AirComp Based Multi-View Sensing for Edge Inference

immediate

I. System Model

A. Network and Sensing Model

Consider a system with one server equipped with a multi-antenna *access point* (AP) and K single-antenna sensing devices, e.g., radar sensors and cameras. The server aims at aggregating the real-time sensing results from the devices to generate a feature vector for finishing an inference task with low latency. Each sensor's observation is a polluted version of the ground-true one, due to the sensing noise and only having the partial view of the information source. Specifically, for device k, its sensed data vector can be written as

$$\mathbf{x}_k = \mathbf{x} + \mathbf{d}_k, \quad 1 \le k \le K, \tag{1}$$

where $\mathbf{x} = [x_1, x_2, ..., x_M]^T$ is the ground-true sensing result of the information source, or called the ground-true feature vector, $\mathbf{x}_k = [x_{k,1}, x_{k,2}, ..., x_{k,M}]^T$, and $\mathbf{d}_k$ is the sensing distortion. According to [x], the sensing distortion $\mathbf{d}_k$ have the following Gaussian distribution:

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

where $\mathcal{N}(\cdot,\cdot)$ is the Gaussian distribution and $\mathbf{D}_k \in \mathbb{R}^{M\times M}$ is the diagonal covariance matrix, given as

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

Besides, the wireless transmission works in a time-division manner, i.e., the total transmit time duration is divided into multiple time slots. In each time slot, the technique of AirComp is used for aggregating certain feature dimensions of $\mathbf{x}$ (which is elaborated later in Section X). Without loss of generality, the channel is assumed to be static in one slot and varies among different slots. The channel gain of device k in the current slot is denoted as $\mathbf{h}_k \in \mathbb{C}^N$, with N being the number of receive antennas at the server and $\mathbb{C}^N$ being a complex vector space with the dimension of N. Moreover, the server is assumed to work as the coordinator and have the ability to acquire the channel gains of all devices' uplink links in the current time slot.

B. Feature Distribution and Inference Model

1) Feature Distribution: Following the setting in [x-x], the ground-true data feature vector, is assumed to follow a *Gaussian mixture* and is used to inference a well trained model with L classes. Given the data distribution, the method of *principle component analysis* (PCA) is applied to extract the principle feature space for alleviating the communication overhead. Thats' to say, for each local observation at each device, only the features projected to the principle space is transmitted to the server. The detailed procedure is described as follows.

- At the training stage, PCA is first performed by the server over the training dataset to extract the principle dimensions of each training sample. The learning model is trained using the principle feature dimensions.
- At the inference stage, before the server aggregates the local observations from each device, the principle eigen-space is broadcast to each device. For each local observation at the devices, only the features projected to the principle subspace is transmitted.

Without loss of generality, the local observations and the ground-true feature vector, say $\{\mathbf{x}_k\}$ and $\mathbf{x}$ defined in (1), are already projected to the principle eigen-space using PCA. As a result, each element of $\mathbf{x}$ is independent of other elements therein. It follows that the distribution of $\mathbf{x}$ can be written as

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

where L is the total number of classes, $\boldsymbol{\mu}_\ell \in \mathbb{R}^M$ is the centroid of the ℓ-th class, given as

$$\boldsymbol{\mu}_\ell = [\mu_{\ell,1}, \mu_{\ell,2}, ..., \mu_{\ell,M}]^T, \quad 1 \leq \ell \leq L, \tag{5}$$

and $\boldsymbol{\Sigma} \in \mathbb{R}^{M\times M}$ is a diagonal covariance matrix, given as

$$\boldsymbol{\Sigma} = \mathrm{diag}\{\sigma_1^2, \sigma_2^2, ..., \sigma_M^2\}. \tag{6}$$

2) Inference Capability: Following the settings of [X], the discriminant gain is used to measure the the inference capability, which is based on the well known *symmetric Kullback-Leibler (KL) divergence* proposed in [1]. Specifically, consider an arbitrary class pair, say classes

ℓ and ℓ', and the feature space expanded by all data samples, say $\{\mathbf{x}\}$. Based on the distribution of $\mathbf{x}$ in (4) and the KL divergence metric in [1], the pair-wise discriminant gain is defined as

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

where x_m is the m-th element in $\mathbf{x}$ and $G_{\ell,\ell'}(x_m)$ is given as

$$G_{\ell,\ell'}(x_m) = \frac{\left(\mu_{\ell,m} - \mu_{\ell',m}\right)^2}{\sigma_m^2}, \quad 1 \leq m \leq M, \tag{8}$$

and the notations follow that in (5) and (6). Then, the overall discriminant gain is defined as the average over the all pair-wise discriminant gains in (7), given as

$$G(\mathbf{x}) = \frac{2}{L(L-1)} \sum_{\ell'=1}^{L} \sum_{\ell<\ell'} G_{\ell,\ell'}(\mathbf{x}) = \frac{2}{L(L-1)} \sum_{\ell'=1}^{L} \sum_{\ell<\ell'} \sum_{m=1}^{M} G_{\ell,\ell'}(x_m) = \sum_{m=1}^{M} G(x_m), \tag{9}$$

where $G(x_m)$ is the discriminant gain of the m-th dimension, given as

$$G(x_m) = \frac{2}{L(L-1)} \sum_{\ell'=1}^{L} \sum_{\ell<\ell'} \frac{\left(\mu_{\ell,m} - \mu_{\ell',m}\right)^2}{\sigma_m^2}, \quad 1 \leq m \leq M. \tag{10}$$

C. AirComp Model

The technique of AirComp is used to aggregate the local observations, say the local feature vectors $\{\mathbf{x}_k\}$, form all device, as it can significantly enhance the communication efficiency. Specifically, in each time slot, each device transmit a complex scalar, as each device has only one transmit antenna. The real part and the imaginary part of the scalar contain one feature element, respectively. At the server, AirComp is performed to aggregate the two feature elements and separately estimate their ground-true versions. Thereby, the whole feature vector can be grouped into different element pairs, which can be sequentially aggregated among different time slots. Obviously, the design of AirComp among different time slots are the same.

Consider an arbitrary time slot to aggregate the m_1-th and m_2-th local elements from all devices, say $\{x_{k,m_1}, x_{k,m_2}, \ \forall k\}$, to estimate the ground-true ones, say $\{x_{m_1}, x_{m_2}\}$. The procedure of AirComp is shown in Fig. 1 and is described as follows. For an arbitrary device, say the k-th, its local observation is first pre-processed by PCA. Then, the m_1-th and m_2-th principle feature dimensions, say x_{k,m_1} and x_{k,m_2}, are combine in one symbol for transmission, as

$$s_k = x_{k,m_1} + j x_{k,m_2}, \quad 1 \leq k \leq K, \tag{11}$$

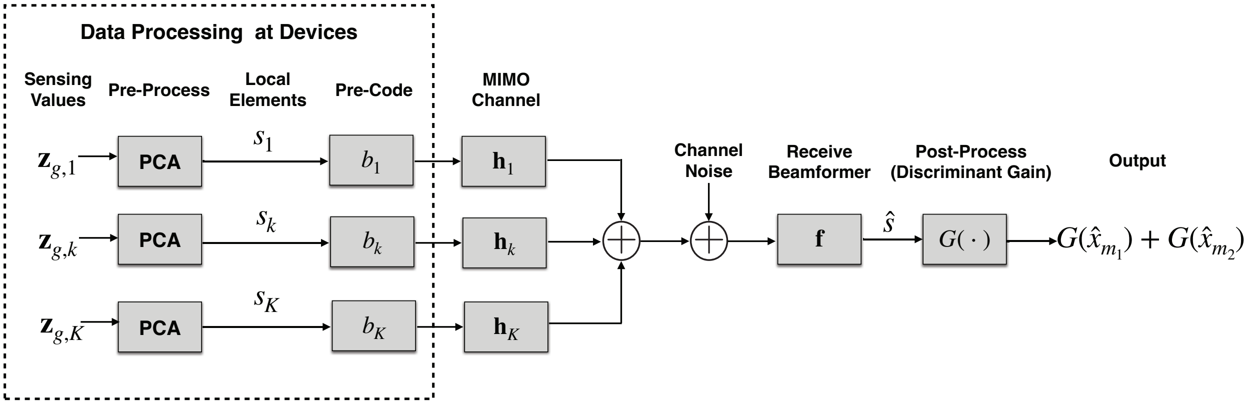

Fig. 1. Block diagram of AirComp for Feature Aggregation.

where $s_k \in \mathbb{C}$ is the transmitted symbol and j represents the imaginary unit. Next, s_k is further pre-coded with a scalar $b_k \in \mathbb{C}$ and transmitted over a *multiple-input-multiple-output* (MIMO) channel. At the server, the receive signal is the aggregation of all local elements, given as

$$\mathbf{y}_m = \sum_{k=1}^{K} \mathbf{h}_k b_k s_k + \mathbf{n}, \tag{12}$$

where s_k is the transmit symbol of device k defined in (11), $\mathbf{h}_k \in \mathbb{C}^N$ is the uplink channel gain vector of device k, b_k is the pre-coding scalar, $\mathbf{n}$ is the additive white Gaussian noise with the following distribution:

$$\mathbf{n} \sim \mathcal{N}\left(\mathbf{0}, \delta_0^2 \mathbf{I}\right), \tag{13}$$

and δ_0^2 is the noise variance. Next, a receive beamforming vector $\mathbf{f} \in \mathbb{C}^N$ is used to aggregate all local s_k to generate the estimates of the ground-true feature elements x_{m_1} and x_{m_2}. Specifically, the received symbol after receive beamforming can be written as

$$\hat{s} = \mathbf{f}^H \mathbf{y}_m = \mathbf{f}^H \sum_{k=1}^{K} \mathbf{h}_k b_k s_k + \mathbf{f}^H \mathbf{n}, \tag{14}$$

where the notations follow that in (12). It follows that the estimates x_{m_1} and x_{m_2} are given by

$$\begin{cases} \hat{x}_{m_1} = \mathsf{Re}\left(\hat{s}\right) = \mathsf{Re}\left(\mathbf{f}^H \sum_{k=1}^{K} \mathbf{h}_k b_k s_k + \mathbf{f}^H \mathbf{n}\right), \\ \hat{x}_{m_2} = \mathsf{Im}\left(\hat{s}\right) = \mathsf{Im}\left(\mathbf{f}^H \sum_{k=1}^{K} \mathbf{h}_k b_k s_k + \mathbf{f}^H \mathbf{n}\right), \end{cases} \tag{15}$$

where $\mathsf{Re}(\cdot)$ and $\mathsf{Im}(\cdot)$ are the functions to extract real part and imaginary part of one complex number, respectively, and other notations follow that in (12). Finally, $\hat{x}_{m_1}$ and $\hat{x}_{m_2}$ are post-processed to output the discriminant gain $G(\hat{x}_{m_1}) + G(\hat{x}_{m_2})$.

II. Problem Formulation & Simplification

A. Problem Formulation

Different from the traditional AirComp design, which aims at minimizing the distortion between the estimate values $\{\hat{x}_{m_1}, \hat{x}_{m_2}\}$ and the ground-true ones $\{x_{m_1}, x_{m_2}\}$ without consideration of the post-processing, in this work, the objective is to maximize the post-processing function, say the sum discriminant gains of $\hat{x}_{m_1}$ and $\hat{x}_{m_2}$. Specifically, the objective is

$$\max \; G = G(\hat{x}_{m_1}) + G(\hat{x}_{m_2}), \tag{16}$$

where $\hat{x}_{m_1}$ and $\hat{x}_{m_2}$ defined in (15) are the estimates of the ground-true feature elements, and $G(\hat{x}_{m_1})$ and $G(\hat{x}_{m_2})$ are the corresponding discriminant gains.

Besides, there is one constraint on the transmit power of each device, given by

$$b_k s_k s_k^H b_k^H \leq P_k, \quad 1 \leq k \leq K, \tag{17}$$

where b_k is the pre-coding scalar at device k, b_k^H is the hermitian of b_k, s_k is the transmit symbol, and P_k is the total transmit power of device k.

Proposition 1 (Diverse Local Symbol Variance)**.** *The variances of different local symbols are not uniform, as the sensing distortion levels of different devices, say $\{\mathbf{d}_k\}$ defined in* (2)*, are different. As a result, the variance of s_k can not be neglected in the power constraint in* (17)*.*

Nonetheless, the transmit symbol variance, say $s_k s_k^H$, is known by each device. Therefore, the power constraint in (17) can be re-written as

$$b_k b_k^H \leq \hat{P}_k, \quad 1 \leq k \leq K, \tag{18}$$

where $\hat{P}_k$ is the maximum transmit pre-coding power, given as

$$\hat{P}_k = \frac{P_k}{s_k s_k^H}, \quad 1 \leq k \leq K. \tag{19}$$

In summary, the discriminant gain maximization problem can be written as

$$\text{(P1)} \qquad \begin{aligned} \max_{\{b_k\}, \mathbf{f}} \;\; & G = G(\hat{x}_{m_1}) + G(\hat{x}_{m_2}), \\ \text{s.t.} \;\; & b_k b_k^H \leq \hat{P}_k, \quad 1 \leq k \leq K, \end{aligned} \tag{20}$$

where the notations follow that in (16) and (18).

Proposition 2 (Challenges for Solving (P1))**.** *To achieve the solution of (P1) is difficult due to the following two reasons. On one hand, the design of the receive beamforming* $\mathbf{f}$ *and the pre-coding scalars* $\{b_k\}$ *are coupled [see* (15)*], making (P1) non-convex. On the other hand, each estimate defined in* (15) *could include the both ground-true elements, say* x_{m_1} *and* x_{m_2}*, due to the channel rotation. This leads to a complicated distribution of* $\hat{x}_{m_1}$ *and* $\hat{x}_{m_2}$*, and thus a complicated expression of the discriminant gains* $G(\hat{x}_{m_1})$ *and* $G(\hat{x}_{m_2})$*.*

III. Discriminant Gains with Zero-Forcing Pro-coders

To address the challenges in Proposition 2, in this section, (P1) is simplified with two steps. The well-known *zero-forcing* pre-coders is first used to simplify the received estimates $\{\hat{x}_{m_1}, \hat{x}_{m_2}\}$. Then, based on the ZF pre-coders, the discriminant gains, i.e., $G(\hat{x}_{m_1})$ and $G(\hat{x}_{m_2})$, are derived to simplify the objective function.

1) ZF pre-coders: First, the ZF design is given by

$$\mathbf{f}^H \mathbf{h}_k b_k = c_k, \quad 1 \leq k \leq K, \tag{21}$$

where $\mathbf{f}^H$ is the receive beam-forming vector, $\mathbf{h}_k$ is the channel vector of device k, b_k is pre-coder of device k, and $c_k \geq 0$ is a real number representing the receive signal strength from device k. Then, the ZF pre-coders can be derived as

$$b_k = \frac{c_k \mathbf{h}_k^H \mathbf{f}}{\mathbf{h}_k^H \mathbf{f}\mathbf{f}^H \mathbf{h}_k}. \tag{22}$$

It follows that the power constraint in (P1) can be re-written as

$$c_k^2 \leq P_k \mathbf{h}_k^H \mathbf{f}\mathbf{f}^H \mathbf{h}_k, \quad 1 \leq k \leq K. \tag{23}$$

Besides, by substituting the pre-coders in (22) and $\hat{s}_k$ in (11) into the estimates, say $\hat{x}_{m_1}$ and $\hat{x}_{m_2}$ in (15), we can obtain

$$\begin{cases} \hat{x}_{m_1} = \mathsf{Re}\left(\sum_{k=1}^{K} c_k s_k + \mathbf{f}^H \mathbf{n}\right) = \sum_{k=1}^{K} c_k x_{k,m_1} + \mathsf{Re}(\mathbf{f}^H \mathbf{n}), \\ \hat{x}_{m_2} = \mathsf{Im}\left(\sum_{k=1}^{K} c_k s_k + \mathbf{f}^H \mathbf{n}\right) = \sum_{k=1}^{K} c_k x_{k,m_2} + \mathsf{Im}(\mathbf{f}^H \mathbf{n}), \end{cases} \tag{24}$$

where the notations follow that in (11), (15), and (22).

2) Discriminat Gains: To achieve the discriminant gain G, in the sequel, the distributions of the local transmit elements, say x_{k,m_1} and x_{k,m_2}, are first derived. Then, based on the ZF pre-coders, the distribution of the received elements, say $\hat{x}_{m_1}$ and $\hat{x}_{m_2}$, are derived. Next, the discriminant gains are obtained, followed by the derivation of a simplified problem of (P1).

First, recall the local elements x_{k,m_1} and x_{k,m_2} are given by

$$x_{k,m_i} = x_{m_i} + d_{k,m_i}, \quad i = 1,2, \ \& \ 1 \le k \le K, \tag{25}$$

where the distribution of the ground-true element $x_{m,i}$ is given by

$$x_{m_i} \sim \frac{1}{L}\mathcal{N}\left(\mu_{\ell,m_i}, \sigma^2_{m_i}\right), \quad i = 1,2, \tag{26}$$

according the distribution of $\mathbf{x}$ in (4), (5), and (6), and the distribution of the distortion d_{k,m_i} is given by

$$d_{k,m_i} \sim \mathcal{N}\left(0, \delta^2_{k,m_i}\right), \quad i = 1,2, \tag{27}$$

according to the distribution of $\mathbf{d}_k$ in (2) and (3). Subsequently, the following lemma in terms of x_{k,m_i}'s distribution can be obtained.

Lemma 1. *The distribution of the local elements* $\{x_{k,m_i}\}$ *can be derived as*

$$x_{k,m_i} \sim \frac{1}{L}\mathcal{N}\left(\mu_{\ell,m_i}, \sigma^2_{m_i} + d^2_{k,m_i}\right), \quad i = 1,2, \ \& \ 1 \le k \le K, \tag{28}$$

Proof: Please see Appendix X.

Then, by substituting the distributions of $\{x_{k,m_1}, x_{k,m_2}\}$ in (28) and the distribution of Gaussian noise $\mathbf{n}$ in (13) into the estimates $\{\hat{x}_{m_1}, \hat{x}_{m_2}\}$ in (24), their distributions can be derived as shown in Lemma 2.

Lemma 2. *The distribution of the estimated elements* $\{x_{k,m_i}\}$ *are given by*

$$\hat{x}_{m_i} \sim \frac{1}{L}\mathcal{N}\left(\hat{\mu}_{\ell,m_i}, \hat{\sigma}^2_{m_i}\right), \quad i = 1,2, \tag{29}$$

where the means $\{\hat{\mu}_{\ell,m_i}\}$ and the variance $\{\hat{\sigma}^2_{m_i}\}$ are

$$\begin{cases} \hat{\mu}_{\ell,m_1} = \sum_{k=1}^{K} c_k \mu_{\ell,m_1}, \\ \hat{\sigma}^2_{m_1} = \sigma^2_{m_1} \left(\sum_{k=1}^{K} c_k \right)^2 + \sum_{k=1}^{K} c_k^2 \delta^2_{k,m_1} + \frac{\delta_0^2}{2} \left(\mathbf{f}_1^T \mathbf{f}_1 + \mathbf{f}_2^T \mathbf{f}_2 \right), \\ \hat{\mu}_{\ell,m_2} = \sum_{k=1}^{K} c_k \mu_{\ell,m_2}, \\ \hat{\sigma}^2_{m_2} = \sigma^2_{m_2} \left(\sum_{k=1}^{K} c_k \right)^2 + \sum_{k=1}^{K} c_k^2 \delta^2_{k,m_2} + \frac{\delta_0^2}{2} \left(\mathbf{f}_1^T \mathbf{f}_1 + \mathbf{f}_2^T \mathbf{f}_2 \right), \end{cases} \tag{30}$$

where c_k is the receive signal strength from device k, μ_{ℓ,m_1} and μ_{ℓ,m_2} are the centroids of the m_1-th and m_2-th elements regarding the ℓ-th class, $\sigma^2_{m_1}$ and $\sigma^2_{m_2}$ are the variance of the m_1-th and m_2-th elements, δ_{k,m_1} and δ_{k,m_2} are the distortion variance of the m_1-th and m_2-th elements at device k, and $\mathbf{f}_1 = \mathsf{Re}(\mathbf{f})$ and $\mathbf{f}_2 = \mathsf{Im}(\mathbf{f})$ are the real part and imaginary part of the receive beamforming $\mathbf{f}$, respectively.

Proof: Please see Appendix X.

Next, based on the distributions in Lemma 2 and the definition of discriminant gain in (10), the discriminant gains of $\{x_{k,m_1}, x_{k,m_2}\}$ can be derived as

$$G(x_{m_i}) = \frac{2}{L(L-1)} \sum_{\ell'=1}^{L} \sum_{\ell<\ell'} \frac{\left(\hat{\mu}_{\ell,m_i} - \hat{\mu}_{\ell',m_i} \right)^2}{\hat{\sigma}^2_{m_i}}, \quad i = 1, 2, \tag{31}$$

where $\{\hat{\mu}_{\ell,m_i}\}$ and $\{\hat{\sigma}^2_{m_i}\}$ are defined in (30).

Finally, by substituting the discriminant gains of $\{x_{k,m_1}, x_{k,m_2}\}$ in (31) and the power constraint in (23) into (P1), it can be equivalently derived as

$$\text{(P2)} \quad \begin{aligned} \max_{\{c_k\},\mathbf{f}_1,\mathbf{f}_2} \quad & G = \frac{2}{L(L-1)} \sum_{i=1}^{2} \sum_{\ell'=1}^{L} \sum_{\ell<\ell'} \frac{\left(\hat{\mu}_{\ell,m_i} - \hat{\mu}_{\ell',m_i} \right)^2}{\hat{\sigma}^2_{m_i}}, \\ \text{s.t.} \quad & c_k^2 \leq P_k \mathbf{h}_k^H \left(\mathbf{f}_1 \mathbf{f}_1^T + \mathbf{f}_2 \mathbf{f}_2^T \right) \mathbf{h}_k, \quad 1 \leq k \leq K, \end{aligned} \tag{32}$$

where the notations follow that in (30).

References

[1] G. Saon and M. Padmanabhan, “Minimum bayes error feature selection for continuous speech recognition,” *Advances in neural information processing systems*, vol. 13, pp. 800–806, 2000.