ are known to the edge server. At the beginning of the inference phase, the computed principal components are broadcasted to all the devices. Each device projects the observed ground-true data into the eigenspace and only the features processed by PCA will be transmitted.

C. Discriminant Gain

The *discriminant gain*, which is proposed by adopting the well known metric of *symmetric Kullback-Leibler* (KL) *divergence*, is used to measure the inference capability of the vector $\mathbf{x}$ aggregated by all the local observation. Based on the settings in [1], in particular, we consider a pairwise discriminant gain of an arbitrary class pair, say class ℓ and ℓ'. The pairwise discriminant gain measures how much the inference model can distinguish class ℓ from class ℓ' if $\mathbf{x}$ is known.

As the feature vector follows the distribution of mixture of Gaussians given in (4), the pairwise discriminant gain is defined as

$$\begin{aligned} G_{\ell,\ell'}(\mathbf{x}) &= \mathsf{KL}[\mathcal{N}(\boldsymbol{\mu}_\ell, \boldsymbol{\Sigma})||\mathcal{N}(\boldsymbol{\mu}_{\ell'}, \boldsymbol{\Sigma})] + \mathsf{KL}[\mathcal{N}(\boldsymbol{\mu}_{\ell'}, \boldsymbol{\Sigma})||\mathcal{N}(\boldsymbol{\mu}_\ell, \boldsymbol{\Sigma})] \\ &= (\boldsymbol{\mu}_\ell - \boldsymbol{\mu}_{\ell'})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_\ell - \boldsymbol{\mu}_{\ell'}) \\ &= \sum_{i=1}^{2N} G_{\ell,\ell'}(x(i)), \end{aligned} \tag{7}$$

where $x(i)$ is the i-th entry of $\mathbf{x}$, and $G_{\ell,\ell'}(x(i))$ is given as

$$G_{\ell,\ell'}(x(i)) = \frac{(\mu_\ell(i) - \mu_{\ell'}(i))^2}{\sigma_i^2}, \qquad i \in \{1, ..., 2N\}. \tag{8}$$

The overall discriminant gain is defined as the average of all the pairwise discriminant gain, given as

$$\begin{aligned} G(\mathbf{x}) &= \frac{2}{L(L-1)} \sum_{\ell'=1}^{L} \sum_{\ell<\ell'} G_{\ell,\ell'}(\mathbf{x}) \\ &= \frac{2}{L(L-1)} \sum_{\ell'=1}^{L} \sum_{\ell<\ell'} \sum_{i=1}^{2N} G_{\ell,\ell'}(x(i)) \\ &= \sum_{i=1}^{2N} G(x(i)), \end{aligned} \tag{9}$$

where $G(x(i))$ is the discriminant gain of the i-th entry of $\mathbf{x}$, given as

$$\begin{aligned} G(x(i)) &= \frac{2}{L(L-1)} \sum_{\ell'=1}^{L} \sum_{\ell<\ell'} G_{\ell,\ell'}(x(i)) \\ &= \frac{2}{L(L-1)} \sum_{\ell'=1}^{L} \sum_{\ell<\ell'} \frac{(\mu_\ell(i) - \mu_{\ell'}(i))^2}{\sigma_i^2}, \end{aligned} \qquad i \in \{1, ..., 2N\}. \tag{10}$$

$G(x(i))$ measures the importance level of feature dimension i for the inference task. The overall $G(\mathbf{x})$, consisting of each $G(x(i))$, is the measurement of the inference capability of vector $\mathbf{x}$.

D. Over-the-air Computation

We use the AirComp technique to aggregate the features observed by the edge devices. In each time slot, suppose device k obtains M principal features processed by PCA and each feature is a real number. Though the device has only one transmit antenna, the OFDM modulation is adopted to transmit multiple features simultaneously. Assume that the whole bandwidth is divided into N orthogonal sub-channels. Therefore, with N sub-carriers, an OFDM symbol consisting of N complex scalars can be transmitted in a single time slot. The real and imaginary parts of a complex scalar each contain one feature element, thus, $2N$ features are carried by one OFDM symbol. We further assume that $M = 2N$. Then, as mentioned above, the obtained M-dimensional feature vector $\mathbf{x}_k$ will be transmitted in one OFDM symbol in a time slot. Since the AirComp design among different time slots are the same, we only elaborate the transmission of an arbitrary time slot here.

Each sub-carrier is assigned to transmit a complex scalar, which contains two features of $\mathbf{x}_k$, as its real and imaginary part. Based on the importance level of different feature dimensions (measured by discriminant gain) and different channel gains in N sub-channels, we design an allocation scheme to allocate $2N$ features of $\mathbf{x}_k$ into a N-dimensional complex vector, which is denoted as $\mathbf{s}_k$. The allocation of features is represented as a simple matrix multiplication, i.e.,

$$\mathbf{s}_k = \mathbf{P}\mathbf{A}\mathbf{x}_k, \ k \in \{1, ..., K\}, \tag{11}$$

where $\mathbf{P} \in \mathbb{C}^{N \times 2N}$ is given by

$$\mathbf{P} = \begin{pmatrix} 1 & j & & & & & \\ & & 1 & j & & & \\ & & & & \ddots & \ddots & \\ & & & & & & 1 & j \end{pmatrix}, \tag{12}$$

and j represents the imaginary unit. $\mathbf{A} \in \mathbb{R}^{2N \times 2N}$ is the permutation matrix. Let $a_{r,c}$ denote the entry of $\mathbf{A}$ in row r and column c. We require that $\mathbf{A}$ is subjected to the following constraints

$$\text{For } c = 1, ..., 2N, \quad \sum_{r=1}^{2N} a_{r,c} = 1, \tag{13}$$

$$\text{For } r = 1, ..., 2N, \quad \sum_{c=1}^{2N} a_{r,c} = 1, \tag{14}$$

and

$$\text{For } r = 1, ..., 2N, \ c = 1, ..., 2N, \quad a_{r,c} \in \{0, 1\}. \tag{15}$$

Constraints (13-15) guarantee that each line of $\mathbf{A}$ (ether a row or a column) contains exactly a entry 1, and the remaining entries of the line equal to 0. Every eligible $\mathbf{A}$ corresponds to a permutation of the features in $\mathbf{x}_k$. The matrix $\mathbf{P}$ is a constant. Multiplied by $\mathbf{P}$, the $(2i-1)$-th and $(2i)$-th features after permutation are combined into the i-th entry of $\mathbf{s}_k$ as its real and imaginary part, respectively. Therefore, every two features are carried by a sub-carrier. The permutation matrix $\mathbf{A}$ will be considered as one of the optimization variables. Depending on different sub-channel conditions and importance level of features, matrix $\mathbf{A}$ is optimized so more important features are more likely to be transmitted in low-noise sub-channels with smooth channel fading. At the beginning of transmission, the allocation matrix $\mathbf{A}$ is computed by the edge server and broadcasted to all the devices.

We use parentheses to denote the index of vector elements, i.e., $s_k(i)$ is the i-th entry of $\mathbf{s}_k$. Next, consider the transmission of an arbitrary element, say the i-th entry $s_k(i)$. $s_k(i)$ contains two features of $\mathbf{x}_k$, given as

$$s_k(i) = x_k(p) + jx_k(q). \tag{16}$$

Based on constraints (13-15), we have $a_{2i-1,p} = 1$ and $a_{2i,q} = 1$. We use $\mathbf{b}_k \in \mathbb{C}^N$ to denote the pre-coding vector for $\mathbf{s}_k$. $s_k(i)$ is further pre-coded with a scalar $b_k(i)$. After pre-coding, as the implementation of OFDM system, the *inverse fast Fourier transform* (IFFT) operation is used to convert all the $b_k(i)s_k(i),\ i = 1, ..., 2N$ into the time domain. To avoid the *intersymbol interference* (ISI), the cyclic extension of the time domain samples is then added as the guard interval.

The pre-coded $s_k(i)$ is transmitted over the i-th sub-channel. At the edge server, the received signal is aggregated by all the local $s_k(i),\ k = 1, ..., K$, given by

$$\mathbf{y}_i = \sum_{k=1}^{K} \mathbf{h}_k^i b_k(i) s_k(i) + \mathbf{n}, \tag{17}$$

where $\mathbf{n}$ is the additive white Gaussian noise with average noise power δ^2, i.e.,

$$\mathbf{n} \sim \mathcal{CN}(0, \delta^2 \mathbf{I}), \tag{18}$$

and $\mathbf{h}_k^i \in \mathbb{C}^R$ represents the uplink channel gain vector of the i-th sub-channel between device k and the edge server. R is the number of receive antennas. The server has the ability to acquire all the devices' uplink channel gains in the current time slot.

The received signal is first converted into the frequency domain by the *fast Fourier transform* (FFT) operation. Then, a receive beamformer is adopted. The received $\mathbf{y}_i$ after beamforming is

$$\hat{s}(i) = \mathbf{f}_i^H \mathbf{y}_i = \mathbf{f}_i^H \sum_{k=1}^{K} \mathbf{h}_k^i b_k(i) s_k(i) + \mathbf{f}_i^H \mathbf{n}, \tag{19}$$

where $\mathbf{f}_i \in \mathbb{C}^R$ is the beamforming vector. The edge server extracts the real and imaginary parts of $\hat{s}(i)$ as the received features, given by

$$\begin{cases} \hat{x}(p) = \mathsf{Re}(\hat{s}(i)) = \mathsf{Re}\left(\mathbf{f}_i^H \sum_{k=1}^{K} \mathbf{h}_k^i b_k(i) s_k(i) + \mathbf{f}_i^H \mathbf{n}\right), \\ \hat{x}(q) = \mathsf{Im}(\hat{s}(i)) = \mathsf{Im}\left(\mathbf{f}_i^H \sum_{k=1}^{K} \mathbf{h}_k^i b_k(i) s_k(i) + \mathbf{f}_i^H \mathbf{n}\right), \end{cases} \tag{20}$$

where $\mathsf{Re}(\cdot)$ and $\mathsf{Im}(\cdot)$ represent the real and imaginary parts. The received feature vector is denoted as $\hat{\mathbf{x}}$. $\hat{x}(p)$ and $\hat{x}(q)$ are the p-th and q-th entry of $\hat{\mathbf{x}}$. Since the allocation matrix $\mathbf{A}$ is computed by the edge server, the server can put $\hat{x}(p)$ and $\hat{x}(q)$ at the p-th and q-th position of $\hat{\mathbf{x}}$ using $\mathbf{A}$. Along with $\hat{x}(p)$ and $\hat{x}(q)$, the other feature elements will be extracted from the received signals and organized into $\hat{\mathbf{x}}$ as well.

The edge server uses the overall discriminant gain of the received vector, i.e., $G(\hat{\mathbf{x}})$, to measure the inference capability of $\hat{\mathbf{x}}$. The distribution of $\hat{\mathbf{x}}$ and the exact form of $G(\hat{\mathbf{x}})$ will be derived in the following section.

II. Problem Formulation

We use the overall discriminant gain of the received feature vector as the objective function to optimize the allocation scheme, the pre-coding and the beamforming vectors. The optimization objective is

$$\text{maximize} \quad G(\hat{\mathbf{x}}) = \sum_{i=1}^{2N} G(\hat{x}(i)). \tag{21}$$

$G(\hat{\mathbf{x}})$ will be maximized to enhance the inference capability of the received feature vector. As with (10), $G(\hat{x}(i))$ is given by

$$G(\hat{x}(i)) = \frac{2}{L(L-1)} \sum_{\ell'=1}^{L} \sum_{\ell<\ell'} G_{\ell,\ell'}(\hat{x}(i)). \tag{22}$$

To obtain the exact form of $G_{\ell,\ell'}(\hat{x}(i))$, the distribution of $\hat{x}(i)$ is first derived.

A. Zero-forcing Pre-coders

Since the pre-coding and the beamforming design are coupled, the well-known *zero-forcing* (ZF) pre-coders is adopted to simplify the optimization problem and the derivation below. The ZF design is

$$\mathbf{f}_i^H \mathbf{h}_k^i b_k(i) = c_{k,i}, \;\; k = 1, ..., K, \;\; i = 1, ..., N, \tag{23}$$

where $c_{k,i} \geq 0$ is a real number, which represents the strength of the i-th received signal after beamforming from device k. The ZF pre-coders can be written as

$$b_k(i) = \frac{c_{k,i} {\mathbf{h}_k^i}^H \mathbf{f}_i}{\mathbf{f}_i^H \mathbf{h}_k^i {\mathbf{h}_k^i}^H \mathbf{f}_i}. \tag{24}$$

Then, the received features given in (20) can be rewritten as

$$\begin{cases} \hat{x}(p) = \sum_{k=1}^{K} c_{k,i} x_k(p) + \mathsf{Re}(\mathbf{f}_i^H \mathbf{n}), \\ \hat{x}(q) = \sum_{k=1}^{K} c_{k,i} x_k(q) + \mathsf{Im}(\mathbf{f}_i^H \mathbf{n}), \end{cases} \tag{25}$$

where the $x_k(p)$ and $x_k(q)$ is given in (16) as the real and imaginary parts of $s_k(i)$.

B. Problem Simplification

To formalize the optimization problem, we consider the allocation matrix $\mathbf{A}$ together in the derivation. $s_k(i)$ is rewritten as

$$s_k(i) = \mathbf{p}_i \mathbf{A} \mathbf{x}_k, \tag{26}$$

where $\mathbf{p}_i$ is the i-th row of $\mathbf{P}$ and $\mathbf{x}_k$ given in (1) is the sensed data of device k. Recall that the feature vector $\mathbf{x}$ processed by PCA follows the distribution of Gaussian mixture, i.e.,

$$(\mathbf{x}, \ell) \sim \frac{1}{L} \mathcal{N}(\boldsymbol{\mu}_\ell, \boldsymbol{\Sigma}), \;\; \ell \in \{1, ..., L\}, \tag{27}$$

where the class label ℓ of $\mathbf{x}$ is assumed as uniformly distributed. The sensed data with distortion is given by $\mathbf{x}_k = \mathbf{x} + \mathbf{d}_k$, which has the following distribution

$$(\mathbf{x}_k, \ell) \sim \frac{1}{L} \mathcal{N}(\boldsymbol{\mu}_\ell, \boldsymbol{\Sigma} + \mathbf{D}_k), \tag{28}$$

where $\mathbf{D}_k$ is the distortion covariance matrix defined in (2). Thus, the distribution of $s_k(i)$ after the allocation is

$$(s_k(i), \ell) \sim \frac{1}{L} \mathcal{CN}(\mathbf{p}_i \mathbf{A} \boldsymbol{\mu}_\ell, \mathbf{p}_i \mathbf{A} (\boldsymbol{\Sigma} + \mathbf{D}_k) \mathbf{A}^T \mathbf{p}_i^H). \tag{29}$$

All devices' observations are independent, therefore, from (19), the distribution of $\hat{s}(i)$ aggregated by all the $s_k(i)$ is derived as

$$(\hat{s}(i), \ell) \sim \frac{1}{L} \mathcal{CN}(\eta_\ell(i), \psi_i^2). \tag{30}$$

The mean $\eta_\ell(i)$ and the variance ψ_i^2 are given by

$$\begin{cases} \eta_\ell(i) = \sum_{k=1}^{K} c_{k,i}\mathbf{p}_i\mathbf{A}\boldsymbol{\mu}_\ell, \\ \psi_i^2 = \delta^2\mathbf{f}_i^H\mathbf{f}_i + \sum_{k=1}^{K} c_{k,i}^2\mathbf{p}_i\mathbf{A}(\mathbf{\Sigma}+\mathbf{D}_k)\mathbf{A}^T\mathbf{p}_i^H, \end{cases} \tag{31}$$

where $c_{k,i}$ is the positive real number defined in (23), δ^2 is the average power of the channel noise, and $\mathbf{f}^H$ is the receive beamforming vector. Hence, the distribution of $\hat{\mathbf{s}} = [s_k(1), ..., s_k(N)]^T$ with class ℓ is given by

$$(\hat{\mathbf{s}}, \ell) \sim \frac{1}{L}\mathcal{CN}(\boldsymbol{\eta}_\ell, \mathbf{\Psi}),\ \ \ell \in \{1, ..., L\}, \tag{32}$$

where the mean vector $\boldsymbol{\eta}_\ell$ is given as

$$\boldsymbol{\eta}_\ell = [\eta_\ell(1), ..., \eta_\ell(N)]^T, \tag{33}$$

and the covariance matrix $\mathbf{\Psi}$ is diagonal, given as

$$\mathbf{\Psi} = \mathrm{diag}\{\psi_1^2, ..., \psi_N^2\}. \tag{34}$$

The edge server will extract the real and imaginary parts of the entries in $\hat{\mathbf{s}}$ to recover the feature vector $\hat{\mathbf{x}}$. It can be derived that $\hat{\mathbf{x}}$ is the real part of $\mathbf{A}^T\mathbf{P}^H\hat{\mathbf{s}}$, i.e.,

$$\hat{\mathbf{x}} = \mathsf{Re}(\mathbf{A}^T\mathbf{P}^H\hat{\mathbf{s}}). \tag{35}$$

Finally, the distribution of $\hat{\mathbf{x}}$ associated with class ℓ is derived as

$$(\hat{\mathbf{x}}, \ell) \sim \frac{1}{L}\mathcal{CN}(\hat{\boldsymbol{\mu}}_\ell, \hat{\mathbf{\Sigma}}),\ \ \ell \in \{1, ..., L\}, \tag{36}$$

where the mean $\hat{\boldsymbol{\mu}}_\ell$ and the covariance $\hat{\mathbf{\Sigma}}$ are given by

$$\begin{cases} \hat{\boldsymbol{\mu}}_\ell = \mathsf{Re}(\mathbf{A}^T\mathbf{P}^H\boldsymbol{\eta}_\ell) = [\hat{\mu}_\ell(1), ..., \hat{\mu}_\ell(2N)]^T, \\ \hat{\mathbf{\Sigma}} = \mathsf{Re}(\mathbf{A}^T\mathbf{P}^H\mathbf{\Psi}\mathbf{P}\mathbf{A}) = \mathrm{diag}\{\hat{\sigma}_1^2, ..., \hat{\sigma}_{2N}^2\}. \end{cases} \tag{37}$$

Next, based on the distribution in (36) and the definition of discriminant gain, the discriminant gain of $\hat{\mathbf{x}}$ can be derived as

$$\begin{aligned} G(\hat{\mathbf{x}}) &= \frac{2}{L(L-1)}\sum_{i=1}^{2N}\sum_{\ell'=1}^{L}\sum_{\ell<\ell'} G_{\ell,\ell'}(\hat{x}(i)) \\ &= \frac{2}{L(L-1)}\sum_{i=1}^{2N}\sum_{\ell'=1}^{L}\sum_{\ell<\ell'} \frac{(\hat{\mu}_\ell(i)-\hat{\mu}_{\ell'}(i))^2}{\hat{\sigma}_i^2}. \end{aligned} \tag{38}$$

Besides, the constraint of transmit power on each device is given as

$$\sum_{i=1}^{N} b_k(i)s_k(i)s_k^H(i)b_k^H(i) \leq P_k,\ \ k=1,...,K, \tag{39}$$

where $s_k(i)$ carrying two features is the complex scalar to transmit, $b_k(i)$ is the pre-coder for $s_k(i)$, and P_k is the total transmit power on device k. The transmit power constraint can be rewritten with the ZF pre-coders,

$$\sum_{i=1}^{N} \frac{c_{k,i}^2 s_k(i)s_k^H(i)}{\mathbf{f}_i^H \mathbf{h}_k^i {\mathbf{h}_k^i}^H \mathbf{f}_i} \leq P_k,\ \ k=1,...,K. \tag{40}$$

By optimizing the allocation scheme, the pre-coders and the beamforming vectors, our goal is to maximize the discriminant gain of the received feature vector. With the constraints constraints (13-15) on the allocation matrix $\mathbf{A}$, we define the optimization problem as

$$\begin{aligned}
\mathrm{P1}:\quad & \underset{\{c_{k,i}\},\{\mathbf{f}_i\},\{a_{r,c}\}}{\text{maximize}} \quad G(\hat{\mathbf{x}}) = \frac{2}{L(L-1)} \sum_{i=1}^{2N} \sum_{\ell'=1}^{L} \sum_{\ell<\ell'} \frac{(\hat{\mu}_\ell(i)-\hat{\mu}_{\ell'}(i))^2}{\hat{\sigma}_i^2}, \\
& \text{subject to} \quad \sum_{i=1}^{N} \frac{c_{k,i}^2 s_k(i)s_k^H(i)}{\mathbf{f}_i^H \mathbf{h}_k^i {\mathbf{h}_k^i}^H \mathbf{f}_i} \leq P_k,\ \ k=1,...,K, \\
& \qquad \sum_{r=1}^{2N} a_{r,c} = 1,\ \ c=1,...,2N, \\
& \qquad \sum_{c=1}^{2N} a_{r,c} = 1,\ \ r=1,...,2N, \\
& \qquad a_{r,c} \in \{0,1\},\ \ r=1,...,2N,\ \ c=1,...,2N,
\end{aligned} \tag{41}$$

where the notations are from (37) and (40).

III. Algorithm

We adopt an alternating method to resolve the optimization problem, which indicates that problem (41) is divided into two decoupled subproblems. The first subproblem focuses on the assignment of features, whose goal is to optimize the allocation matrix while holding the beamforming vectors and the $\{c_{k,i}\}$ as constants. In contrast, the second subproblem focuses on optimizing the beamforming vectors and the $\{c_{k,i}\}$, while the allocation matrix is fixed. Each subproblem aims at maximizing the discriminant gain of the received feature vector $\hat{\mathbf{x}}$. In each time slot, to obtain the optimal of the original optimization problem, these two subproblem are solved alternatively to optimize each optimization variable.

A. Assignment of Features

In the first subproblem, $\{c_{k,i}\},\ k=1,...,K,\ i=1,...,N$ and the beamforming vectors $\mathbf{f}_i,\ i=1,...,N$ are assumed to be fixed. Only the allocation matrix $\mathbf{A}$ is considered as the optimization variable. We rewrite the mean vector and the covariance matrix in (37) as

$$\begin{cases} \hat{\boldsymbol{\mu}}_\ell = \sum_{k=1}^{K}\sum_{i=1}^{N} c_{k,i}(\mathbf{a}_{2i-1}^T\mathbf{a}_{2i-1}+\mathbf{a}_{2i}^T\mathbf{a}_{2i})\boldsymbol{\mu}_\ell, \\ \hat{\boldsymbol{\Sigma}} = \sum_{i=1}^{N}\left(\delta^2\mathbf{f}_i^H\mathbf{f}_i\mathbf{I}+\sum_{k=1}^{K}c_{k,i}^2(\boldsymbol{\Sigma}+\mathbf{D}_k)\right)(\mathbf{a}_{2i-1}^T\mathbf{a}_{2i-1}+\mathbf{a}_{2i}^T\mathbf{a}_{2i}) \end{cases} \tag{42}$$

where $\mathbf{a}_i$ is the i-th row of matrix $\mathbf{A}$, $\mathbf{I}\in\mathbb{R}^{2N\times 2N}$ is the identity matrix, $\{c_{k,i}\}$ is the real numbers given in (23) and δ^2 is the average power of the channel noise. Then, the discriminant gain of $\hat{\mathbf{x}}$ is rewritten as

$$\begin{aligned} G(\hat{\mathbf{x}}) &= \frac{2}{L(L-1)}\sum_{\ell=1}^{L}\sum_{\ell<\ell'}(\hat{\boldsymbol{\mu}}_\ell-\hat{\boldsymbol{\mu}}_{\ell'})^T\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_\ell-\hat{\boldsymbol{\mu}}_{\ell'}) \\ &= -\sum_{r=1}^{2N}\sum_{c=1}^{2N} t_{r,c}a_{r,c} \end{aligned} \tag{43}$$

where the $a_{r,c}$ is the entry of $\mathbf{A}$ in row r and column c and $t_{r,c}$ denotes the value of $a_{r,c}$ to the discriminant gain if $a_{r,c}=1$, which is given by

$$\begin{aligned} & t_{2r-1,c} = t_{2r,c} \\ &\quad = -\frac{2}{L(L-1)}\left(\sum_{k=1}^{K}c_{k,r}\right)^2\left(\delta^2\mathbf{f}_r^H\mathbf{f}_r+\sum_{k=1}^{K}c_{k,r}^2(\sigma_c^2+\delta_{k,c}^2)\right)^{-1}\sum_{\ell'=1}^{L}\sum_{\ell<\ell'}(\mu_\ell(c)-\mu_{\ell'}(c))^2, \\ &\quad r=1,...,N,\ c=1,...,2N \end{aligned} \tag{44}$$

where the $\mu_\ell(c)$ is the c-th entry of the mean vector $\boldsymbol{\mu}_\ell$ given in (5). The notations of $\delta_{k,r}^2$ and σ_r^2 follow the definition in (3) and (6). We formulate the equivalent form of the first subproblem as

$$\begin{aligned} \text{sub P1}: \quad \underset{\{a_{r,c}\}}{\text{minimize}} \quad & \sum_{r=1}^{2N}\sum_{c=1}^{2N} t_{r,c}a_{r,c}, \\ \text{subject to} \quad & \sum_{r=1}^{2N} a_{r,c}=1,\ c=1,...2N, \\ & \sum_{c=1}^{2N} a_{r,c}=1,\ r=1,...,2N, \\ & a_{r,c}\in\{0,1\},\ r=1,...,2N,\ c=1,...2N, \end{aligned} \tag{41a}$$

Based on that, we relax the last constraint $a_{r,c} \in \{0,1\}$ to $a_{r,c} \geq 0$ and solve the corresponding linear programming problem. The dual problem of the linear program is given by

$$\begin{aligned} \text{dual}: \quad & \underset{\{a_{r,c}\}}{\text{maximize}} \quad \sum_{r=1}^{2N} u_r + \sum_{c=1}^{2N} v_c, \\ & \text{subject to} \quad u_r + v_c \leq t_{r,c}, \;\; r = 1, ..., 2N, \;\; c = 1, ..., 2N. \end{aligned} \tag{45}$$

The optimal values of the primal and dual problems are attained if the *Karush-Kuhn-Tucker* (KKT) conditions are satisfied, which is given as

$$\begin{aligned} & \sum_{r=1}^{2N} a_{r,c} = 1, && c = 1, ..., 2N, \\ & \sum_{c=1}^{2N} a_{r,c} = 1, && r = 1, ..., 2N, \\ & a_{r,c} \geq 0, && r = 1, ..., 2N, \;\; c = 1, ..., 2N, \\ & u_r + v_c \leq t_{r,c}, && r = 1, ..., 2N, \;\; c = 1, ..., 2N, \\ & a_{r,c}(t_{r,c} - u_r - v_c) = 0, && r = 1, ..., 2N, \;\; c = 1, ..., 2N, \end{aligned} \tag{46}$$

The optimal values can be attained iteratively with *The Hungarian method*. We briefly describe the algorithm here

Algorithm 1 The Hungarian method

Input: $N, \; \{t_{r,c}\}$

Output: $\{a_{r,c}\}$

```
initialize the matching set M to ∅
u_r ← min_{1≤c≤2N}{t_{r,c}}, for 1 ≤ r ≤ 2N
v_c ← min_{1≤r≤2N}{t_{r,c} − u_r}, for 1 ≤ c ≤ 2N
repeat
    t_{r,c} ← t_{r,c} − u_r − v_c, for 1 ≤ r ≤ 2N, 1 ≤ c ≤ 2N
    for r = 1 to N and c = 1 to 2N do
        if t_{r,c} is 0 then
            a_{r,c} ← 1
        end if
    end for
    find a minimal cover S in A
until |M| = 2N
```

B. sub problem II

Whereas, with fixed allocation matrix $\mathbf{A}$, the second subproblem focuses on optimizing the $\{c_{k,i}\}$ and the beamforming vectors, which is equivalently formulated as

$$\begin{aligned} \text{sub P2}: \quad & \underset{\{c_{k,i}\},\{\mathbf{f}_i\}}{\text{maximize}} \quad G(\hat{\mathbf{x}}) = \frac{2}{L(L-1)} \sum_{\ell'=1}^{L} \sum_{\ell<\ell'} (\hat{\boldsymbol{\mu}}_\ell - \hat{\boldsymbol{\mu}}_{\ell'})^T \hat{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{\mu}}_\ell - \hat{\boldsymbol{\mu}}_{\ell'}), \\ & \text{subject to} \quad \sum_{i=1}^{N} \frac{c_{k,i}^2 s_k(i) s_k^H(i)}{\mathbf{f}_i^H \mathbf{h}_k^i {\mathbf{h}_k^i}^H \mathbf{f}_i} \leq P_k, \ k = 1, ..., K, \end{aligned} \tag{41b}$$

where the notations are form (41), the mean $\hat{\boldsymbol{\mu}}_\ell$ and the covariance $\hat{\boldsymbol{\Sigma}}$ are given in (42). The second subproblem can be further simplified as

$$\begin{aligned} \text{sub P2}: \quad & \underset{\{c_{k,i}\},\{\mathbf{f}_i\}}{\text{maximize}} \quad \sum_{i=1}^{N} \sum_{j=2i-1}^{2i} u_{\theta(j)} \left(\sum_{k=1}^{K} c_{k,i} \right)^2 \left(\delta^2 \mathbf{f}_i^H \mathbf{f}_i + \sum_{k=1}^{K} c_{k,i}^2 (\sigma_{\theta(j)}^2 + \delta_{k,\theta(j)}^2) \right)^{-1}, \\ & \text{subject to} \quad \sum_{i=1}^{N} \frac{c_{k,i}^2 s_k(i) s_k^H(i)}{\mathbf{f}_i^H \mathbf{h}_k^i {\mathbf{h}_k^i}^H \mathbf{f}_i} \leq P_k, \ k = 1, ..., K, \end{aligned} \tag{41b}$$

where $\theta \in \mathbb{Z}_+^{2N}$ is a vector of the rearranged indices from $\{1, 2, ..., 2N\}$, which is given by

$$\theta = \mathbf{A}(1, 2, ..., 2N)^T, \tag{47}$$

and $u_{\theta(j)}$ is given by

$$u_{\theta(j)} = \frac{2}{L(L-1)} \sum_{\ell'=1}^{L} \sum_{\ell<\ell'} (\mu_\ell(\theta(j)) - \mu_{\ell'}(\theta(j)))^2, \tag{48}$$

To make the constraint in (41b) to be concave, we use $\hat{\mathbf{f}}_{k,i}$ to substitute $\mathbf{f}_i$ and $c_{k,i}$, given by

$$\hat{\mathbf{f}}_{k,i} = \frac{1}{c_{k,i}} \mathbf{f}_i \tag{49}$$

and the constraint is rewritten as

$$\sum_{i=1}^{N} \frac{||s_k(i)||^2}{\hat{\mathbf{f}}_{k,i}^H \mathbf{h}_k^i {\mathbf{h}_k^i}^H \hat{\mathbf{f}}_{k,i}} - P_k \leq 0, \ k = 1, ..., K, \tag{50}$$

$$\begin{aligned} \max_{\{c_{k,i}\},\{\mathbf{f}_i\},\{\alpha_{i,j}\}} \quad & \sum_{i=1}^{N} \sum_{j=2i-1}^{2i} \alpha_{i,j}, \\ \text{s.t.} \quad & \sum_{i=1}^{N} \frac{c_{k,i}^2 s_k(i) s_k^H(i)}{\mathbf{f}_i^H \mathbf{h}_k^i {\mathbf{h}_k^i}^H \mathbf{f}_i} \leq P_k, \ k = 1, ..., K, \\ & u_{\theta(j)} \left(\sum_{k=1}^{K} c_{k,i} \right)^2 \geq \alpha_{i,j} \left(\delta^2 \mathbf{f}_i^H \mathbf{f}_i + \sum_{k=1}^{K} c_{k,i}^2 (\sigma_{\theta(j)}^2 + \delta_{k,\theta(j)}^2) \right), \ i = 1, ..., N, \ j = 2i-1, 2i \end{aligned}$$

$$
\begin{aligned}
\max_{\{c_{k,i}\},\{\mathbf{f}_i\},\{\alpha_{i,j}\}} \quad & \sum_{i=1}^{N}\sum_{j=2i-1}^{2i}\alpha_{i,j}, \\
\mathrm{s.\,t.} \quad & \sum_{i=1}^{N}\frac{c_{k,i}^2 s_k(i)s_k^H(i)}{\mathbf{f}_i^H\mathbf{h}_k^i{\mathbf{h}_k^i}^H\mathbf{f}_i} \le P_k,\ \ k=1,...,K, \\
& {u_{\theta(j)}}^{-1}\left(\delta^2\mathbf{f}_i^H\mathbf{f}_i+\sum_{k=1}^{K}c_{k,i}^2(\sigma_{\theta(j)}^2+\delta_{k,\theta(j)}^2)\right)\le\frac{1}{\alpha_{i,j}},\ \ i=1,...,N,\ \ j=2i-1,2i, \\
& \sum_{k=1}^{K}c_{k,i}=1,\ \ i=1,...,N
\end{aligned}
$$

$$
\frac{1}{\alpha_{i,j}^{(t+1)}}-\frac{1}{\alpha_{i,j}^{(t)}}\ge-\left(\frac{1}{\alpha_{i,j}^{(t)}}\right)^2(\alpha_{i,j}^{(t+1)}-\alpha_{i,j}^{(t)})
$$

$$
{\mathbf{f}_i^{(t+1)}}^H\mathbf{h}_k^i{\mathbf{h}_k^i}^H\mathbf{f}_i^{(t+1)}-{\mathbf{f}_i^{(t)}}^H\mathbf{h}_k^i{\mathbf{h}_k^i}^H\mathbf{f}_i^{(t)}\ge 2(\mathbf{h}_{k,R}^i+\mathbf{h}_{k,I}^i)^T(\mathbf{f}_{i,R}^{(t+1)}-\mathbf{f}_{i,R}^{(t)})+2(\mathbf{h}_{k,I}^i-\mathbf{h}_{k,R}^i)^T(\mathbf{f}_{i,I}^{(t+1)}-\mathbf{f}_{i,I}^{(t)})
$$

$$
\begin{aligned}
\max_{\{c_{k,i}^{(t+1)}\},\{\mathbf{f}_i^{(t+1)}\},\{\alpha_{i,j}^{(t+1)}\}} \quad & \sum_{i=1}^{N}\sum_{j=2i-1}^{2i}\alpha_{i,j}^{(t+1)}, \\
\mathrm{s.\,t.} \quad & \sum_{k=1}^{K}c_{k,i}^{(t+1)}=1,\ \ i=1,...,N, \\
& \sum_{i=1}^{N}\frac{{c_{k,i}^{(t+1)}}^2 s_k(i)s_k^H(i)}{{\mathbf{f}_i^{(t)}}^H\mathbf{h}_k^i{\mathbf{h}_k^i}^H\mathbf{f}_i^{(t)}+g(\mathbf{f}_{i,R}^{(t+1)})+g(\mathbf{f}_{i,I}^{(t+1)})}\le P_k,\ \ k=1,...,K, \\
& g(\mathbf{f}_{i,R}^{(t+1)})=2(\mathbf{h}_{k,R}^i+\mathbf{h}_{k,I}^i)^T(\mathbf{f}_{i,R}^{(t+1)}-\mathbf{f}_{i,R}^{(t)}), \\
& g(\mathbf{f}_{i,I}^{(t+1)})=2(\mathbf{h}_{k,I}^i-\mathbf{h}_{k,R}^i)^T(\mathbf{f}_{i,I}^{(t+1)}-\mathbf{f}_{i,I}^{(t)}), \\
& {u_{\theta(j)}}^{-1}\left(\delta^2{\mathbf{f}_i^{(t+1)}}^H\mathbf{f}_i^{(t+1)}+\sum_{k=1}^{K}{c_{k,i}^{(t+1)}}^2(\sigma_{\theta(j)}^2+\delta_{k,\theta(j)}^2)\right)\le\frac{1}{\alpha_{i,j}^{(t)}}-\left(\frac{1}{\alpha_{i,j}^{(t)}}\right)^2(\alpha_{i,j}^{(t+1)}-\alpha_{i,j}^{(t)}), \\
& i=1,...,N,\ \ j=2i-1,2i
\end{aligned}
$$

$$\begin{aligned}
\min_{\{c_{k,i}\}} \quad & -\sum_{i=1}^{N}\sum_{j=2i-1}^{2i}\frac{u_{\theta(j)}}{\delta^2\mathbf{f}_i^H\mathbf{f}_i+\sum_{k=1}^{K}c_{k,i}^2(\sigma_{\theta(j)}^2+\delta_{k,\theta(j)}^2)},\\
\text{s.t.} \quad & \sum_{i=1}^{N}\frac{c_{k,i}^2 s_k(i)s_k^H(i)}{\mathbf{f}_i^H\mathbf{h}_k^i{\mathbf{h}_k^i}^H\mathbf{f}_i}-P_k\le 0,\ \ k=1,...,K,\\
& \sum_{k=1}^{K}c_{k,i}=1,\ \ i=1,...,N
\end{aligned}$$

$$\begin{aligned}
\max_{\{\mathbf{f}_i\}} \quad & \sum_{i=1}^{N}\sum_{j=2i-1}^{2i}\frac{u_{\theta(j)}}{\delta^2\mathbf{f}_i^H\mathbf{f}_i+\sum_{k=1}^{K}c_{k,i}^2(\sigma_{\theta(j)}^2+\delta_{k,\theta(j)}^2)},\\
\text{s.t.} \quad & \sum_{i=1}^{N}\frac{c_{k,i}^2 s_k(i)s_k^H(i)}{\mathbf{f}_i^H\mathbf{h}_k^i{\mathbf{h}_k^i}^H\mathbf{f}_i}-P_k\le 0,\ \ k=1,...,K
\end{aligned}$$

$$\begin{aligned}
\max_{\{c_{k,i}\},\{\mathbf{f}_i\}} \quad & \sum_{i=1}^{N}\sum_{j=2i-1}^{2i}\frac{u_{\theta(j)}}{\delta^2\mathbf{f}_i^H\mathbf{f}_i+\sum_{k=1}^{K}c_{k,i}^2(\sigma_{\theta(j)}^2+\delta_{k,\theta(j)}^2)},\\
\text{s.t.} \quad & \sum_{i=1}^{N}\frac{c_{k,i}^2 s_k(i)s_k^H(i)}{\mathbf{f}_i^H\mathbf{h}_k^i{\mathbf{h}_k^i}^H\mathbf{f}_i}\le P_k,\ \ k=1,...,K,\\
& \sum_{k=1}^{K}c_{k,i}=1,\ \ i=1,...,N
\end{aligned}$$

$$\begin{aligned}
\max_{\{c_{k,i}\},\{\mathbf{f}_i\}} \quad & \sum_{i=1}^{N}\sum_{j=2i-1}^{2i}u_{\theta(j)}\left(\sum_{k=1}^{K}c_{k,i}\right)^2,\\
\text{s.t.} \quad & \sum_{i=1}^{N}\frac{c_{k,i}^2 s_k(i)s_k^H(i)}{\mathbf{f}_i^H\mathbf{h}_k^i{\mathbf{h}_k^i}^H\mathbf{f}_i}\le P_k,\ \ k=1,...,K,\\
& \delta^2\mathbf{f}_i^H\mathbf{f}_i+\sum_{k=1}^{K}c_{k,i}^2(\sigma_{\theta(j)}^2+\delta_{k,\theta(j)}^2)=1,\ \ i=1,...,N,\ \ j=2i-1,2i
\end{aligned}$$

$$x_k(p) \sim \frac{1}{L}\mathcal{N}(\mu_\ell(p), \sigma_p^2 + \delta_{k,p}^2)$$
$$x_k(q) \sim \frac{1}{L}\mathcal{N}(\mu_\ell(q), \sigma_q^2 + \delta_{k,q}^2)$$

$$s_k(i) = x_k(p) + jx_k(q)$$
$$s_k(i) \sim \frac{1}{L}\mathcal{CN}(\mu_\ell(p) + j\mu_\ell(q), \sigma_p^2 + \sigma_q^2 + \delta_{k,p}^2 + \delta_{k,q}^2)$$

$$s(i) \sim \frac{1}{L}\mathcal{CN}\left((\mu_\ell(p) + j\mu_\ell(q))\sum_{k=1}^{K} c_{k,i}, \delta^2 \mathbf{f}_i^H \mathbf{f}_i + \sum_{k=1}^{K}(\sigma_p^2 + \sigma_q^2 + \delta_{k,p}^2 + \delta_{k,q}^2)c_{k,i}^2\right)$$

$$\hat{x}(p) \sim \frac{1}{L}\mathcal{N}\left(\mu_\ell(p)\sum_{k=1}^{K} c_{k,i}, \frac{\delta^2}{2}\mathbf{f}_i^H \mathbf{f}_i + \sigma_p^2 \sum_{k=1}^{K} c_{k,i}^2 + \sum_{k=1}^{K} \delta_{k,p}^2 c_{k,i}^2\right)$$
$$\hat{x}(q) \sim \frac{1}{L}\mathcal{N}\left(\mu_\ell(q)\sum_{k=1}^{K} c_{k,i}, \frac{\delta^2}{2}\mathbf{f}_i^H \mathbf{f}_i + \sigma_q^2 \sum_{k=1}^{K} c_{k,i}^2 + \sum_{k=1}^{K} \delta_{k,q}^2 c_{k,i}^2\right)$$

REFERENCES

[1] Q. Lan, Q. Zeng, P. Popovski, D. Gündüz, and K. Huang, "Progressive Feature Transmission for Split Inference at the Wireless Edge," *arXiv:2112.07244 cs, eess, math*, Dec. 2021.