Task-Oriented Over-the-Air Computation for Multi-Device Edge AI

Abstract

In this paper, a task-oriented over-the-air computation (AirComp) technique is proposed for multi-device edge split inference systems in order to support low-latency artificial intelligent (AI) services at the network edge. In this system, local feature vectors are extracted from the real-time noise-corrupted sensory data on each device. Targeting suppressing the sensing noise for enhancing the inference accuracy, the local feature vectors are aggregated at the server via AirComp to generate a global one, which is used for finishing the remaining inference task. Under this setup and by considering classification tasks, an approximate but tractable approach, called discriminant gain, is adopted to measure the inference accuracy. It is defined as the centroids distance between two classes normalized by their covariance in the Euclidean feature space. To maximize the discriminant gain, we first quantify the influence of sensing noise and channel noise on it by deriving a closed-form expression. This task-oriented AirComp scheme, however, leads to a non-convex optimization problem due to the complicated form of discriminant gain and the joint design of transmit precoding and receive beamforming. To address the problem, variables transformation is applied to equivalently convert it into a problem, which has a convex objective function and is shown to contain convex feasible sub-regions. Thanks to this convex structure, it is solved by the method of successive convex approximation (SCA). The performance of the proposed scheme is verified by extensive experimental results, which are performed based on a concrete human motion recognition task.

I. Introduction

Edge artificial intelligence (AI) emerges as a promising technique towards 6G, as it can support various of intelligent services, such as Metaverse, auto-driving, and virtual reality (VR), at the network edge [1], [2], [3], [4], [5]. The realization of edge AI includes two stages: edge (machine) learning, which targets distilling intelligence from massive distributed data (see e.g., [6], [7], [8]), and edge inference, which deploys well-trained models and uses their prediction capabilities for decision making (see e.g., [9], [10]). Although edge learning has attracted much research focus (see e.g., [11], [12], [13]), the implementation of AI services ultimately depends on edge inference [14], [15], [16], [17], which can assist edge devices with making real-time intelligent

decisions. Hence, edge inference is a practical issue for enabling AI services and deserves more research attentions. However, the deployment of edge inference faces new challenges. On one hand, the goal of edge inference systems is no longer the conventional throughput maximization but inference performance, i.e., high accuracy and low latency. This calls for new *task-oriented design principles* [18], [19], [20], [21], [22]. On the other hand, it faces the communication bottleneck, since the real-time AI service (e.g., human motion recognition in auto-driving) has a low-latency requirement but needs to access multiple devices for suppressing the sensing noise caused on each device. To address these challenges, this work aims to design an *inference-task-oriented* communication-efficient multiple access technique, called over-the-air computation (AirComp).

There have been several techniques targets the efficient deployment of edge inference. Among others, the most popular is edge split inference, which divides an AI model into two parts: one deployed on resource-limited devices for feature extraction [e.g., principal component analysis (PCA) and convolution], and the other at an edge server for finishing the remaining computation-intensive inference task. As a result, the split inference technique enjoys the advantages of keeping data privacy via avoiding raw data transmission and low computation overhead on devices by task offloading, and is adopted in this work. One main research focus on edge split inference is the trade-off between the computation and communication overhead on edge device via, e.g., compressing the feature map of the split layer [23], [24], [25], two-step pruning strategy [26], progressive feature transmission [27], setting early existing point [28], [29], and joint source and channel coding using deep neural networks [30]. However, the aforementioned designs focus on the case of single edge device, which either senses the source in a narrow view to obtain highly accurate sensory data by focusing a single angle, or obtains a noise-corrupted wide-view sensory data for wide angle object detection by e.g., scanning from angle to angle [31]. To address the inefficient data caused by the narrow view, a multi-device cooperated multi-view edge inference technique is proposed in [32] to maximize the inference accuracy via the design of sensing, computation, and communication integration. However, the issue of suppressing the noise of wide-view sensory data for inference accuracy enhancement remains unresolved and is the theme of this paper.

In this work, a multi-device edge inference system is considered. Each device obtains a noise-corrupted version of the ground-true wide-view sensory data and extracts from it a noisy local feature vector using simple linear operations like convolution, PCA, etc.. To suppress the sensing

noise, we adopt a common approach, which derives a weighted sum of all local feature vectors [33], [34]. To this end, the technique of over-the-air computation (AirComp) can be employed to enhance the communication efficiency. It can support fast data aggregation from a large number of devices [35]. Specifically, in AirComp, signals from all devices are allowed to transmit simultaneously over the same frequency band. At the receiver, a functional value of all signals is directly calculated using the waveform superposition property of wireless channels, instead of decoding the data stream from each device. There have been comprehensive research for the efficient implementation of AirComp, including the design of beamforming in multi-input-multi-output (MIMO) system (see e.g., [36], [37], [38], [39], [40]), power control for complementing channel diverse channel fading (see e.g., [41]), the investigation of tradeoff between computation and energy efficiency (see e.g., [42]), the design of unmanned aerial vehicle (UAV) assisted AirComp (see e.g., [43]), etc.. Furthermore, due to the superior ability of accessing massive devices with low latency, AirComp has been a promising technique widely utilized in federated edge learning systems, as it can greatly alleviate the communication overhead via fast data aggregation (see .g., [44], [45], [46], [47], [48]). Besides, authors in [49] proposed a AirComp based edge inference system, where the same inference task is performed in multiple devices and a server aggregates the local inference results and make a final decision based on majority vote. However, the design is the traditional on-device inference, which causes huge computation overhead at the resource limited edge devices. It remains an uncharted area to implement edge split inference using AirComp, and thus motivates the work of this paper.

In the considered multi-device edge inference system, the server is equipped with multiple antennas and all devices equip with one antenna. The server aggregates all local feature vectors via AirComp to estimate a global one, which is further used for the remaining inference task. The inference-task-oriented AirComp faces new challenges. To begin with, the traditional AirComp design criterion, i.e., minimum mean square error (MMSE) used in existing literature, is no longer effective for edge split inference systems. To be specific, the schemes based on MMSE minimize the distortion between the received symbol and the target one, which generally is the weighted sum of all transmitted symbols. It ignores the influence of the post-processing. However, in the context of edge inference, minimal distortion of the feature vector not necessarily leads to high performance, i.e., inference accuracy, since same distortion level on different feature elements have various influence on the accuracy. As an example, a classification task is shown in Fig. 1, whose feature vector has two elements (dimensions). It is observed that feature element

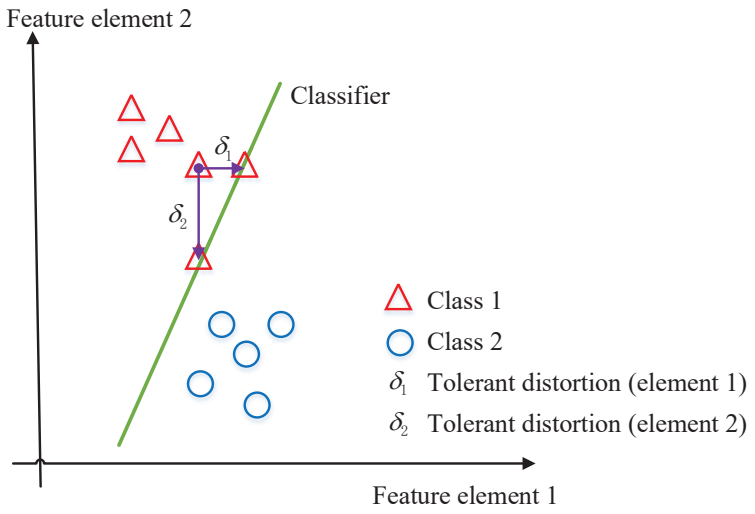


Fig. 1. Various distortion tolerance of different feature elements in classification tasks: For a distortion level δ_1 obtained under the MMSE criterion, incorrect inference occurs if it is on element 1, but the inference result is correct if it is on element 2.

2 is more tolerant to distortion than element 1 in terms of violating the inference accuracy. Obviously, in the case of MMSE, the various importance level is ignored and thus may lead to a bad performance. To address this issue, the best approach is directly maximizing the inference accuracy in the AirComp design. However, there is lack of metric to measure the instantaneous inference accuracy, especially before the feature sample is inputted to the AI model. Alternatively, this work adopts an approximate but tractable statistic model, namely discriminant gain, as the accuracy metric for classification tasks. It is proposed in [27] and is built on the well-known KL divergence [50]. Specifically, for arbitrary two classes in the Euclidean feature space, discriminant gain is the distance of their centroids normalized by their covariance. With a larger distance, the two classes are better separated, which leads to a higher inference accuracy. However, the design of AirComp under the criterion of discriminant gain maximization still faces the challenges arising from the complicated form of the objective function, as well as the joint design of transmit precoding and receive beamforming.

To address the challenges above, the technique of inference-task-oriented AirComp is proposed in this paper. The detailed contributions are listed below.

- **AirComp based Edge Split Inference Systems**: An AirComp based multi-device edge split inference system is established, where the feature vector used for inference at the server is estimated by aggregating all noisy local feature vectors via AirComp. In each time

slot, two feature elements are transmitted using a complex scalar, since each device has only one transmit antenna. The whole feature vector can be transmitted sequentially over different time slots using the same approach. Under the system settings, the influence of sensing noise and channel noise on the inference accuracy measured by discriminant gain is quantified using a closed-form expression.

- **Inference Accuracy Maximization via AirComp**: The task-oriented principle is adopted in this work. That's to say, the AirComp design in the multi-device edge inference system aims at maximizing the inference accuracy measured by discriminant gain instead of the conventional MMSE, via the joint design of transmit precoding and receive beamforming.
- **Joint Design of Transmit Precoding and Receive Beamforming**: The formulated inference accuracy maximization problem is non-convex. To tackle the problem, variables transformation is first applied to equivalently derive it into a problem, whose optimum is shown to have a convex lower bound. Benefiting from the convexity, the problem is solved using the successive convex approximation (SCA) approach.
- **Performance Evaluation**: In order to assess the effectiveness of our proposed scheme, we run extensive simulations over the high-fidelity wireless sensing simulator proposed in [51] while taking into account the specific task of wide-view human motion recognition with two inference models, i.e., support vector machine (SVM) and multi-layer perception (MLP) neural network, respectively. It is demonstrated that, for both models using SVM and MLP neural networks, maximizing the discriminant gain is effective in maximizing the inference accuracy. Additionally, it is demonstrated that the proposed scheme greatly outperforms the benchmark schemes based on MMSE and random design in terms of inference accuracy. It is also shown that aggregating the local feature vectors from multiple devices can efficiently suppress the sensing noise to enhance the inference accuracy.

The organization of this paper is as follows. The system model is introduced in Section II. The problem formulation and simplification is in Section III. The joint design of receive power control and receive beamforming is proposed in Section IV. The performance evaluation is in Section V, and Section VI concludes the paper.

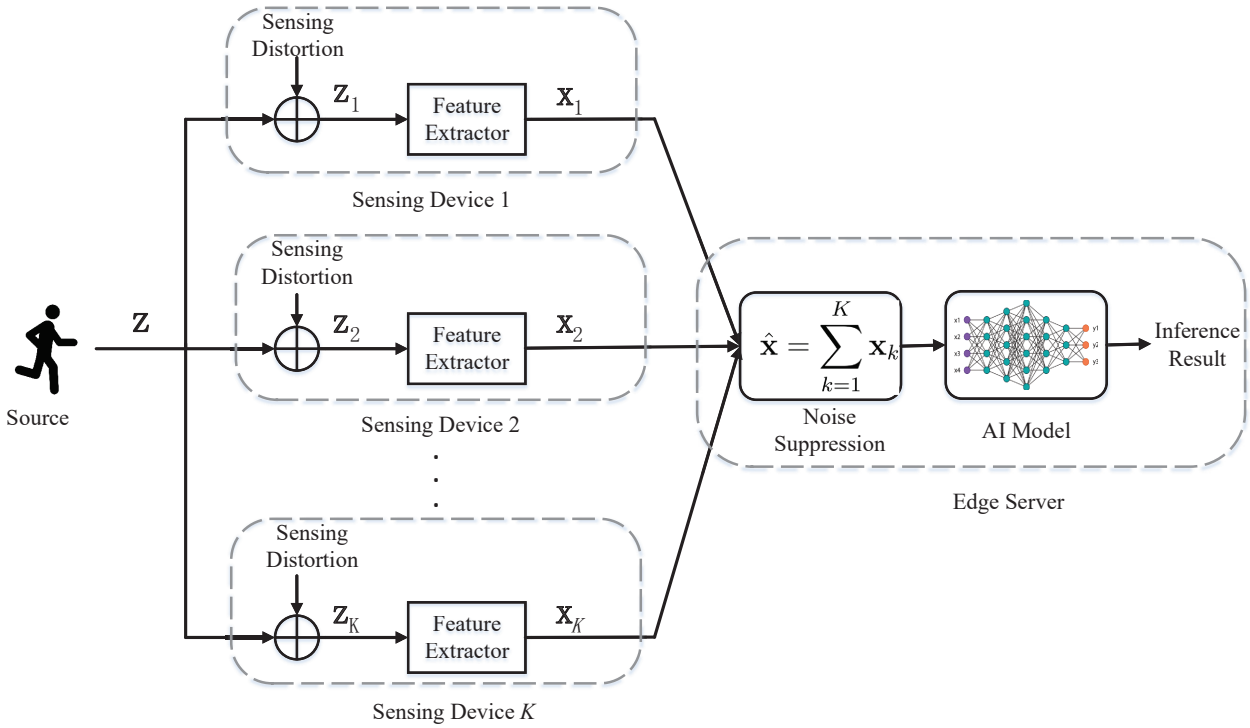


Fig. 2. Model of Over-the-air Computation Based Edge Inference Systems.

II. System Model

A. Network and Sensing Model

Consider an edge inference system where there are one server equipped with a multi-antenna access point (AP) and K single-antenna sensing devices, e.g., radar sensors and cameras, as shown in Fig. 2. The server aims at aggregating the local feature vectors, which are extracted from the real-time sensory data, on all devices to form a global one for finishing the remaining inference task. As each device need to sense a wide-angle range (e.g., for object detection), their observations are noise-corrupted versions of the ground-true one [31]. The sensory data of device k is given as

$$\mathbf{z}_k = \mathbf{z} + \mathbf{e}_k, \tag{1}$$

where $\mathbf{z} = [z_1, z_2, ..., z_S]^T$ is the ground-true sensory data of the source, $\mathbf{z}_k = [z_{k,1}, z_{k,2}, ..., z_{k,S}]^T$ is the local observation of device k, $\mathbf{e}_k$ is the sensing noise, and S is the dimension of the raw sensory data. According to [33], [34], different elements of the sensing noise vector follow

indentical and independent zero-mean Gaussian distributions:

$$\mathbf{e}_k \sim \mathcal{N}\left(\mathbf{0}, \epsilon^2 \mathbf{I}\right), \tag{2}$$

where $\mathcal{N}(\cdot, \cdot)$ is the Gaussian distribution, ϵ^2 is the sensing noise power, and $\mathbf{I} \in \mathbb{R}^{S \times S}$ is the identical matrix.

Besides, the server and the sensing devices communicate via wireless links. Time-division multiple access is adopted. The channels are assumed to be static in each time slot and varying among different slots. The channel gain of device k 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.

B. Feature Generation and Distribution

In this part, the feature generation procedure is first introduced, followed by the description of the feature distribution.

1) Feature Generation: As transmitting the raw sensory data with large dimensions causes large communication overhead, as well as violates the data privacy, an alternative approach is to move the feature extraction part (e.g., PCA and convolutional operations) of an AI model on devices. In this work, PCA is adopted to extract a latent low-dimensional feature sub-space from the raw sensory data on each device. The detailed procedure is described as follows.

- At the training stage, PCA is first performed by the server over the offline training dataset to extract the principal dimensions of each sample. The learning model is trained using the principal feature dimensions.
- At the inference stage, before the server aggregates the local observations from each device, the principal eigen-space is broadcast to each device. For each device, the local feature vector is extracted by projecting the sensory data into the principal eigen-space, and then transmitted.

Thereby, the extracted local feature vectors from the sensory data $\mathbf{z}_k$ can be expressed as

$$\mathbf{x}_k = \mathbf{U}^T \mathbf{z}_k = \mathbf{U}^T \mathbf{z} + \mathbf{U}^T \mathbf{e}_k = \mathbf{x} + \mathbf{d}_k, \quad 1 \leq k \leq K, \tag{3}$$

where $\mathbf{U}$ is a $S \times M$ real column unitary matrix representing the principal eigen-space of PCA, M is the dimension of the principal feature eigen-space,

$$\mathbf{x} = \mathbf{U}^T \mathbf{z} = [x_1, x_2, ..., x_M]^T, \tag{4}$$

is the ground-true feature vector, and

$$\mathbf{d}_k = \mathbf{U}^T \mathbf{e}_k, \quad 1 \leq k \leq K, \tag{5}$$

is the projected noise vector of device k. By substituting the distribution of $\mathbf{e}_k$ in (2), the distribution of $\mathbf{d}d_k$ can be derived as

$$\mathbf{d}_k \sim \mathcal{N}\left(\mathbf{0}, \epsilon^2 \mathbf{I}\right), \quad 1 \leq k \leq K, \tag{6}$$

where the variance is derived from

$$\mathbb{E}\left(\mathbf{d}_k^T \mathbf{d}_k\right) = \mathbb{E}\left(\mathbf{U}^T \mathbf{e}_k \mathbf{e}_k^T \mathbf{U}\right) = \mathbf{U}^T \mathbb{E}(\mathbf{e}_k \mathbf{e}_k^T)\mathbf{U} = \epsilon^2 \mathbf{I}. \tag{7}$$

2) Feature Distribution: Following the setting in [27], the ground-true data feature vector $\mathbf{x}$ is assumed to follow a Gaussian mixture to finish the model inference of a classification task with L classes, given as

$$\mathcal{F}(\mathbf{x}) = \frac{1}{L} \sum_{\ell=1}^{L} \mathcal{F}_\ell(\mathbf{x}), \tag{8}$$

where $\mathcal{F}_\ell(\mathbf{x})$ is the Gaussian distribution of $\mathbf{x}$ in terms of the ℓ-th class. As PCA is performed, different feature elements are linearly independent. It follows that $\mathcal{F}_\ell(\mathbf{x})$ is given by

$$\mathcal{F}_\ell(\mathbf{x}) = \mathcal{N}\left(\boldsymbol{\mu}_\ell, \boldsymbol{\Sigma}\right), \quad 1 \leq \ell \leq L, \tag{9}$$

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

$$\boldsymbol{\mu}_\ell = [\mu_{\ell,1}, \mu_{\ell,2}, ..., \mu_{\ell,M}]^T, \quad 1 \leq \ell \leq L, \tag{10}$$

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

$$\boldsymbol{\Sigma} = \mathrm{diag}\{\sigma_1^2, \sigma_2^2, ..., \sigma_M^2\}. \tag{11}$$

C. Inference Capability

In this work, the metric *discriminant gain* proposed in [27] is adopted as the inference accuracy for classification tasks. For arbitrary two classes, the discriminant gain represents the distance between their centroids in the Euclidean feature space under normalized covariance, as presented in Fig. 3. That says, a larger discriminant gain between two classes indicates that they are more likely to be differentiated, and thus leads to a higher inference accuracy. In the sequel, the mathematical model of discriminant gain is introduced.

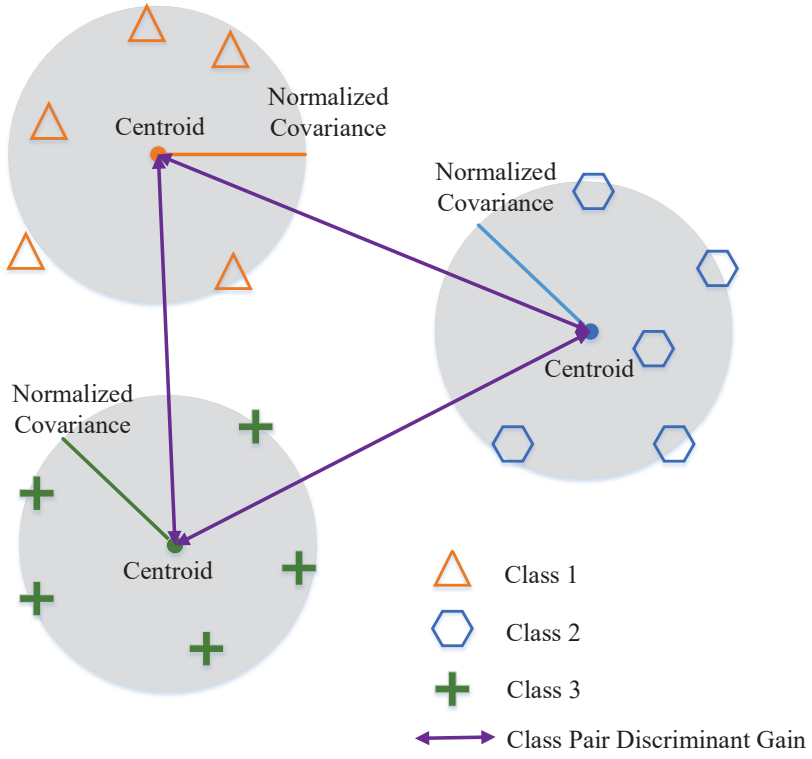


Fig. 3. Geometric interpretation of discriminant gain in the feature space.

Discriminant gain is derived from the well known KL divergence proposed in [50]. 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 (8) and $\mathcal{F}_\ell(\mathbf{x})$ in (9), the pair-wise discriminant gain is defined as

$$\begin{aligned} G_{\ell,\ell'}(\mathbf{x}) &= \mathsf{D}_{KL}\left[\mathcal{F}_\ell(\mathbf{x}) \,\|\, \mathcal{F}_{\ell'}(\mathbf{x})\right] + \mathsf{D}_{KL}\left[\mathcal{F}_{\ell'}(\mathbf{x}) \,\|\, \mathcal{F}_\ell(\mathbf{x})\right], \\ &= \int_{\mathbf{x}} \mathcal{F}_\ell(\mathbf{x}) \log\left[\frac{\mathcal{F}_{\ell'}(\mathbf{x})}{\mathcal{F}_\ell(\mathbf{x})}\right] \mathrm{d}\mathbf{x} + \int_{\mathbf{x}} \mathcal{F}_{\ell'}(\mathbf{x}) \log\left[\frac{\mathcal{F}_\ell(\mathbf{x})}{\mathcal{F}_{\ell'}(\mathbf{x})}\right] \mathrm{d}\mathbf{x}, \\ &= \left(\boldsymbol{\mu}_\ell - \boldsymbol{\mu}_{\ell'}\right)^T \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{\mu}_\ell - \boldsymbol{\mu}_{\ell'}\right), \quad \forall(\ell, \ell'), \end{aligned} \tag{12}$$

where $\mathsf{D}_{KL}(\cdot\|\cdot)$ is the KL divergence defined in [50]. As different feature elements are independent, it follows that

$$G_{\ell,\ell'}(\mathbf{x}) = \sum_{m=1}^{M} G_{\ell,\ell'}(x_m), \tag{13}$$

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

$$G_{\ell,\ell'}(x_m) = \frac{\left(\mu_{\ell,m} - \mu_{\ell',m}\right)^2}{\sigma_m^2}, \quad 1 \le m \le M, \tag{14}$$

and the other notations follow that in (10) and (11). Then, the overall discriminant gain is defined as the average of all pair-wise discriminant gains in (12), given as

$$\begin{aligned} G(\mathbf{x}) &= \frac{2}{L(L-1)} \sum_{\ell'=1}^{L} \sum_{\ell<\ell'} G_{\ell,\ell'}(\mathbf{x}), \\ &= \frac{2}{L(L-1)} \sum_{\ell'=1}^{L} \sum_{\ell<\ell'} \sum_{m=1}^{M} G_{\ell,\ell'}(x_m), \\ &= \sum_{m=1}^{M} G(x_m), \end{aligned} \tag{15}$$

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

$$G(x_m) = \frac{2}{L(L-1)} \sum_{\ell'=1}^{L} \sum_{\ell<\ell'} \frac{\left(\mu_{\ell,m} - \mu_{\ell',m}\right)^2}{\sigma_m^2}, \quad 1 \le m \le M. \tag{16}$$

D. AirComp Model

The technique of AirComp is used to aggregate the local feature vectors $\{\mathbf{x}_k\}$ form all devices, as it can suppress the sensing noise and significantly enhance the communication efficiency. Specifically, in each time slot, each device transmit a complex scalar symbol, as each device has only one transmit antenna. The whole feature vector can be aggregated sequentially over different time slots using the same approach. The real part and the imaginary part of the complex scalar symbol 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 transmitted in a time-division way over several time slots. Obviously, the design of AirComp in all time slots are the same. Without loss of generality, in the sequel, the transmission in an arbitrary time slot is considered.

Consider the case where the server aggregates the m_1-th and m_2-th local elements from all devices, say $\{x_{k,m_1}, x_{k,m_2},\ 1 \le k \le K\}$, to estimate the ground-true ones, say $\{x_{m_1}, x_{m_2}\}$. The procedure of AirComp is shown in Fig. 4 and is described as follows. For an arbitrary device, say the k-th, its local sensory data is first pre-processed by PCA to extract the principal feature elements. Then, the m_1-th and m_2-th principal feature dimensions, say x_{k,m_1} and x_{k,m_2}, are combined in one symbol for transmission, as

$$s_k = x_{k,m_1} + j x_{k,m_2}, \quad 1 \le k \le K, \tag{17}$$

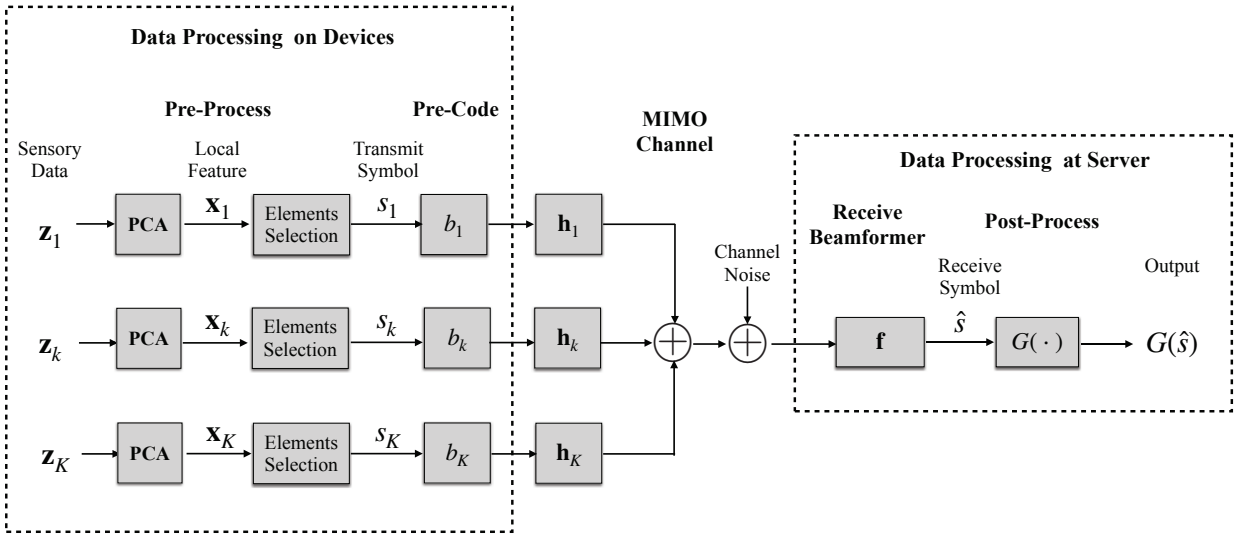


Fig. 4. Block diagram of AirComp for Feature Aggregation.

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 MIMO channel. At the server, the receive signal is the aggregation of all transmit symbols, given as

$$\mathbf{y}_m = \sum_{k=1}^{K} \mathbf{h}_k b_k s_k + \mathbf{n}, \tag{18}$$

where $\mathbf{n}$ is the additive white Gaussian noise with the following distribution:

$$\mathbf{n} \sim \mathcal{N}\left(\mathbf{0}, \delta_0^2 \mathbf{I}\right), \tag{19}$$

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 symbols $\{s_k\}$ 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

$$\hat{s} = \mathbf{f}^H \mathbf{y}_m = \mathbf{f}^H \sum_{k=1}^{K} \mathbf{h}_k b_k s_k + \mathbf{f}^H \mathbf{n}. \tag{20}$$

It follows that the estimates x_{m_1} and x_{m_2} are given by

$$\begin{cases} \hat{x}_{m_1} = \mathsf{Re}\left(\hat{s}\right) = \mathsf{Re}\left(\mathbf{f}^H \sum_{k=1}^{K} \mathbf{h}_k b_k s_k + \mathbf{f}^H \mathbf{n}\right), \\ \hat{x}_{m_2} = \mathsf{Im}\left(\hat{s}\right) = \mathsf{Im}\left(\mathbf{f}^H \sum_{k=1}^{K} \mathbf{h}_k b_k s_k + \mathbf{f}^H \mathbf{n}\right), \end{cases} \tag{21}$$

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 (18). Finally, $\hat{x}_{m_1}$ and $\hat{x}_{m_2}$ are post-processed to output the discriminant gain $G(\hat{x}_{m_1}) + G(\hat{x}_{m_2})$.

III. Problem Formulation and Simplification

A. Problem Formulation

Different from the traditional AirComp design, which aims at minimizing the distortion between the estimated feature elements $\{\hat{x}_{m_1}, \hat{x}_{m_2}\}$ 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 $\hat{x}_{m_2}$, given as

$$\max \; G = G(\hat{x}_{m_1}) + G(\hat{x}_{m_2}), \tag{22}$$

where $\hat{x}_{m_1}$ and $\hat{x}_{m_2}$ defined in (21) 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

$$b_k \mathbb{E}(s_k s_k^H) b_k^H \leq P_k, \quad 1 \leq k \leq K, \tag{23}$$

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. The transmit symbol variance, say $\mathbb{E}(s_k s_k^H)$, can be estimated from the offline training data samples, and thus is known by the edge server as a prior information. Therefore, the power constraint in (23) can be re-written as

$$b_k b_k^H \leq \hat{P}_k, \quad 1 \leq k \leq K, \tag{24}$$

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

$$\hat{P}_k = \frac{P_k}{\mathbb{E}(s_k s_k^H)}, \quad 1 \leq k \leq K. \tag{25}$$

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

$$\text{(P1)} \quad \begin{aligned} &\max_{\{b_k\}, \mathbf{f}} \; G = G(\hat{x}_{m_1}) + G(\hat{x}_{m_2}), \\ &\text{s.t.} \; \; b_k b_k^H \leq \hat{P}_k, \quad 1 \leq k \leq K. \end{aligned} \tag{26}$$

The formulation of (P1) follows the task-oriented principle. To be specific, the inference accuracy measured by the discriminant gain is maximized instead of using the MMSE criterion. That's because MMSE does not necessarily lead to high performance due to the fact that same distortion

on different feature elements has different impact on the accuracy. The task-oriented formulation, however, causes new challenges. To begin with, the discriminant gain has a complicated non-convex sum-of-ratios form. Besides, the design of the receive beamforming $\mathbf{f}$ and the pre-coding scalars $\{b_k\}$ are coupled [see (21)]. Moreover, the feature elements in the received symbol are cross coupled, i.e., each estimated feature element defined in (21) could be a linear combination of the ground-true elements x_{m_1} and x_{m_2}, due to 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 $G(\hat{x}_{m_2})$.

B. Discriminant Gains with Zero-Forcing Pre-coders

To address the challenges mentioned above, in this section, (P1) is simplified with two steps. The well-known zero-forcing (ZF) pre-coders is first used to simplify the estimated feature elements $\{\hat{x}_{m_1}, \hat{x}_{m_2}\}$. 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

$$\mathbf{f}^H \mathbf{h}_k b_k = c_k, \quad 1 \leq k \leq K, \tag{27}$$

where $\mathbf{f}^H$ is the receive beamforming 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 (power) from device k. Then, the ZF pre-coders can be derived as

$$b_k = \frac{c_k \mathbf{h}_k^H \mathbf{f}}{\mathbf{h}_k^H \mathbf{f}\mathbf{f}^H \mathbf{h}_k}, \quad 1 \leq k \leq K. \tag{28}$$

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

$$c_k^2 \leq P_k \mathbf{h}_k^H \mathbf{f}\mathbf{f}^H \mathbf{h}_k, \quad 1 \leq k \leq K. \tag{29}$$

Besides, by substituting the pre-coders in (28) and $\hat{s}_k$ in (17) into the estimates, say $\hat{x}_{m_1}$ and $\hat{x}_{m_2}$ in (21), we can obtain

$$\begin{cases} \hat{x}_{m_1} = \mathsf{Re}\left(\sum_{k=1}^{K} c_k s_k + \mathbf{f}^H \mathbf{n}\right) = \sum_{k=1}^{K} c_k x_{k,m_1} + \mathsf{Re}(\mathbf{f}^H \mathbf{n}), \\ \hat{x}_{m_2} = \mathsf{Im}\left(\sum_{k=1}^{K} c_k s_k + \mathbf{f}^H \mathbf{n}\right) = \sum_{k=1}^{K} c_k x_{k,m_2} + \mathsf{Im}(\mathbf{f}^H \mathbf{n}), \end{cases} \tag{30}$$

where the notations follow that in (17), (21), and (28).

2) Discriminant Gains: To achieve the discriminant gain G, in the sequel, the distributions of the local transmit feature elements $\{x_{k,m_1}, x_{k,m_2}\}$ are first derived. Then, based on the ZF precoders, the distribution of the received elements $\{\hat{x}_{m_1}, \hat{x}_{m_2}\}$ 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

$$x_{k,m_i} = x_{m_i} + d_{k,m_i}, \quad i = 1, 2, \ \& \ 1 \leq k \leq K, \tag{31}$$

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

$$x_{m_i} \sim \frac{1}{L} \sum_{\ell=1}^{L} \mathcal{N}\left(\mu_{\ell,m_i}, \sigma_{m_i}^2\right), \quad i = 1, 2, \tag{32}$$

according the distribution of $\mathbf{x}$ in (8), (10), and (11), and the distribution of the noise d_{k,m_i} is given by

$$d_{k,m_i} \sim \mathcal{N}\left(0, \epsilon^2\right), \quad i = 1, 2, \tag{33}$$

according to the distribution of $\mathbf{d}_k$ in (6). 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*

$$x_{k,m_i} \sim \frac{1}{L} \sum_{\ell=1}^{L} \mathcal{N}\left(\mu_{\ell,m_i}, \sigma_{m_i}^2 + \epsilon^2\right), \quad i = 1, 2, \ \& \ 1 \leq k \leq K, \tag{34}$$

Proof: Please see Appendix A.

Then, by substituting the distributions of $\{x_{k,m_1}, x_{k,m_2}\}$ in (34) and the distribution of the channel noise $\mathbf{n}$ in (19) into the receive feature elements $\{\hat{x}_{m_1}, \hat{x}_{m_2}\}$ in (30), 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*

$$\hat{x}_{m_i} \sim \frac{1}{L} \mathcal{N}\left(\hat{\mu}_{\ell,m_i}, \hat{\sigma}_{m_i}^2\right), \quad i = 1, 2, \tag{35}$$

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

$$\begin{cases} \hat{\mu}_{\ell,m_1} = \sum\limits_{k=1}^{K} c_k \mu_{\ell,m_1}, \\ \hat{\sigma}^2_{m_1} = \sigma^2_{m_1} \left(\sum\limits_{k=1}^{K} c_k \right)^2 + \sum\limits_{k=1}^{K} c_k^2 \delta^2_{k,m_1} + \dfrac{\delta_0^2}{2} \left(\mathbf{f}_1^T \mathbf{f}_1 + \mathbf{f}_2 \mathbf{f}_2 \right), \\ \hat{\mu}_{\ell,m_2} = \sum\limits_{k=1}^{K} c_k \mu_{\ell,m_2}, \\ \hat{\sigma}^2_{m_2} = \sigma^2_{m_2} \left(\sum\limits_{k=1}^{K} c_k \right)^2 + \sum\limits_{k=1}^{K} c_k^2 \delta^2_{k,m_2} + \dfrac{\delta_0^2}{2} \left(\mathbf{f}_1^T \mathbf{f}_1 + \mathbf{f}_2^T \mathbf{f}_2 \right), \end{cases} \tag{36}$$

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 noise 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 B.

Next, based on the distributions in Lemma 2 and the definition of discriminant gain in (16), the discriminant gains of $\{x_{k,m_1}, x_{k,m_2}\}$ can be derived as

$$G(\hat{x}_{m_i}) = \frac{2}{L(L-1)} \sum_{\ell'=1}^{L} \sum_{\ell<\ell'} \frac{\left(\hat{\mu}_{\ell,m_i} - \hat{\mu}_{\ell',m_i}\right)^2}{\hat{\sigma}^2_{m_i}}, \quad i = 1, 2, \tag{37}$$

where $\{\hat{\mu}_{\ell,m_i}\}$ and $\{\hat{\sigma}^2_{m_i}\}$ are defined in (36).

Finally, by substituting the discriminant gains of $\{x_{k,m_1}, x_{k,m_2}\}$ in (37) and the power constraint in (29) into (P1), it can be equivalently derived as

$$\text{(P2)} \quad \begin{aligned} \max_{\{c_k\},\mathbf{f}_1,\mathbf{f}_2} \quad & G = \frac{2}{L(L-1)} \sum_{i=1}^{2} \sum_{\ell'=1}^{L} \sum_{\ell<\ell'} \frac{\left(\hat{\mu}_{\ell,m_i} - \hat{\mu}_{\ell',m_i}\right)^2}{\hat{\sigma}^2_{m_i}}, \\ \text{s.t.} \quad & c_k^2 \le P_k \mathbf{h}_k^H \left(\mathbf{f}_1 \mathbf{f}_1^T + \mathbf{f}_2 \mathbf{f}_2^T \right) \mathbf{h}_k, \quad 1 \le k \le K, \end{aligned} \tag{38}$$

where the notations follow that in (36).

IV. Joint Receive Power Control and Receive Beamforming

In this section, variables transformation is first applied to derive an equivalent problem of (P2). Then, the method of SCA is adopted to address it and obtain the joint design of receive power control and receive beamforming.

A. An Equivalent Problem

In this part, to simplify (P2), the following variables are first defined:

$$\alpha_{\ell,\ell',m_i} = \frac{\left(\hat{\mu}_{\ell,m_i} - \hat{\mu}_{\ell',m_i}\right)^2}{\hat{\sigma}^2_{m_i}}, \quad \forall(\ell,\ell',m_i), \tag{39}$$

where α_{ℓ,ℓ',m_i} represents the per class pair discriminant gain of the m_i-th element. It follows that the problem can be equivalently derived as

$$\begin{aligned} \max_{\{c_k\},\mathbf{f}_1,\mathbf{f}_2,\left\{\alpha_{\ell,\ell',m_i}\right\}} \quad & G = \frac{2}{L(L-1)} \sum_{i=1}^{2} \sum_{\ell'=1}^{L} \sum_{\ell<\ell'} \alpha_{\ell,\ell',m_i}, \\ \text{s.t.} \quad & c_k^2 \leq P_k \mathbf{h}_k^H \left(\mathbf{f}_1\mathbf{f}_1^T + \mathbf{f}_2\mathbf{f}_2^T\right)\mathbf{h}_k, \quad 1 \leq k \leq K, \\ & \left(\hat{\mu}_{\ell,m_i} - \hat{\mu}_{\ell',m_i}\right)^2 = \alpha_{\ell,\ell',m_i}\hat{\sigma}^2_{m_i}, \quad \forall(\ell,\ell',m_i), \end{aligned} \tag{40}$$

where

$$\left(\hat{\mu}_{\ell,m_i} - \hat{\mu}_{\ell',m_i}\right)^2 = \left(\sum_{k=1}^{K} c_k\right)^2 \left(\mu_{\ell,m_i} - \mu_{\ell',m_i}\right)^2, \forall(\ell,\ell',m_i), \tag{41}$$

and

$$\hat{\sigma}^2_{m_i} = \left[\sigma^2_{m_i}\left(\sum_{k=1}^{K} c_k\right)^2 + \sum_{k=1}^{K} c_k^2\epsilon^2 + \frac{\delta_0^2}{2}\left(\mathbf{f}_1^T\mathbf{f}_1 + \mathbf{f}_2^T\mathbf{f}_2\right)\right], \ \forall(\ell,\ell',m_i). \tag{42}$$

Then, it can be shown that using symmetric real and imiginary receive beamformers can achieve the optimal solution of the probelm in (40), as presented in the following lemma.

Lemma 3 (Symmetric Receive Beamformers)**.** *Symmetric real and imaginary receive beamformers, as in* (43)*, will not influence the optimality of the problem in* (40)*.*

$$\mathbf{f}_1 = \mathbf{f}_2 = \hat{\mathbf{f}}. \tag{43}$$

Proof: Please see Appendix C.

Besides, it can be further proved that extending the feasible region of the second constraint of the problem in (40), i.e., the equality constraint, has no influence on its optimal solution, as equality should be achieved to obtain the optimum, as presented in Lemma 4.

Lemma 4 (Equivalent Extended Feasible Region)**.** *A problem, which extends the feasible region of the second constraint of the problem in* (40) *as*

$$\left(\sum_{k=1}^{K} c_k\right)^2 \left[\frac{\left(\mu_{\ell,m_i} - \mu_{\ell',m_i}\right)^2}{\alpha_{\ell,\ell',m_i}} - \sigma^2_{m_i}\right] \geq \sum_{k=1}^{K} c_k^2\epsilon^2 + \delta_0^2\hat{\mathbf{f}}^T\hat{\mathbf{f}}, \quad \forall(\ell,\ell',m_i), \tag{44}$$

and keeps the other constraints and the objective function, achieves the same optimal solution to the problem in (40)*.*

Proof: Please see Appendix D.

Next, based on Lemmas 3 and 4, the problem in (40) can be equally derived as

$$\text{(P3)} \quad \begin{aligned} \max_{\{c_k\},\hat{\mathbf{f}},\{\alpha_{\ell,\ell',m_i}\}} \quad & G = \frac{2}{L(L-1)} \sum_{i=1}^{2} \sum_{\ell'=1}^{L} \sum_{\ell<\ell'} \alpha_{\ell,\ell',m_i}, \\ \text{s.t.} \quad & c_k^2 \le P_k 2\mathbf{h}_k^H \hat{\mathbf{f}}\hat{\mathbf{f}}^T \mathbf{h}_k, \quad 1 \le k \le K, \\ & \left(\sum_{k=1}^{K} c_k\right)^2 \left[\frac{\left(\mu_{\ell,m_i} - \mu_{\ell',m_i}\right)^2}{\alpha_{\ell,\ell',m_i}} - \sigma_{m_i}^2\right] \ge \sum_{k=1}^{K} c_k^2 \epsilon^2 + \delta_0^2 \hat{\mathbf{f}}^T \hat{\mathbf{f}}, \ \forall(\ell,\ell',m_i). \end{aligned}$$

(P3) is non-convex due to the two constraints therein. In the sequel, the SCA method is used to tackle this problem.

B. SCA Based Joint Receive Power Control and Receive Beamforming

In this part, the SCA approach is used to address (P3) for obtaining a sub-optimal solution by iterating over the following two steps:

- *Convex relaxation*: Based on a feasible reference point, Problem (P3) is relaxed into a convex problem, whose feasible region is a subset of that of Problem (P3). Hence, the solution of relaxed problem is guaranteed to be feasible for Problem (P3).
- *Reference point updating*: The solution of the relaxed problem is used as the new reference point.

This process iterates till convergence and the final result can be guaranteed to satisfy the KKT conditions of (P3). In the sequel, the approach of convex relaxation is first presented, followed by the summary of the overall joint receive power control and receive beamforming algorithm.

1) Convex Relaxation of Problem (P3): First, for notation simplification, denote

$$\begin{cases} R_k(\hat{\mathbf{f}}) = 2P_k \mathbf{h}_k^H \hat{\mathbf{f}}\hat{\mathbf{f}}^T \mathbf{h}_k, \quad q \le k \le K, \\ Q_{\ell,\ell',m_i}\left(\{c_k\}, \alpha_{\ell,\ell',m_i}\right) = \left(\sum_{k=1}^{K} c_k\right)^2 \times \dfrac{\left(\mu_{\ell,m_i} - \mu_{\ell',m_i}\right)^2}{\alpha_{\ell,\ell',m_i}}, \quad \forall(\ell,\ell',m_i). \end{cases} \tag{45}$$

Then, consider an arbitrary iteration $(t+1)$, the reference point is denoted as $\{\hat{\mathbf{f}}^{[t]}, c_k^{[t]}, \alpha_{\ell,\ell',m_i}^{[t]}\}$. It follows that the following lemma can be derived.

Lemma 5. *$R_k(\hat{\mathbf{f}})$ and $Q_{\ell,\ell',m_i}\left(\{c_k\},\alpha_{\ell,\ell',m_i}\right)$ are differntiable and convex, and hence are no less than their corresponding first-order Taylor expansion at the point $\{\hat{\mathbf{f}}^{[t]},c_k^{[t]},\alpha_{\ell,\ell',m_i}^{[t]}\}$, i.e.,*

$$\begin{cases} R_k\left(\hat{\mathbf{f}}\right) \geq \hat{R}_k^{[t]}\left(\hat{\mathbf{f}}\right), \\ Q_{\ell,\ell',m_i}\left(\{c_k\},\alpha_{\ell,\ell',m_i}\right) \geq \hat{Q}_{\ell,\ell',m_i}^{[t]}\left(\{c_k\},\alpha_{\ell,\ell',m_i}\right), \quad \forall(\ell,\ell',m_i). \end{cases} \tag{46}$$

In the equation above, $\hat{R}_k^{[t]}\left(\hat{\mathbf{f}}\right)$ and $\hat{Q}_{\ell,\ell',m_i}^{[t]}\left(\{c_k\},\alpha_{\ell,\ell',m_i}\right)$ are the corresponding first-order linear expansion functions, given by

$$\hat{R}_k^{[t]}\left(\hat{\mathbf{f}}\right) = R(\hat{\mathbf{f}}^{[t]}) + 4P_k(\hat{\mathbf{f}}-\hat{\mathbf{f}}^{[t]})^H\left(\mathbf{h}_k^H\hat{\mathbf{f}}^{[t]}\mathbf{h}_k\right), \quad 1\leq k\leq K, \tag{47}$$

and

$$\begin{aligned} \hat{Q}_{\ell,\ell',m_i}^{[t]}\left(\{c_k\},\alpha_{\ell,\ell',m_i}\right) =& Q\left(\{c_k^{[t]}\},\alpha_{\ell,\ell',m_i}^{[t]}\right) + \sum_{k=1}^{K} A_k^{[t]}\left(c_k-c_k^{[t]}\right) \\ &+ B_{\ell,\ell',m_i}^{[t]}(\alpha_{\ell,\ell',m_i}-\alpha_{\ell,\ell',m_i}^{[t]}). \end{aligned} \tag{48}$$

where

$$\begin{cases} A_k^{[t]} = \left.\dfrac{\partial Q}{\partial c_k}\right|_{c_k=c_k^{[t]}} = \dfrac{2\sum\limits_{k=1}^{K} c_k^{[t]}\left(\mu_{\ell,m_i}-\mu_{\ell',m_i}\right)^2}{\alpha_{\ell,\ell',m_i}^{[t]}}, \\ B_{\ell,\ell',m_i}^{[t]} = \left.\dfrac{\partial Q}{\partial \alpha_{\ell,\ell',m_i}}\right|_{\alpha_{\ell,\ell',m_i}=\alpha_{\ell,\ell',m_i}^{[t]}} = -\left[\dfrac{\left(\sum_{k=1}^{K} c_k^{[t]}\right)\left(\mu_{\ell,m_i}-\mu_{\ell',m_i}\right)}{\alpha_{\ell,\ell',m_i}^{[t]}}\right]^2. \end{cases} \tag{49}$$

Proof: Please see Appendix E.

Next, by substituting the inequalities in (45) into Problem (P3), it can be derived as

$$\begin{aligned} \max_{\{c_k\},\hat{\mathbf{f}},\{\alpha_{\ell,\ell',m_i}\}} \quad & G = \frac{2}{L(L-1)}\sum_{i=1}^{2}\sum_{\ell'=1}^{L}\sum_{\ell<\ell'}\alpha_{\ell,\ell',m_i}, \\ \text{(P4)} \qquad \text{s.t.} \quad & c_k^2 \leq \hat{R}_k^{[t]}\left(\hat{\mathbf{f}}\right), \quad 1\leq k\leq K, \\ & \hat{Q}_{\ell,\ell',m_i}^{[t]}\left(\{c_k\},\alpha_{\ell,\ell',m_i}\right) - \left(\sum_{k=1}^{K} c_k\right)^2\sigma_{m_i}^2 \geq \sum_{k=1}^{K} c_k^2\epsilon^2 + \delta_0^2\hat{\mathbf{f}}^T\hat{\mathbf{f}}, \ \forall(\ell,\ell',m_i), \end{aligned}$$

where $\hat{R}_k^{[t]}\left(\hat{\mathbf{f}}\right)$ and $\hat{Q}_{\ell,\ell',m_i}^{[t]}\left(\{c_k\},\alpha_{\ell,\ell',m_i}\right)$ are defined in (47) and (48), respectively. (P4) is convex. The proof is straightforward and hence omitted. To address (P4), the well-known cvx toolbox can be used [52].

2) Algorithm of Joint Receive Power Control and Receive Beamforming: Based on the convex relaxation approach above, (P3) can be addressed by using the SCA method, which iteratively solves the relaxed convex problem (P4), and updates the reference point using the obtained solution. The detailed procedure is summarized in Algorithm 1.

Algorithm 1: Joint Tansmit Power Control and Receive Beamforming

1: **Input:** Channel gains $\{\mathbf{h}_k\}$.

2: **Initialize** $t=0$ and $\{\hat{\mathbf{f}}^{[0]}, c_k^{[0]}, \alpha_{\ell,\ell',m_i}^{[0]}\}$, which is in the feasible region of (P3).

3: **Loop**

4: $t=t+1$.

5: Derive Problem (P4), based on the reference point $\{\hat{\mathbf{f}}^{[t-1]}, c_k^{[t-1]}, \alpha_{\ell,\ell',m_i}^{[t-1]}\}$.

5: Solve Problem (P4) and obtain the optimum as $\{\hat{\mathbf{f}}^{[t]}, c_k^{[t]}, \alpha_{\ell,\ell',m_i}^{[t]}\}$.

6: **Until Convergence**

7: The solution is

$$\begin{cases} \hat{\mathbf{f}}^* = \hat{\mathbf{f}}^{[t]}, \\ c_k^* = c_k^{[t]}, \quad 1 \le k \le K, \\ \alpha_{\ell,\ell',m_i}^* = \alpha_{\ell,\ell',m_i}^{[t]}, \quad \forall(\ell,\ell',m_i). \end{cases}$$

8: **Output**: $\hat{\mathbf{f}}^*$, $\{c_k^*\}$, and $\{\alpha_{\ell,\ell',m_i}^*\}$.

3) Receive Beamformer and Transmit Precoders: Based on Algorithm 1 and the ZF precoding design in (28), the receive beamformer and transmit precoders of the priginal problem, say (P1), can be derived as

$$\boxed{\begin{aligned} &\mathbf{f}^* = \hat{\mathbf{f}}^* + j\hat{\mathbf{f}}^*, \\ &b_k^* = \frac{c_k^* \mathbf{h}_k^H \mathbf{f}^*}{\mathbf{h}_k^H \mathbf{f}^* \mathbf{f}^{*H} \mathbf{h}_k}, \quad 1 \le k \le K. \end{aligned}} \tag{50}$$

V. Performance Evaluation

A. Experiment Setup

1) Communication model: In this experiment, a multi-user single-input multiple-output network is taken into consideration. K devices are distributed randomly within a circle with a radius of 50 meters in the network. The AP is located 450 meters from the circle's center. The channel gain H_k is modeled as $H_k = |\varphi_k h_k|^2$, where φ_k and h_k stand for the large-scale and

TABLE I

SIMULATION PARAMETERS

Parameter	**Value**	**Parameter**	**Value**
Number of ISAC devices, K	3	Channel noise variance, δ_0^2	1
feature noise variance, ϵ^2	0.4	Number of receive antennas, N_r	8
Number of dimension after PCA, N_K	12	Number of classes, L	4
Training data sizes, B	6400	Transmit power, P_k	12 mdB
Variance of shadow fading, σ_ζ^2	8 dB	Communication channel noise power, δ_c^2	10^{-11} W

small-scale fading propagation coefficient, respectively. The large-scale propagation coefficient in dB from device k to the edge server is modeled as $[\varphi_k]_{\text{dB}} = -[\text{PL}_k]_{\text{dB}} + [\zeta_k]_{\text{dB}}$, where $[\text{PL}_k]_{\text{dB}} = 128.1 + 37.6 \log_{10} \text{dist}_k$ (dist_k is the distance in kilometer) is the path loss in dB, and $[\zeta_k]_{\text{dB}}$ accounts for the shadowing in dB. In the simulation, $[\zeta_k]_{\text{dB}}$ is Gauss-distributed random variable with mean zero and variance σ_ζ^2. It is assumed that the small-scale fading is Rayleigh fading, i.e., $h_k \sim \mathcal{CN}(0, 1)$.

2) Inference task: The wireless sensing simulator proposed in [51] is adopted to produce datasets for diverse high-fidelity human motions. Identification of four distinct human motions, i.e., *child walking*, *child pacing*, *adult walking*, and *adult pacing* is required for the inference task. The heights of children and adults are assumed to be uniformly distributed in interval $[0.9\text{m}, 1.2\text{m}]$ and $[1.6\text{m}, 1.9\text{m}]$, respectively, similar to the setting in [53]. Standing, walking, and pacing have respective speeds of 0 m/s, $0.5H$ m/s, and $0.25H$ m/s, where H is the height value. The heading of the moving human is set to be uniformly distributed in $[-180°, 180°]$.

3) Inference model: SVM and MLP neural networks are two machine learning models that are taken into account for inference in the experiments, respectively. The neural network model consists of two hidden layers, each with 80 and 40 neurons. On a total of 6400 data samples, both models are trained without the use of channel or data noise. In addition, the neural network model underwent 160 iterations of training. There are 1600 data samples used in the inference studies for testing accuracy.

Unless specified otherwise, other simulation parameters are stated in Table I. All experiments are implemented using Python 3.8 on a Linux server with one NVIDIA® GeForce® RTX 3090 GPU 24GB and one Intel® Xeon® Gold 5218 CPU.

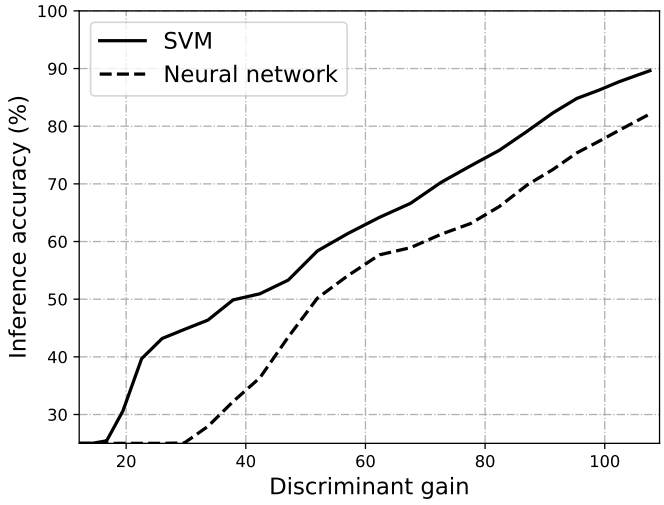


Fig. 5. Inference accuracy versus discriminant gain.

B. Inference Algorithms

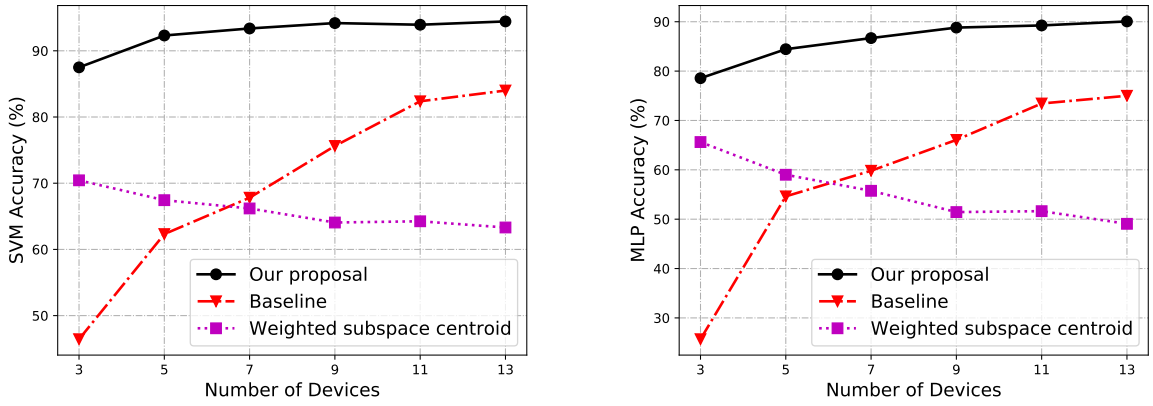
For comparison, we consider three schemes as follows.

- *Baseline*: This is a heuristic algorithm. The receive beamforming is firstly allocated randomly, then the other parameters are allocated randomly under the constrains of receive beamforming.
- *Weighted subspace centroid*: All the parameters are allocated following the AirComp scheme in [37], whose design criterion is MMSE.
- *Joint design of transmit precoding and receive beamforming (our proposal)*: All parameters are set follow the proposed scheme in (50).

C. Experimental Results

The section begins by outlining the relationships between the two model's discriminant gains and inference accuracy. The SVM and the neural network are then used to compare the three algorithms. To be more precise, we first investigate how the number of participating devices affects the inference accuracy. Next, we concentrate on the impact of transmit power. We finally delve deeply into the connection between the PCA dimension and inference accuracy.

1) Inference accuracy v.s. discriminant gain: In Fig. 5, the relations between inference accuracy and discriminant gain for the SVM model and the MLP neural network are presented. To investigate the relation, different values of discrminant gain are obtained by using different



(a) Inference accuracy with SVM versus number of devices (b) Inference accuracy with MLP versus number of devices

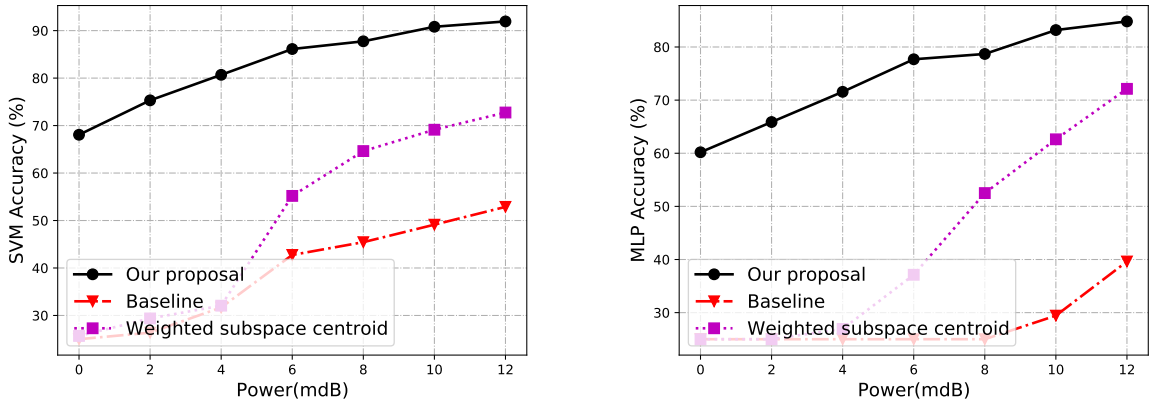
Fig. 6. Inference accuracy comparison among different models under differenct number of devices.

transmit power on devices. From the figure, for both models, it is seen that the inference accuracy increases as the discriminant gain grows. Additionally, the SVM beats the neural network, as the training of the latter is overfitting, which has a complex model compared to the simple dataset.

2) Inference accuracy v.s. number of devices: The inference accuracy of both models is shown in Fig. 6 in terms of various number of devices. it is observed that our proposed scheme has the best performance. Besides, the performance of the weighted subspace centroid scheme decreases with the number of devices. The reason is as follows. Although more data diversity is provided, channel equalization is performed among all devices under the target of MMSE in this scheme while the possibility of deep fading channels increases with increasing number of devices. As a result, the distortion level increases. However, better inference accuracy is obtained in the baseline scheme and our proposed scheme, as the number of devices increases. This is because under the task-oriented principle, different receive powers are enabled for different devices and the data diversity can be fully exploited.

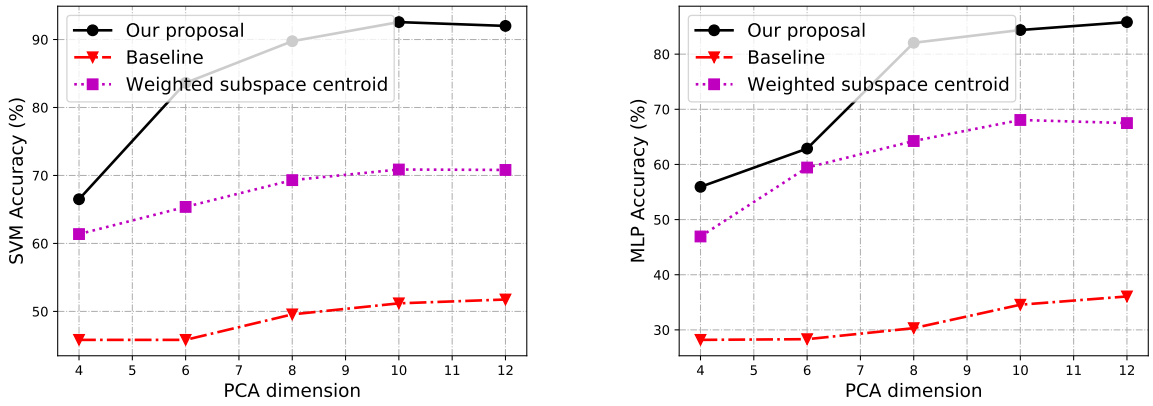
3) Inference accuracy v.s. transmit power: The inference accuracy of both models under various transmit power conditions is shown in Fig. 7. In both cases, improved inference accuracy is acquired as the transmit power rises. This is because more transmit power can suppresses the channel noise so that the discriminant gain is enhanced. As well, our proposed scheme outperforms the other two schemes.

4) Inference accuracy v.s. number of feature elements: The inference accuracy of both models is shown in Fig. .8 in terms of the number of used feature elements extracted using PCA.



(a) Inference accuracy with SVM versus transmit power (b) Inference accuracy with MLP versus transmit power

Fig. 7. Inference accuracy comparison among different models under different transmit power.



(a) Inference accuracy with SVM versus PCA dimension (b) Inference accuracy with MLP versus transmit power

Fig. 8. Inference accuracy comparison among different models under differenct PCA dimension.

Specifically, the number of used feature elements is sequentially increased following the order of the PCA dimensions from the largest to the least. From the figure, the inference accuracy increases as the number of used feature elements in the inference task. That's because more feature elements increase the dimensions of feature space so that different classes can be better differentiated and a better discriminant gain can be achieved. In addition, the accuracy turns to be saturate at a large number of feature elements, since least important feature elements are added.

The extensive experimental results presented above demonstrate the best performance of the proposed optimal scheme and verify our theoretical analysis.

VI. Conclusion

To enhance the performance of mutlti-device edge inference systems, this paper proposed a task-oriented AirComp scheme. It aggregated the noise-corrupted local feature vectors for suppressing the sensing noise. Following the task-oriented principle instead of using the conventional criterion MMSE, the inference accruacy measured by discriminant gain was maximized. This task-oriented problem, however, is non-convex due to the complicated form of discriminant gain as well as the couple of transmit pre-coding and receive beamforming. To taclke this problem, variables transformation was first applied to derive an equivalent problem with convex objective function and feasible sub-regions. Then, a joint scheme of transmit pre-coding and receive beamforming was proposed based on the SCA approach. The performance of the proposed scheme was verified using extensive numerical results of a human motion recognition task.

This work opens several interesting directions. One is the device selection for further accuracy enhancement by not accessing the devices with weak channel or high sensing noise. Another is to extend the current design to the case where devices are equipped with multi antennas.

Appendix

A. *Proof of Lemma 1*

First, according to (31), x_{m_i} can be decomposed into the avegrage of L indepdent Gaussian variables, as

$$x_{m_i} = \frac{1}{L}\sum_{\ell=1}^{L} x_{\ell,m_i}, \quad i = 1, 2, \tag{51}$$

where the distribution of $x_{m_i,\ell}$ is

$$x_{\ell,m_i} \sim \mathcal{N}\left(\mu_{\ell,m_i}, \sigma_{m_i}^2\right), \quad 1 \leq \ell \leq L, \ \& \ i = 1, 2. \tag{52}$$

Then, by substituting the above equation into the local elements defined in (32), we have

$$x_{k,m_i} = \frac{1}{L}\sum_{\ell=1}^{L} x_{\ell,m_i} + d_{k,m_i}, \quad 1 \leq k \leq K, \ \& \ i = 1, 2. \tag{53}$$

It follows that

$$x_{k,m_i} = \frac{1}{L}\sum_{\ell=1}^{L} x_{\ell,k,m_i}, \quad 1 \leq k \leq K, \ \& \ i = 1, 2. \tag{54}$$

where

$$x_{\ell,k,m_i} = x_{\ell,m_i} + d_{k,m_i}, \quad 1 \leq k \leq K, \ \& \ \leq \ell \leq L, \ \& \ i = 1, 2. \tag{55}$$

Next, by substituting the distributions of x_{ℓ,m_i} in (52) and the distribution of d_{k,m_i} in (33), the distribution of x_{ℓ,k,m_i} can be derived as

$$x_{\ell,k,m_i} \sim \mathcal{N}\left(\mu_{\ell,m_i}, \sigma_{m_i}^2 + \epsilon^2\right), \quad 1 \leq k \leq K, \ \& \ \leq \ell \leq L, \ \& \ i = 1, 2. \tag{56}$$

It follows that the distribution of x_{k,m_i} can be derived as

$$x_{k,m_i} \sim \frac{1}{L} \sum_{\ell=1}^{L} \mathcal{N}\left(\mu_{\ell,m_i}, \sigma_{m_i}^2 + \epsilon^2\right), \quad i = 1, 2, \ \& \ 1 \leq k \leq K. \tag{57}$$

This ends the proof.

B. Proof of Lemma 2

First, for the received symbol $\hat{s}$ in (20), the received noise can be derived as

$$\begin{aligned} \mathbf{f}^H \mathbf{n} &= (\mathbf{f}_1 + j\mathbf{f}_2)^H (\mathbf{n}_1 + j\mathbf{n}_2), \\ &= \mathbf{f}_1^T \mathbf{n}_1 + \mathbf{f}_2^T \mathbf{n}_2 + j(\mathbf{f}_1^T \mathbf{n}_2 - \mathbf{f}_2^T \mathbf{n}_1), \end{aligned} \tag{58}$$

where $\mathbf{f}_1$ and $\mathbf{f}_2$ are the real part and imiginary part of $\mathbf{f}$ respectively, and $\mathbf{n}_1$ and $\mathbf{n}_2$ are the real part and imiginary part of the Guassian noise $\mathbf{n}$ respectively. Specifically, we have

$$\begin{cases} \mathbf{n}_1 \sim \mathcal{N}\left(\mathbf{0}, \frac{\delta_0^2}{2}\mathbf{I}\right), \\ \mathbf{n}_2 \sim \mathcal{N}\left(\mathbf{0}, \frac{\delta_0^2}{2}\mathbf{I}\right) \end{cases} \tag{59}$$

where δ_0^2 is the noise variance. Then, for the real part of the received noise, its expectation can co-variance can be derived as

$$\mathbb{E}\left[\mathsf{Re}(\mathbf{f}^H \mathbf{n})\right] = \mathbb{E}\left[\mathbf{f}_1^T \mathbf{n}_1 + \mathbf{f}_2^T \mathbf{n}_2\right] = \mathbf{0}, \tag{60}$$

and

$$\begin{aligned} \mathbb{C} = \left[\mathsf{Re}(\mathbf{f}^H \mathbf{n})\right] &= \mathbb{E}\left[\left(\mathbf{f}_1^T \mathbf{n}_1 + \mathbf{f}_2^T \mathbf{n}_2\right)\left(\mathbf{f}_1^T \mathbf{n}_1 + \mathbf{f}_2^T \mathbf{n}_2\right)^T\right], \\ &= \frac{\delta_0^2}{2}\left(\mathbf{f}_1^T \mathbf{f}_1 + \mathbf{f}_2^T \mathbf{f}_2\right). \end{aligned} \tag{61}$$

That's to say,

$$\mathsf{Re}(\mathbf{f}^H \mathbf{n}) \sim \mathcal{N}\left(\mathbf{0}, \frac{\delta_0^2}{2}\left(\mathbf{f}_1^T \mathbf{f}_1 + \mathbf{f}_2^T \mathbf{f}_2\right)\right). \tag{62}$$

Similarly, it can be derived that the imiginary part of the received noise has the same distribution:

$$\mathsf{Im}(\mathbf{f}^H \mathbf{n}) \sim \mathcal{N}\left(\mathbf{0}, \frac{\delta_0^2}{2}\left(\mathbf{f}_1^T \mathbf{f}_1 + \mathbf{f}_2^T \mathbf{f}_2\right)\right). \tag{63}$$

By subsituting the noise distributions in (62) and (63) into the global estimates in (21) and using the similar method in Appendix B, i.e., decompose the local estimate x_{k,m_i} into the average of L independent Gaussia variables, the distributions of the global estimates can be derived as in (35). This ends the proof.

C. Proof of Lemma 3

The Lagrange function of the problem in (40) can be written as

$$\begin{aligned}\mathcal{L} = &- \frac{2}{L(L-1)} \sum_{i=1}^{2} \sum_{\ell'=1}^{L} \sum_{\ell<\ell'} \alpha_{\ell,\ell',m_i} + \sum_{k=1}^{K} \beta_k \left[c_k^2 - P_k \mathbf{h}_k^H \left(\mathbf{f}_1 \mathbf{f}_1^T + \mathbf{f}_2 \mathbf{f}_2^T \right) \mathbf{h}_k \right] \\ &+ \sum_{i=1}^{2} \sum_{\ell'=1}^{L} \sum_{\ell<\ell'} \lambda_{\ell,\ell',m_i} \left[\alpha_{\ell,\ell',m_i} \hat{\sigma}_{m_i}^2 - \left(\hat{\mu}_{\ell,m_i} - \hat{\mu}_{\ell',m_i} \right)^2 \right],\end{aligned} \tag{64}$$

where $\left(\hat{\mu}_{\ell,m_i} - \hat{\mu}_{\ell',m_i} \right)^2$ is defined in (41) and $\hat{\sigma}_{m_i}^2$ is defined in (42). KKT conditions are necessary to achieve the optimal solution. Some useful KKT conditions are given below.

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \mathbf{f}_1} &= -2 \sum_{k=1}^{K} \beta_k P_k \mathbf{h}_k^H \mathbf{h}_k \mathbf{f}_1 + \sum_{i=1}^{2} \sum_{\ell'=1}^{L} \sum_{\ell<\ell'} \lambda_{\ell,\ell',m_i} \alpha_{\ell,\ell',m_i} \delta_0^2 \mathbf{f}_1 = 0, \\ \frac{\partial \mathcal{L}}{\partial \mathbf{f}_2} &= -2 \sum_{k=1}^{K} \beta_k P_k \mathbf{h}_k^H \mathbf{h}_k \mathbf{f}_2 + \sum_{i=1}^{2} \sum_{\ell'=1}^{L} \sum_{\ell<\ell'} \lambda_{\ell,\ell',m_i} \alpha_{\ell,\ell',m_i} \delta_0^2 \mathbf{f}_2 = 0.\end{aligned} \tag{65}$$

It can be observed from the above equations that $\mathbf{f}_1 = \mathbf{f}_2$ won't influence the optimality of the problem. This ends the proof.

D. Proof of Lemma 4

First, the second constraint in (40) can be equally written as

$$\left(\hat{\mu}_{\ell,m_i} - \hat{\mu}_{\ell',m_i} \right)^2 \geq \alpha_{\ell,\ell',m_i} \hat{\sigma}_{m_i}^2, \quad \forall (\ell, \ell', m_i), \tag{66}$$

The reason is straightforward. In (66), if the equality is not achieved, the value of α_{ℓ,ℓ',m_i} can be increased to make the objective function in (40) larger. In otehr words, it's necessary to achieve equality for obtaining the optimal solution. Then, by substituting $\left(\hat{\mu}_{\ell,m_i} - \hat{\mu}_{\ell',m_i} \right)^2$ in (41) and $\hat{\sigma}_{m_i}^2$ in (42) into (66), it can be derived as

$$\left(\sum_{k=1}^{K} c_k \right)^2 \left(\mu_{\ell,m_i} - \mu_{\ell',m_i} \right)^2 \geq \alpha_{\ell,\ell',m_i} \left[\sigma_{m_i}^2 \left(\sum_{k=1}^{K} c_k \right)^2 + \sum_{k=1}^{K} c_k^2 \epsilon^2 + \delta_0^2 \hat{\mathbf{f}}^T \hat{\mathbf{f}} \right], \ \forall (\ell, \ell', m_i),$$

It follows that

$$\left(\sum_{k=1}^{K} c_k\right)^2 \left[\frac{\left(\mu_{\ell,m_i} - \mu_{\ell',m_i}\right)^2}{\alpha_{\ell,\ell',m_i}} - \sigma_{m_i}^2\right] \geq \sum_{k=1}^{K} c_k^2 \epsilon^2 + \delta_0^2 \hat{\mathbf{f}}^T \hat{\mathbf{f}}, \quad \forall(\ell, \ell', m_i). \tag{67}$$

This ends the proof.

E. Proof of Lemma 5

First, it can be observed that $P(\hat{\mathbf{f}})$ defined in (46) is a qudratic function, and thus is convex. Then, $Q\left(\{c_k\}, \alpha_{\ell,\ell',m_i}\right)$ can be linearly transformed from the following function,

$$f(x, y) = \frac{x^2}{y}, \ x > 0, \ y > 0, \tag{68}$$

whose Hessian matrix is

$$\begin{bmatrix} \frac{2}{y}, & -\frac{2x}{y^2} \\ -\frac{2x}{y^2}, & \frac{2x^3}{y^3} \end{bmatrix}. \tag{69}$$

Its eigenvalues are $\lambda_1 = 0$ and $\lambda_2 = 2(x^2 + y^2)/y^3 > 0$. Hence, $f(x, y)$ is convex. As linear transformation preserves convexity, $Q\left(\{c_k\}, \alpha_{\ell,\ell',m_i}\right)$ is convex. Furthermore, according to the property of convex functions, they are no less than the corresponding first-order Taylor expansion. This ends the proof.

REFERENCES

[1] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y.-J. A. Zhang, "The roadmap to 6g: Ai empowered wireless networks," *IEEE Communications Magazine*, vol. 57, no. 8, pp. 84–90, 2019.

[2] K. B. Letaief, Y. Shi, J. Lu, and J. Lu, "Edge artificial intelligence for 6G: Vision, enabling technologies, and applications," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 5–36, Jan. 2022.

[3] G. Zhu, D. Liu, Y. Du, C. You, J. Zhang, and K. Huang, "Toward an intelligent edge: Wireless communication meets machine learning," *IEEE Commun. Mag.*, vol. 58, no. 1, pp. 19–25, Jan. 2020.

[4] Y. Shi, K. Yang, T. Jiang, J. Zhang, and K. B. Letaief, "Communication-efficient edge AI: Algorithms and systems," *IEEE Commun. Surv. & Tut.*, vol. 22, no. 4, pp. 2167–2191, Fourth quarter 2020.

[5] J. Park, S. Samarakoon, M. Bennis, and M. Debbah, "Wireless network intelligence at the edge," *Proc. IEEE*, vol. 107, no. 11, pp. 2204–2239, Nov. 2019.

[6] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, "A joint learning and communications framework for federated learning over wireless networks," *IEEE Trans. Wireless Commun.*, 2020.

[7] D. Wen, M. Bennis, and K. Huang, "Joint parameter-and-bandwidth allocation for improving the efficiency of partitioned edge learning," *IEEE Trans. Wireless Commun.*, vol. 68, pp. 2128–2142, 2020.

[8] Z. Jiang, G. Yu, Y. Cai, and Y. Jiang, "Decentralized edge learning via unreliable device-to-device communications," *IEEE Transactions on Wireless Communications*, pp. 1–1, 2022.

[9] K. Yang, Y. Shi, W. Yu, and Z. Ding, "Energy-efficient processing and robust wireless cooperative transmission for edge inference," *IEEE Internet Things J.*, vol. 7, no. 10, pp. 9456–9470, Oct. 2020.

[10] M. Lee, G. Yu, and H. Dai, "Decentralized inference with graph neural networks in wireless communication systems," *IEEE Trans. Mobile Comput.*, early access, Nov. 2021.

[11] P. Liu, G. Zhu, W. Jiang, W. Luo, J. Xu, and S. Cui, "Vertical federated edge learning with distributed integrated sensing and communication," *IEEE Commun. Lett.*, early access, Jun. 2022.

[12] S. Liu, G. Yu, R. Yin, J. Yuan, L. Shen, and C. Liu, "Joint model pruning and device selection for communication-efficient federated edge learning," *IEEE Transactions on Communications*, vol. 70, no. 1, pp. 231–244, 2022.

[13] D. Wen, K.-J. Jeon, and K. Huang, "Federated dropout–A simple approach for enabling federated learning on resource constrained devices," *IEEE Wireless Commun. Lett.*, vol. 11, no. 5, pp. 923–927, 2022.

[14] C.-J. Wu, D. Brooks, K. Chen, D. Chen, S. Choudhury, M. Dukhan, K. Hazelwood, E. Isaac, Y. Jia, B. Jia, T. Leyvand, H. Lu, Y. Lu, L. Qiao, B. Reagen, J. Spisak, F. Sun, A. Tulloch, P. Vajda, X. Wang, Y. Wang, B. Wasti, Y. Wu, R. Xian, S. Yoo, and P. Zhang, "Machine learning at Facebook: Understanding inference at the edge," in *IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 2019, pp. 331–344.

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

[16] J. Soifer, J. Li, M. Li, J. Zhu, Y. Li, Y. He, E. Zheng, A. Oltean, M. Mosyak, C. Barnes, T. Liu, and J. Wang, "Deep learning inference service at Microsoft," in *2019 USENIX Conference on Operational Machine Learning (OpML 19)*. Santa Clara, CA: USENIX Association, May 2019, pp. 15–17. [Online]. Available: https://www.usenix.org/conference/opml19/presentation/soifer

[17] S. Jang, B. Kostadinov, and D. Lee, "Microservice-based edge device architecture for video analytics," in *2021 IEEE/ACM Symposium on Edge Computing (SEC)*. Los Alamitos, CA, USA: IEEE Computer Society, dec 2021, pp. 165–177. [Online]. Available: https://doi.ieeecomputersociety.org/10.1145/3453142.3491283

[18] E. C. Strinati and S. Barbarossa, "6G networks: Beyond shannon towards semantic and goal-oriented communications," *Comput. Netw.*, vol. 190, May 2021.

[19] D. Ma, N. Shlezinger, T. Huang, Y. Liu, and Y. C. Eldar, "FRaC: FMCW-based joint radar-communications system via index modulation," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 6, pp. 1348–1364, Nov. 2021.

[20] Q. Lan, D. Wen, Z. Zhang, Q. Zeng, X. Chen, P. Popovski, and K. Huang, "What is semantic communication? A view on conveying meaning in the era of machine intelligence," *J. Commun. Inf. Netw.*, vol. 6, no. 4, pp. 336–371, 2021.

[21] H. Xie, Z. Qin, X. Tao, and K. B. Letaief, "Task-oriented multi-user semantic communications," *[Online]. Available: https://arxiv.org/pdf/2112.10255.pdf*, 2021.

[22] M. Merluzzi, M. C. Filippou, L. G. Baltar, and E. C. Strinati, "Effective goal-oriented 6G communications: the energy-aware edge inferencing case," *[Online]. Available: https://arxiv.org/abs/2204.09447*, 2022.

[23] J. Shao and J. Zhang, "Communication-computation trade-off in resource-constrained edge inference," *IEEE Commun. Mag.*, vol. 58, no. 12, pp. 20–26, Jan. 2020.

[24] X. Huang and S. Zhou, "Dynamic compression ratio selection for edge inference systems with hard deadlines," *IEEE Internet Things J.*, vol. 7, no. 9, pp. 8800–8810, Sep. 2020.

[25] J. Shao, Y. Mao, and J. Zhang, "Learning task-oriented communication for edge inference: An information bottleneck approach," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 197–211, Jan. 2022.

[26] W. Shi, S. Zhou, and Z. Niu, "Device scheduling with fast convergence for wireless federated learning," in *2020 IEEE Int. Conf. Commun. (ICC)*. IEEE, 2020, pp. 1–6.

[27] Q. Lan, Q. Zeng, P. Popovski, D. Gündüz, and K. Huang, "Progressive feature transmission for split inference at the wireless edge," *[Online]. Available: https://arxiv.org/abs/2112.07244*, 2021.

[28] E. Li, L. Zeng, Z. Zhou, and X. Chen, "Edge AI: On-demand accelerating deep neural network inference via edge computing," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 447–457, 2019.

[29] Z. Liu, Q. Lan, and K. Huang, "Resource allocation for multiuser edge inference with batching and early exiting (extended version)," *[Online]. Available: https://arxiv.org/abs/2204.05223*, Apr. 2022.

[30] M. Jankowski, D. Gündüz, and K. Mikolajczyk, "Joint device-edge inference over wireless links with pruning," in *IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*. IEEE, May 2020, pp. 1–5.

[31] Z. Yi, R. Zhange, B. Xu, Y. Chen, L. Zhu, F. Li, G. Yang, and Y. Luo, "A wide-angle beam scanning antenna in e-plane for k-band radar sensor," *IEEE Access*, vol. 7, pp. 171 684–171 690, Nov. 2019.

[32] D. Wen, P. Liu, G. Zhu, Y. Shi, S. Cui, and Y. C. Eldar, "Task-oriented sensing, computation, and communication integration for multi-device edge ai," *[Online]. Available: https://arxiv.org/pdf/2207.00969.pdf*, Jul. 2022.

[33] Z.-Q. Luo, "Universal decentralized estimation in a bandwidth constrained sensor network," *IEEE Trans. Inf. Theory*, vol. 51, no. 6, pp. 2210–2219, Jun. 2005.

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

[35] G. Zhu, J. Xu, K. Huang, and S. Cui, "Over-the-air computing for wireless data aggregation in massive IoT," *IEEE Wireless Commun.*, vol. 28, no. 4, pp. 57–65, 2021.

[36] L. Chen, N. Zhao, Y. Chen, F. R. Yu, and G. Wei, "Over-the-air computation for iot networks: Computing multiple functions with antenna arrays," *IEEE Internet Things J.*, vol. 5, no. 6, pp. 5296–5306, 2018.

[37] G. Zhu and K. Huang, "MIMO over-the-air computation for high-mobility multimodal sensing," *IEEE Internet Things J.*, vol. 6, no. 4, pp. 6089–6103, Sep. 2019.

[38] X. Li, G. Zhu, Y. Gong, and K. Huang, "Wirelessly powered data aggregation for IoT via over-the-air function computation: Beamforming and power control," *IEEE Trans. Wireless Commun.*, vol. 18, no. 7, pp. 3437–3452, 2019.

[39] D. Wen, G. Zhu, and K. Huang, "Reduced-dimension design of mimo over-the-air computing for data aggregation in clustered iot networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5255–5268, 2019.

[40] X. Zhai, X. Chen, J. Xu, and D. W. Kwan Ng, "Hybrid beamforming for massive mimo over-the-air computation," *IEEE Trans. Commun.*, vol. 69, no. 4, pp. 2737–2751, 2021.

[41] X. Cao, G. Zhu, J. Xu, and K. Huang, "Optimized power control for over-the-air computation in fading channels," *IEEE Trans. on Wireless Commun.*, vol. 19, no. 11, pp. 7498–7513, 2020.

[42] W. Liu, X. Zang, Y. Li, and B. Vucetic, "Over-the-air computation systems: Optimization, analysis and scaling laws," *IEEE Trans. Wireless Commun.*, vol. 19, no. 8, pp. 5488–5502, 2020.

[43] M. Fu, Y. Zhou, Y. Shi, T. Wang, and W. Chen, "UAV-assisted over-the-air computation," in *IEEE Int. Conf. Commun.*, early access, Jun. 2021, pp. 1–6.

[44] G. Zhu, Y. Wang, and K. Huang, "Broadband analog aggregation for low-latency federated edge learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 491–506, 2020.

[45] K. Yang, T. Jiang, Y. Shi, and Z. Ding, "Federated learning via over-the-air computation," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 2022–2035, 2020.

[46] M. M. Amiri and D. Gündüz, "Federated learning over wireless fading channels," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3546–3557, 2020.

[47] Y. Sun, S. Zhou, Z. Niu, and D. Gündüz, "Dynamic scheduling for over-the-air federated edge learning with energy constraints," *IEEE J. Sel. Areas in Commun.*, vol. 40, no. 1, pp. 227–242, 2022.

[48] D. Liu and O. Simeone, "Privacy for free: Wireless federated learning via uncoded transmission with adaptive power control," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 170–185, 2021.

[49] S. F. Yilmaz, B. Hasircioglu, and D. Gunduz, "Over-the-air ensemble inference with model privacy," *[Online]. Available: https://arxiv.org/pdf/2202.03129.pdf*, May 2022.

[50] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[51] G. Li, S. Wang, J. Li, R. Wang, X. Peng, and T. X. Han, "Wireless sensing with deep spectrogram network and primitive based autoregressive hybrid channel model," in *IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Sep. 2021, pp. 481–485.

[52] M. Grant, S. Boyd, and Y. Ye, "Cvx users' guide," *[Online]. Available: http://www. stanford. edu/boyd/software. html*, 2009.

[53] MathWorks, "Pedestrian and bicyclist classification using deep learning," *[Online]. Available: https://ww2.mathworks.cn/help/radar/ug/pedestrian-and-bicyclist-classification-using-deep-learning.html*, 2022.