

# Distributed Over-the-air Computing for Fast Distributed Optimization: Beamforming Design and Convergence Analysis

Zhenyi Lin, Yi Gong, and Kaibin Huang

## Abstract

Distributed optimization concerns the optimization of a common function in a distributed network, which finds a wide range of applications ranging from machine learning to vehicle platooning. Its key operation is to aggregate all *local state information* (LSI) at devices to update their states. The required extensive message exchange and many iterations cause a communication bottleneck when the LSI is high dimensional or at high mobility. To overcome the bottleneck, we propose in this work the framework of distributed *over-the-air computing* (AirComp) to realize a one-step aggregation for distributed optimization by exploiting simultaneous multicast beamforming of all devices and the property of analog waveform superposition of a multi-access channel. Equivalently, the technique superimposes multiple instances of conventional AirComp processes, giving rise to the challenge of jointly designing multicast beamforming at devices to rein in errors due to interference and channel distortion. We consider two design criteria. The first one is to minimize the sum AirComp error (i.e., sum mean-squared error (MSE)) with respect to the desired average-functional values. An efficient solution approach is proposed by transforming the non-convex beamforming problem into an equivalent concave-convex fractional program and solving it by nesting convex programming into a bisection search. The second criterion, called zero-forcing (ZF) multicast beamforming, is to force the received over-the-air aggregated signals at devices to be equal to the desired functional values. In this case, the optimal beamforming admits closed form. Both the MMSE and ZF beamforming exhibit a *centroid* structure resulting from averaging columns of conventional MMSE/ZF precoding. Last, the convergence of a classic distributed optimization algorithm is analyzed. The distributed AirComp is found to accelerate convergence by dramatically reducing communication latency. Another key finding is that the ZF beamforming outperforms the MMSE design as the latter is shown to cause bias in subgradient estimation.

Z. Lin and K. Huang are with the Dept. of Electrical and Electronic Engr. at The University of Hong Kong, Hong Kong. Z. Lin is also with the Dept. of EEE at Southern University of Science and Technology, China. Y. Gong is with the same department. Contact: K. Huang (Email: huangk@eee.hku.hk), Y. Gong (Email: gongy@sustech.edu.cn).

## I. INTRODUCTION

Distributed optimization concerns the optimization of a common function in a distributed network comprising a cluster of edge devices connected by *device-to-device* (D2D) links [1]. The broad field covers two popular areas: *distributed machine learning* aiming at leveraging local data and computer resources at edge devices to train a global artificial-intelligence (AI) model [2], and *distributed consensus* finding a wide range of applications in, for example, swarms of drones or robots and vehicle platooning [3]. A fundamental operation in distributed optimization, which is usually implemented using an iterative gradient-descent algorithm, is to aggregate all *local state information* (LSI) and use the result to update the states of all devices in each round. The required extensive message exchange and many rounds cause a communication bottleneck when the LSI is high dimensional (e.g., distributed learning) or at high mobility (e.g., drones). To overcome the bottleneck, we propose in this work the framework of distributed *over-the-air computing* (AirComp) to realize a *one-step* aggregation for distributed optimization by exploiting simultaneous multicast beamforming by all devices and the property of analog waveform superposition of a multi-access channel. To develop the framework, beamforming designs to rein in channel distortion are presented and the convergence of distributed optimization adopting distributed AirComp is studied.

Recent years have witnessed fast-growing interests in distributed machine learning as driven by the distillation of enormous mobile data into artificial intelligence (AI) to power next-generation applications ranging from industrial automation to smart cities and extended reality. Relevant research mainly focuses on the efficient deployment of federated learning, arguably the most popular distributed learning framework, at the network edge, giving rise to the fast growing area of *federated edge learning* (FEEL) [4]. FEEL gains popularity for its capabilities of leveraging local data while helping to preserve their privacy and distributed computation resources. Typically implemented in networks with a star topology, the FEEL algorithm iterates between 1) aggregation at an edge server over local models, which are updated by edge devices using local data and uploaded over wireless channels, to update a global model, and 2) broadcast of the model to all devices for updating, until the global model converges. The main challenge confronting efficient FEEL is the communication bottleneck caused by uploading high-dimensional local modes (or stochastic-gradients) from potentially many devices over a multi-access channel. The required large number of rounds/iterations (e.g., tens to hundreds of rounds)

exacerbates the issue. Attempts to overcome the bottleneck have led to the proposal of different techniques including radio resource management [5], [6], model quantization [7], [8], and device scheduling [9], [10]. One particular class of techniques of our interest, known as *over-the-air FEEL*, features the application of AirComp to realize over-the-air model aggregation in FEEL [11]–[14]. The principle underpinning AirComp (as well as over-the-air FEEL) is to exploit the waveform superposition property such that the signal received at the server approximates a desired aggregation function of linear analog modulated data (e.g., local models/gradients) simultaneously transmitted by devices [12], [15].

Most recently, researchers have studied distributed FEEL targeting a cluster of devices without coordination by a sever and connected by D2D links [16]–[18]. The original techniques for server-assisted FEEL can be adopted by arranging the devices to take turn or use orthogonal channels to play the role of edge server [17], [19]. As the resultant sequential aggregation is time consuming, attempts on realizing parallel operations have been made by selecting multiple weakly coupled clusters of devices to perform simultaneous intra-cluster aggregation [16], [18]. Such approaches are ineffective for a single cluster of devices with tightly coupled links such as a drone swarm or a vehicle platoon, and still require multiple time slots to complete aggregation over all devices. Fundamentally, the drawback of the existing approaches is rooted in attempting to orthogonalize multiple aggregation processes in distributed optimization. On the contrary, we advocate fusing them into a single multi-aggregation process to enable a *one-step* distributed aggregation over all devices. Thereby, the resultant design, termed *distributed AirComp*, supports fast distributed FEEL and distributed optimization at large.

Consider the aggregation operation of distributed optimization among  $K$  devices. The goal of designing distributed AirComp is to realize *one-step* updating of the states of all devices via over-the-air aggregation over their LSI. To this end, all devices simultaneously multicast LSI to their peers and at the same time receive aggregated LSI via full-duplex communication [20]. We propose provisioning devices with transmit antenna arrays to enable multicast beamforming for reining in the sum AirComp error. On the other hand, the use of receive arrays can support aggregation beamforming as studied in [15]. To simplify our design, we assume single receive antenna at devices. The extension of the current design to include aggregation beamforming is straightforward but make the design tedious without new insight. Usually targeting a single-cell system, traditional multicast beamforming at the base station aims to efficiently deliver information to multiple receivers under their quality-of-service requirements. Mathematically,

the beamforming design can be formulated as an optimization problem with the objective of minimizing the required transmission resources at the base station (e.g., power and array size) under the constraints of receive signal-to-noise ratios (SNRs) [21], [22]. Another popular “max-min” formulation targets maximizing the lowest rate or receive SNR among the receivers under a transmission power constraint [23], [24]. Consider the context of distributed AirComp. Define the AirComp error as the deviation of an aggregated signal from the desired average-functional value due to channel fading and noise. Then the multicast beamforming from the perspective of an individual device aims to minimize the sum AirComp error over other devices under a transmission power constraint. Solving such a problem is no more difficult than the mentioned traditional ones. However, the new challenge arises from  $K$  simultaneous over-the-air aggregation processes coupling multicast beamforming at  $K$  devices. This necessitates the joint beamforming design that should account for all D2D-channel states in the system. To be precise, the multicast radiation patterns of  $K$  devices should be jointly optimized under individual power constraints such that their over-the-air superposition leads to the minimization of sum AirComp error. AirComp requires the aggregated signal arriving at a device to be scaled by a factor, called *alignment level*, to balance approaching a desired functional value and noise amplification [25]. The said problem is further complicated by the need of optimization over the alignment level. In contrast, transmit beamforming for single-aggregation AirComp is simple as its main purpose is to overcome the fading of an associated single-user channel, and hence is irrelevant to the current problem [15], [26].

In this work, besides proposing the principle of distributed AirComp as described earlier, we design distributed multicast beamforming to materialize the principle and further apply the design to distributed optimization. The key contributions and findings are summarized as follows.

- *MMSE beamforming for distributed AirComp*: Consider the design criterion of minimizing the sum AirComp error defined as the sum *mean-squared error* (MSE) over all devices with respect to the desired average-functional values. We propose an efficient approach for optimally solving the mentioned problem of distributed multicast beamforming. Without compromising the optimality, the optimal alignment factor conditioned on beamforming is derived in a closed form and substituted into the original problem. Even though the resultant problem is still non-convex, it can be transformed into an equivalent concave-convex fractional program. The transformation admits an efficient solution method that nests solving a convex problem into a bisection search. The results reveal that the

optimal multicast beamforming at a device is steered along the *centroid* of the column vectors of a traditional MMSE precoder for the associated D2D multiuser channel to the peers, thereby balancing their AirComp errors; at least one device transmits with full power while some devices transmit with partial power.

- *Zero-forcing (ZF) beamforming for distributed AirComp*: Consider the design criterion of forcing all receive signal power to approach a uniform level for the purpose of aggregation without attempting to avoid potential noise amplification as in the MMSE case. To minimize the resultant sum AirComp error, the ZF multicast beamformers conditioned on the alignment factor can be reduced into a single-device design with the solution derived in a closed form. It is observed to have a similar centroid form as the MMSE counterpart but is based on a traditional ZF precoder. Given the beamformers, the alignment factor is obtained as the minimum of a derived expression over devices, which represents an attempt to cope with the weakest set of D2D links limiting the performance of distributed AirComp.
- *Convergence of distributed optimization*: The preceding distributed AirComp framework is applied to implement distributed optimization over D2D links, which is based on the widely-used distributed dual averaging algorithm. The convergence analysis shows that AirComp errors induce bias terms in the gap between the minimized loss function and its ground truth. The key finding is that the MMSE and ZF designs for distributed multicast beamforming lead to *biased* and *unbiased* estimations of ground-truth stochastic subgradients at devices, respectively. Consequently, at a low-to-medium receive SNR, the former results in a converged test accuracy substantially lower than that in the ideal (noise-free) case while the loss of the ZF counterpart is negligible. This is opposite to the fact that ZF is sub-optimal in terms of minimizing the sum MMSE AirComp error. For both designs, the said loss is shown to diminish as the transmit SNR grows by being inversely proportional to its square root.

The remainder of this paper is organized as follows. An overview of distributed AirComp is given in Section II. The MMSE and ZF designs of distributed multicast beamforming are presented in Sections III and IV, respectively. The application of distributed AirComp to distributed optimization is studied in Section V. Experimental results are presented in Section VI, followed by concluding remarks in Section VII.

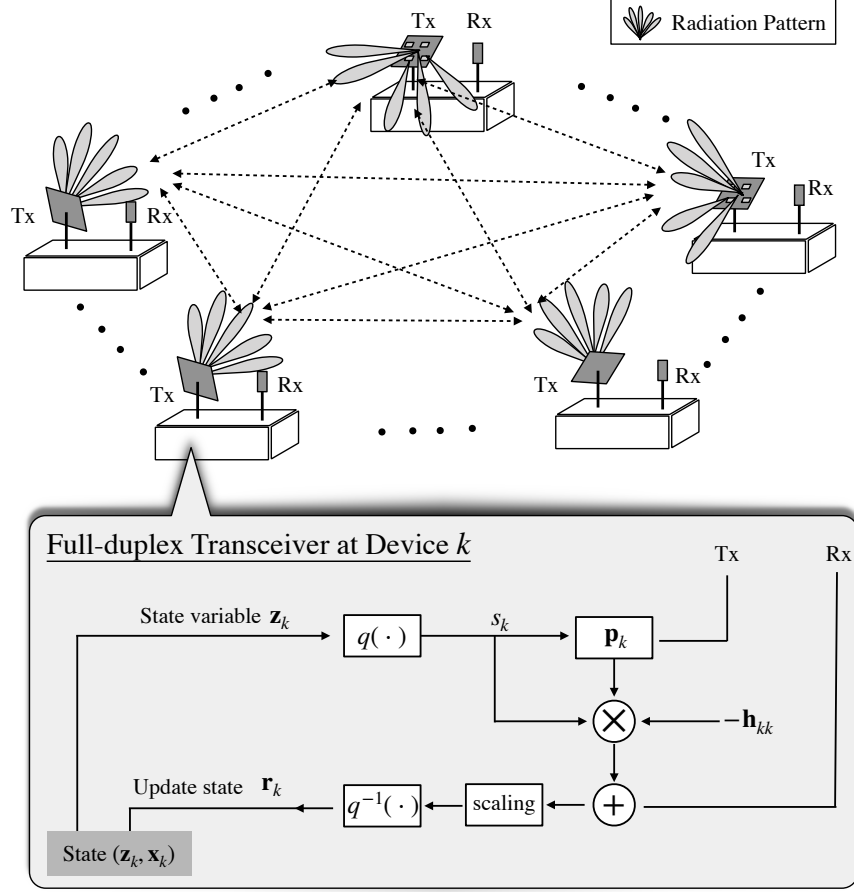


Fig. 1. Distributed AirComp system featuring distributed multicast beamforming and full-duplex communication.

## II. OVERVIEW OF DISTRIBUTED AIRCOMP

### A. System Model

As illustrated in Fig. 1, the considered distributed system comprises  $K$  edge devices without coordination by a server. Each device is equipped with  $N_t$  transmit antennas and one single receive antenna to support multicast beamforming and full-duplex communication. The number of transmit antennas is assumed sufficiently large, i.e.,  $(N_t \geq K - 1)$ , to provide enough degrees-of-freedom (DoF) for each device to support simultaneous aggregations at  $(K - 1)$  peers. For full-duplex communication, it is assumed that a transmitter is perfectly decoupled from a receiver at the same device via passive and/or active self-interference cancellation [27] as illustrated in Fig. 1.

As the process of distributed optimization spans multiple rounds, time is divided into rounds with a fixed duration denoted as  $T_{\text{cmm}}$ ; each round is further divided into  $D$  symbol slots. Channels are assumed to be static in each round and varying over rounds. For simplicity,

each link is assumed frequency non-selective with the bandwidth represented by  $B$ , giving the symbol duration  $T_s = \frac{1}{B}$ . Global *channel state information* (CSI) is available at each device via estimation exploiting channel reciprocity or efficient feedback [15]. The required training overhead is assumed negligible compared with high-dimensional LSI exchange. In each round, each device, say device  $k$ , transmits a symbol vector, denoted as  $\mathbf{s}_k$ , that comprises  $D$  symbols and spans the round duration of  $T_{\text{cmm}} = DT_s$ .

### B. Distributed AirComp Design

An iterative algorithm for distributed optimization (see Appendix A) comprises multiple communication rounds. In each round, all devices broadcast their LSI simultaneously using linear analog modulation and multicast beamforming, and at the same time receive the over-the-air aggregated signals from other devices. All devices are assumed to be synchronized via a common clock. Consider an arbitrary round, for each device, say device  $k$ , let  $\mathbf{p}_k \in \mathbb{C}^{N_t \times 1}$  denote the multicast beamformer under the power constraint,  $\|\mathbf{p}_k\|^2 \leq P_0$ , with  $P_0$  representing the maximum power. Then the over-the-air aggregated signal vector received by the device, denoted as  $\mathbf{y}_k$ , is given as

$$\mathbf{y}_k = \sum_{\ell=1, \ell \neq k}^K \mathbf{h}_{\ell k}^H \mathbf{p}_\ell \mathbf{s}_\ell + \tilde{\mathbf{w}}_k, \quad (1)$$

where  $\mathbf{s}_\ell$  denotes the symbol row vector transmitted by device  $\ell$ ,  $\mathbf{h}_{\ell k} \in \mathbb{C}^{N_t \times 1}$  is the channel vector for the link from device  $\ell$  to  $k$ , and  $\tilde{\mathbf{w}}_k \sim \mathcal{CN}(\mathbf{0}, \sigma^2 \mathbf{I})$  is channel noise. Upon receiving the aggregated signal, it is scaled as shown below to yield the desired aggregation form:

$$\mathbf{r}_k = \frac{\mathbf{y}_k}{(K-1)\sqrt{\eta}}, \quad (2)$$

where  $\eta$  is the alignment factor that helps to reduce the AirComp error.

Consider round  $n$  and device  $k$ . The transmitted symbol vector is a linear analog modulated local state variable ( $D \times 1$  vector),  $\mathbf{z}_k(n)$  (see Appendix A). Normalization of the variable is necessary for meeting the transmission power constraint and avoiding a nonzero DC level, which unnecessarily increases the power. For distributed optimization with a general objective function, typically the statistics of the elements of  $\mathbf{z}_k(n)$  are different and also vary over rounds (see e.g., [28]). Define the round-dependent mean and variance of local state variables, which are assumed uniform for all devices, as  $M(n) = \mathbb{E} \left[ \frac{1}{D} \sum_d z_k^{(d)}(n) \right]$  and  $V^2(n) = \text{Var} \left[ \frac{1}{D} \sum_d z_k^{(d)}(n) \right]$ . These parameters can be estimated offline and stored at devices [28]. Let  $q(\cdot)$  denote the normalization

function such as  $\mathbf{s}_k(n) = q(\mathbf{z}_k(n)) = \frac{\mathbf{z}_k(n) - M(n)}{V(n)}$  for all  $k$ . Consequently,  $\text{Var} \left[ \frac{1}{D} \sum_d s_k^{(d)}(n) \right] = 1$  and  $\mathbb{E}[\frac{1}{D} \sum_d s_k^{(d)}(n)] = 0$  with  $s_k^{(d)}(n)$  being the  $d$ th element of  $\mathbf{s}_k(n)$ . Note that  $\mathbb{E}[\mathbf{s}_k(n)] \neq \mathbf{0}$  due to the non-uniform distributions of the elements in  $\mathbf{s}_k(n)$ . Then the received signal is de-normalized as

$$\mathbf{r}_k = q^{-1} \left( \frac{\mathbf{y}_k}{(K-1)\sqrt{\eta}} \right) = \frac{V}{(K-1)\sqrt{\eta}} \sum_{\ell=1, \ell \neq k}^K \mathbf{h}_{\ell k}^H \mathbf{p}_\ell \mathbf{s}_\ell + M + \mathbf{w}_k, \quad (3)$$

where the channel noise  $\mathbf{w}_k \sim \mathcal{CN} \left( \mathbf{0}, \frac{V^2 \sigma^2}{(K-1)^2 \eta} \mathbf{I} \right)$ .

### C. Distributed Optimization and Performance Metric

In distributed optimization, there exist local objective functions associated with individual devices. The function for device  $k$  is denoted as  $f_k : \mathbb{R}^D \rightarrow \mathbb{R}$ . Following the common assumptions, each function is assumed to be convex and hence sub-differentiable, but not necessarily smooth. The global objective, denoted as  $f$ , is related to local objectives by  $f(\mathbf{x}) = \frac{1}{K} \sum_{k=1}^K f_k(\mathbf{x})$  for a common state of devices. Then the goal of distributed optimization is to find the optimal common state  $\mathbf{x}^*$ , which represents a consensus among devices, that minimizes the objective. Mathematically,

$$\mathbf{x}^* = \arg \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) = \arg \min_{\mathbf{x} \in \mathcal{X}} \frac{1}{K} \sum_{k=1}^K f_k(\mathbf{x}), \quad (4)$$

where the state space  $\mathcal{X}$  is a closed,  $D$ -dimensional convex set.

A classic algorithm, distributed dual averaging, for solving (4) is discussed in Appendix A. In this algorithm, the local state of each device, say device  $\ell$ , contains a primal variable and a dual variable, which are denoted as  $\mathbf{x}_\ell(n) \in \mathbb{R}^D$  and  $\mathbf{z}_\ell(n) \in \mathbb{R}^D$  for round  $n$ , respectively. Each round, say round  $n$ , of the algorithm comprises a three-step procedure simultaneously executed at each device, say device  $\ell$ : (1) calculate a stochastic subgradient  $\tilde{\mathbf{g}}_\ell(n)$  of its local loss function at  $\mathbf{x}_\ell(n)$ ; (2) aggregate the dual variables of all other devices and apply the result together with  $\tilde{\mathbf{g}}_\ell(n)$  to update the local dual variable as  $\mathbf{z}_\ell(n+1)$ ; (3) project  $\mathbf{z}_\ell(n+1)$  onto the feasible set  $\mathcal{X}$  to obtain the updated primal variable  $\mathbf{x}_\ell(n+1)$  [29]. Only Step (2) requires information exchange among devices. We focus on its one-step implementation using distributed AirComp in the preceding subsection. To this end, (27) for Step (2) at device  $\ell$  is modified as

$$\mathbf{z}_\ell(n+1) = (1 - \beta)\mathbf{z}_\ell(n) + \beta\mathbf{r}_\ell(n) + \tilde{\mathbf{g}}_\ell(n) \quad (5)$$



where the received signal  $\mathbf{r}_\ell(n)$  is given in (3) and  $\beta$  is a constant weight (see Appendix A). Note that the ideal received signal, namely the ground truth, is  $\frac{1}{K-1} \sum_{k \neq \ell}^K \mathbf{z}_k(n)$ . Then the sum AirComp error is suitably defined as the sum *mean-square-error* (MSE) between the received signal and its ground-truth:

$$\begin{aligned}
\text{MSE} &= \mathbb{E} \left[ \sum_{\ell=1}^K \sum_{d=1}^D \left( r_\ell^{(d)} - \frac{1}{K-1} \sum_{k \neq \ell}^K z_k^{(d)} \right)^H \left( r_\ell^{(d)} - \frac{1}{K-1} \sum_{k \neq \ell}^K z_k^{(d)} \right) \right] \\
&= \frac{V^2 D}{(K-1)^2} \mathbb{E} \left[ \sum_{\ell=1}^K \left( \sum_{k \neq \ell}^K \frac{\mathbf{h}_{k\ell}^H \mathbf{p}_k}{\sqrt{\eta}} s_k^{(d)} - \sum_{k \neq \ell}^K s_k^{(d)} + w_k^{(d)} \right)^H \left( \sum_{k \neq \ell}^K \frac{\mathbf{h}_{k\ell}^H \mathbf{p}_k}{\sqrt{\eta}} s_k^{(d)} - \sum_{k \neq \ell}^K s_k^{(d)} + w_k^{(d)} \right) \right] \\
&= \frac{V^2 D}{(K-1)^2} \sum_{\ell=1}^K \left[ \sum_{k \neq \ell}^K \left\| \frac{\mathbf{h}_{k\ell}^H \mathbf{p}_k}{\sqrt{\eta}} - 1 \right\|^2 + \frac{\sigma^2}{\eta} \right] \\
&= \frac{V^2 D}{(K-1)^2} \sum_{k=1}^K \sum_{\ell \neq k}^K \left\| \frac{\mathbf{h}_{k\ell}^H \mathbf{p}_k}{\sqrt{\eta}} - 1 \right\|^2 + \frac{K D V^2 \sigma^2}{(K-1)^2} \frac{1}{\eta},
\end{aligned} \tag{6}$$

where the expectation is taken over the distribution of the transmitted symbol and the received noise  $w_k^{(d)}$  and the round index  $(n)$  is omitted here for brevity.

### III. MMSE DESIGN OF DISTRIBUTED MULTICAST BEAMFORMING

In this section, distributed multicast beamforming for distributed AirComp as well as the alignment factor are jointly optimized under the criterion of minimum sum AirComp error. The result is called minimum MSE (MMSE) multicast beamforming. An efficient solution approach is developed based on fractional programming.

#### A. Optimal Distributed Multicast Beamforming

Consider an arbitrary round with its index omitted for ease of notation. With the sum AirComp error in (6), the mentioned optimization problem is formulated as follows

$$\begin{aligned}
&\min_{\{\mathbf{p}_k\}, \eta} \text{MSE}(\mathbf{p}_1, \dots, \mathbf{p}_K, \eta) \\
&\text{s.t.} \quad \|\mathbf{p}_k\|^2 \leq P_0, \quad \forall k, \\
&\quad \eta \geq 0.
\end{aligned} \tag{P1}$$

Due to the coupling of  $\{\mathbf{p}_k\}$  and  $\eta$ , Problem (P1) is non-convex and hence directly solving it is difficult. To overcome the difficulty, an alternative approach is presented as follows.

First, given  $\{\mathbf{p}_k\}$ , Problem (P1) is reduced to the following conditional problem:

$$\min_{\eta \geq 0} \quad \frac{V^2 D}{(K-1)^2} \sum_{k=1}^K \sum_{\ell \neq k}^K \left\| \frac{\mathbf{h}_{k\ell}^H \mathbf{p}_k}{\sqrt{\eta}} - 1 \right\|^2 + \frac{K D V^2 \sigma^2}{(K-1)^2} \frac{1}{\eta}. \quad (7)$$

The objective in (7) is found to be convex and then the optimal  $\eta_{\text{mmse}}^*$  is obtained as

$$\eta_{\text{mmse}}^* = \left( \frac{2K\sigma^2 + 2 \sum_{\ell=1}^K \sum_{k \neq \ell}^K \mathbf{p}_k^H \mathbf{h}_{k\ell} \mathbf{h}_{k\ell}^H \mathbf{p}_k}{\sum_{\ell=1}^K \sum_{k \neq \ell}^K (\mathbf{h}_{k\ell}^H \mathbf{p}_k + \mathbf{p}_k^H \mathbf{h}_{k\ell})} \right)^2. \quad (8)$$

Next, we proceed to optimize  $\{\mathbf{p}_k\}$ . Substituting  $\eta_{\text{mmse}}^*$  in (8) into Problem (P1) gives

$$\begin{aligned} \min_{\{\mathbf{p}_k\}} \quad & \frac{K}{K-1} - \frac{1}{(K-1)^2} \cdot \frac{\left( \sum_{\ell=1}^K \sum_{k \neq \ell}^K (\mathbf{h}_{k\ell}^H \mathbf{p}_k + \mathbf{p}_k^H \mathbf{h}_{k\ell}) \right)^2}{4(K\sigma^2 + \sum_{\ell=1}^K \sum_{k \neq \ell}^K \mathbf{h}_{k\ell}^H \mathbf{p}_k \mathbf{p}_k^H \mathbf{h}_{k\ell})} \\ \text{s.t.} \quad & \|\mathbf{p}_k\|^2 \leq P_0, \quad \forall k. \end{aligned} \quad (\text{P1.1})$$

Problem (P1.1) remains non-convex. Nevertheless, the result in the following lemma can transform this problem into a tractable equivalent form.

**Lemma 1.** Given  $\sum_{\ell=1}^K \sum_{k \neq \ell}^K (\mathbf{h}_{k\ell}^H \mathbf{p}_k + \mathbf{p}_k^H \mathbf{h}_{k\ell}) \geq 0$ , Problem (P1.1) is equivalent to the following *concave-convex fractional program*:

$$\begin{aligned} \max_{\{\mathbf{p}_k\}} \quad & \frac{\left( \sum_{\ell=1}^K \sum_{k \neq \ell}^K (\mathbf{h}_{k\ell}^H \mathbf{p}_k + \mathbf{p}_k^H \mathbf{h}_{k\ell}) \right)}{2\sqrt{(K\sigma^2 + \sum_{\ell=1}^K \sum_{k \neq \ell}^K \mathbf{h}_{k\ell}^H \mathbf{p}_k \mathbf{p}_k^H \mathbf{h}_{k\ell})}} \\ \text{s.t.} \quad & \|\mathbf{p}_k\|^2 \leq P_0 \quad \forall k, \\ & \sum_{\ell=1}^K \sum_{k \neq \ell}^K (\mathbf{h}_{k\ell}^H \mathbf{p}_k + \mathbf{p}_k^H \mathbf{h}_{k\ell}) \geq 0. \end{aligned} \quad (\text{P1.2})$$

*Proof.* See Appendix B. □

It can be straightforwardly proved by a simple variable substitution. When  $\sum_{\ell=1}^K \sum_{k \neq \ell}^K (\mathbf{h}_{k\ell}^H \mathbf{p}_k + \mathbf{p}_k^H \mathbf{h}_{k\ell}) < 0$ , the optimal solution of (P1.1) can be written as  $\tilde{\mathbf{p}}_k^* = -\mathbf{p}_k^*$ , where  $\mathbf{p}_k^*$  is the optimal solution of (P1.2). Since  $\tilde{\mathbf{p}}_k^*$  and  $\mathbf{p}_k^*$  correspond to the same optimal value of the objective of (P1.1), considering the case in Lemma 1 where  $\sum_{\ell=1}^K \sum_{k \neq \ell}^K (\mathbf{h}_{k\ell}^H \mathbf{p}_k + \mathbf{p}_k^H \mathbf{h}_{k\ell}) \geq 0$  is sufficient for the purpose of solving Problem (P1.1). This allows a property of fractional program to be applied. To this end, define the *upper contour set* of a function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  at  $c = f(x_0)$  for some  $x_0 \in \mathbb{R}^n$  as the set  $\{x \in \mathbb{X} : f(x) \geq c\}$ . One useful property of the concave-convex fractional

program in Lemma 1 is its strict quasi-concavity [30]. Specifically, the upper contour sets of the objective of Problem (P1.2), are *convex*. By introducing an auxiliary variable  $\alpha$  ( $\alpha > 0$ ), which corresponds to the aligned fraction, Problem (P1.2) can be written as the convex problem of maximization over  $\alpha$  as follows:

$$\begin{aligned} \max_{\alpha, \{\mathbf{p}_k\}} \quad & \alpha \\ \text{s.t.} \quad & \|\mathbf{p}_k\|^2 \leq P_0, \quad \forall k, \\ & \alpha \leq \frac{\left( \sum_{\ell=1}^K \sum_{k \neq \ell}^K (\mathbf{h}_{k\ell}^H \mathbf{p}_k + \mathbf{p}_k^H \mathbf{h}_{k\ell}) \right)}{2\sqrt{(K\sigma^2 + \sum_{\ell=1}^K \sum_{k \neq \ell}^K \mathbf{h}_{k\ell}^H \mathbf{p}_k \mathbf{p}_k^H \mathbf{h}_{k\ell})}}. \end{aligned} \quad (\text{P1.3})$$

To solve the above problem requires consideration of a related problem of transmission power minimization described as follows. Define the maximum power as  $p_{\max} = \max_k \|\mathbf{p}_k\|_2^2$ . Then given  $\alpha$ , the mentioned problem is given as

$$\begin{aligned} \min_{\{\mathbf{p}_k\}, p_{\max}} \quad & p_{\max} \\ \text{s.t.} \quad & 2\alpha \sqrt{(K\sigma^2 + \sum_{\ell=1}^K \sum_{k \neq \ell}^K \mathbf{h}_{k\ell}^H \mathbf{p}_k \mathbf{p}_k^H \mathbf{h}_{k\ell})} \leq \left( \sum_{\ell=1}^K \sum_{k \neq \ell}^K (\mathbf{h}_{k\ell}^H \mathbf{p}_k + \mathbf{p}_k^H \mathbf{h}_{k\ell}) \right), \\ & \|\mathbf{p}_k\|_2^2 \leq p_{\max}, \quad \forall k. \end{aligned} \quad (\text{P1.4})$$

The solution of Problem (P1.4) is denoted as a function of  $\alpha$ ,  $p^*(\alpha)$ . Note that given the desired aligned fraction  $\alpha$ , Problem (P1.4) is feasible if and only if the solution for Problem (P1.4) satisfies  $p^*(\alpha) \leq P_0$ . Two useful lemmas for relating Problems (P1.3) and (P1.4) are given as follows, which are proved in Appendices C and D, respectively.

**Lemma 2.** The minimum transmission power  $p^*(\alpha)$  over devices is a monotonously increasing function of  $\alpha$ .

It follows from the result in Lemma 2 that the solution for Problem (P1.3) is the maximum alignment level, denoted as  $\alpha^*$ , such that the corresponding minimum transmission power  $p^*(\alpha^*)$  is no larger than  $P_0$ . This suggests a solution method of Problem (P1.3) by a search for  $\alpha^*$  over the range of  $0 < p^*(\alpha) \leq P_0$ . This requires solving Problem (P1.4) so as to compute the function  $p^*(\alpha)$ . To this end, the following result is useful.

**Lemma 3.** Given  $\alpha$ , Problem (P1.4) is convex.

---

**Algorithm 1** Optimal Algorithm for MMSE Distributed Multicast Beamforming
 

---

```

1: Inputs:  $K, \sigma, \{\mathbf{h}_{k\ell}\}$ .
2: Initialization:
   Select  $\alpha_u$  so that  $\alpha = \alpha_u$  makes  $p^*(\alpha_u)$  defined in (P1.4) larger than  $P_0$ .
   Select  $\alpha_l$  so that  $\alpha = \alpha_l$  makes  $p^*(\alpha_l) < P_0$ .
3: while  $|\alpha_u - \alpha_l| \geq \epsilon$  do
4:   Let  $\alpha_m = (\alpha_u + \alpha_l)/2$  and substitute  $\alpha = \alpha_m$  into (P1.4).
5:   Solve (P1.4) by CVX toolbox to obtain  $p^*(\alpha_m)$  and  $\{\mathbf{p}_k^*\}$ .
6:   if  $p^*(\alpha_m) > P_0$  then
7:      $\alpha_u = \alpha_m$ .
8:   else
9:      $\alpha_l = \alpha_m$ .
10:  end if
11: end while
12: return  $\{\mathbf{p}_k^*\}$ .

```

---

The convexity of Problem (P1.4) allows it to be solved efficiently using rich existing techniques for convex programming, for example, the primal-dual method.

In summary, Problem (P1.3) for optimal transmit beamforming can be solved efficient by nesting a one-dimensional search over  $\alpha$  and the solution of the convex Problem (P1.4). The detailed algorithm is presented in Algorithm 1.

### B. Centroid Beamforming

To shed light on the structure of the optimal MMSE design of distributed multicast beamformer, we consider a special case when the number of antennas  $N_t = K-1$ . Let  $\mathbf{H}_k \in \mathbb{C}^{N_t \times (K-1)}$  be the matrix with its  $\ell$ -th column being  $\mathbf{h}_{k\ell}^H$  with  $\ell \neq k$ . In the current case,  $\mathbf{H}_k$  is an invertible full-rank matrix. The convexity of Problem (P1.4) allows its KKT conditions to be a sufficient and necessary condition for its optimal solution. Based on the KKT conditions, Lemmas 4 and 5 are obtained as follows with proofs in Appendix E.

**Lemma 4.** At least one device performs full-power transmission, i.e.  $\exists k, \|\mathbf{p}_k\|^2 = P_0$ .

**Lemma 5** (Centroid Beamforming). Define  $[\mathbf{H}]_\ell$  as the  $\ell$ -th column of matrix  $\mathbf{H}$ . When  $N_t = K - 1$ , the beamforming directions of devices with partial-power transmission are given as

$$\frac{\mathbf{p}_k^*}{\|\mathbf{p}_k^*\|} = \frac{\sum_{\ell \neq k}^K [(\mathbf{H}_k^H)^{-1}]_\ell}{\left\| \sum_{\ell \neq k}^K [(\mathbf{H}_k^H)^{-1}]_\ell \right\|}, \quad (9)$$

and those with full-power transmission are given as

$$\frac{\mathbf{p}_k^*}{\|\mathbf{p}_k^*\|} = \frac{\sum_{\ell \neq k}^K [(\mathbf{H}_k \mathbf{H}_k^H + \mu_k \mathbf{I})^{-1} \mathbf{H}_k]_\ell}{\left\| \sum_{\ell \neq k}^K [(\mathbf{H}_k \mathbf{H}_k^H + \mu_k \mathbf{I})^{-1} \mathbf{H}_k]_\ell \right\|}, \quad (10)$$

where  $\mu_k$  is the regularization term that diminishes as the SNR increases.

The result in (9) shows that the direction of the optimal beamforming of a device with partial-power transmission points to the centroid of the column vectors of a ZF or a regularized channel-inversion precoder. Such a design overcomes fading of multiuser channels to facilitate simultaneous signal alignments at other devices while balancing their AirComp errors.

#### IV. ZERO-FORCING DESIGN OF DISTRIBUTED MULTICAST BEAMFORMING

In this section, we consider the ZF design criterion of forcing the received signals to approach the desired ground-truth values without consideration of channel noise. Under this criterion, a low-complexity design of distributed multicast beamforming is presented as follows. Conditioned on the alignment factor  $\eta$ , the ZF criterion allows the joint beamforming design to be decoupled into the following individual beamforming problems formulated for device  $k$  as

$$\mathbf{p}_k^* = \arg \min_{\mathbf{p}_k} \left\| \mathbf{H}_k^H \mathbf{p}_k - \sqrt{\eta} \mathbf{1}_{(K-1)} \right\|^2, \quad \forall k, \quad (\text{P2})$$

where the objective function is termed as *misalignment error*. Note that in this case, since  $N_t \geq K - 1$ , there is sufficient DoF to align the channels, i.e.,  $\mathbf{H}_k^H \mathbf{p}_k = \sqrt{\eta} \mathbf{1}_{(K-1)}$ ,  $\forall k$  is always satisfied. Therefore  $\left\| \mathbf{H}_k^H \mathbf{p}_k - \sqrt{\eta} \mathbf{1}_{(K-1)} \right\|^2 = 0$  always has non-trivial solutions. Then following lemma gives a solution with the smallest norm.

**Proposition 1.** When  $N_t \geq K - 1$ , given the alignment factor  $\eta$ , the misalignment error of device  $k$  is minimized by following ZF multicast beamforming:

$$\mathbf{p}_k = \sqrt{\eta} \mathbf{H}_k (\mathbf{H}_k^H \mathbf{H}_k)^{-1} \mathbf{1}_{(K-1)}, \quad \forall k, \quad (11)$$

where  $\eta$  is chosen to meet the power constraint.

*Proof.* See Appendix F. □

The beamformer in (11) can be written in a centroid form similarly as its MMSE counterpart, i.e.,  $\mathbf{p}_k = \sqrt{\eta} \sum_{\ell \neq k}^K \left[ \mathbf{H}_k (\mathbf{H}_k^H \mathbf{H}_k)^{-1} \right]_{\ell}$ . Specifically, this multicast beamforming is given by the centroid of the column vectors of a traditional ZF precoder for the associated D2D multiuser channel to the peers. However, it should be noted that compared with MMSE beamforming, this ZF scheme can potentially amplify the noise and is therefore vulnerable to deep fading.

By substituting (11) in Proposition 1, the problem of alignment-error minimization over the alignment factor  $\eta$  is

$$\begin{aligned} \min_{\{\eta\}} \quad & \frac{K D V^2 \sigma^2}{(K-1)^2 \eta}, \\ \text{s.t.} \quad & \eta \mathbf{1}_{(K-1)}^H (\mathbf{H}_k^H \mathbf{H}_k)^{-1} \mathbf{1}_{(K-1)} \leq P_0, \quad \forall k. \end{aligned} \quad (\text{P2.1})$$

Since the objective of Problem (P2.1) is a monotonically decreasing function of  $\eta$ , the optimal  $\eta_{\text{ZF}}^*$  is directly given as

$$\eta_{\text{ZF}}^* = \min_k \frac{P_0}{\mathbf{1}_{(K-1)}^H (\mathbf{H}_k^H \mathbf{H}_k)^{-1} \mathbf{1}_{(K-1)}}. \quad (12)$$

By constructing a Rayleigh quotient as following, we have

$$\begin{aligned} \eta_{\text{ZF}}^* &= \min_k \frac{P_0 \mathbf{1}_{(K-1)}^H \mathbf{1}_{(K-1)}}{(K-1) \mathbf{1}_{(K-1)}^H (\mathbf{H}_k^H \mathbf{H}_k)^{-1} \mathbf{1}_{(K-1)}} \geq \min_k \frac{P_0}{(K-1) \lambda_{\max} \left( (\mathbf{H}_k^H \mathbf{H}_k)^{-1} \right)} \\ &\stackrel{(a)}{=} \min_k \frac{P_0}{(K-1) \lambda_{\min}^{-1} (\mathbf{H}_k^H \mathbf{H}_k)}, \end{aligned} \quad (13)$$

where (a) is due to that  $\mathbf{H}_k^H \mathbf{H}_k$  is a hermitian matrix and  $\lambda_{\min}(\cdot)$ ,  $\lambda_{\max}(\cdot)$  correspond to the minimum and maximum eigenvalue of given matrix separately. Using this upper bound on  $\eta_{\text{ZF}}^*$  leads to an upper bound on the sum AirComp error in the current case:

$$\text{MSE}_{\text{ZF}} \leq \frac{K D V^2 \sigma^2}{(K-1) P_0 \min_k \lambda_{\min} (\mathbf{H}_k^H \mathbf{H}_k)}. \quad (14)$$

Note that the minimum eigenvalue  $\lambda_{\min}(\mathbf{H}_k^H \mathbf{H}_k)$  is a monotonically increasing function of  $N_t$  [26]. This demonstrates the diversity gain of increasing the transmit array size for reducing the sum AirComp error. Next, comparing the sum AirComp errors of the MMSE and ZF designs, it is worth mentioning that  $\text{MSE}_{\text{mmse}} \leq \text{MSE}_{\text{ZF}}$ .

## V. APPLICATION TO DISTRIBUTED OPTIMIZATION

In this section, we consider the application of distributed AirComp designed in the preceding sections to provide an efficient air interface for implementing the distributed-optimization algorithm in Appendix A in a D2D network. The effects of AirComp error on its convergence are characterized mathematically. For tractable analysis, several assumptions commonly made in the literature (see e.g., [31]) are also adopted in this work.

**Assumption 1** (Continuity). Each local objective function is L-Lipschitz continuous:

$$|f_k(\mathbf{x}) - f_k(\mathbf{y})| \leq L\|\mathbf{x} - \mathbf{y}\| \quad \text{for } \mathbf{x}, \mathbf{y} \in \mathcal{X} \quad (15)$$

where  $L$  is a constant.

**Assumption 2** (Local Gradient Estimation). The stochastic subgradient  $\tilde{\mathbf{g}}_k(n)$  at the  $k$ -th device is an *unbiased* estimate of the ground-truth local subgradient,  $\mathbf{g}_k(n) \in \partial f_k(\mathbf{x}_k(n))$ , with bounded second moment:

$$\mathbb{E}[\tilde{\mathbf{g}}_k(n)] = \mathbf{g}_k(n) \quad \text{and} \quad \mathbb{E}[\|\tilde{\mathbf{g}}_k(n)\|^2] \leq \Omega^2, \quad (16)$$

where  $\Omega$  is a given constant.

### A. Effects of AirComp Error

In this subsection, we investigate the effects of AirComp error on the subgradient estimation and state updating, which are the key operations of distributed optimization. The variables in the subsequent analysis follow those defined in Appendix A.

First, consider round  $n$ , the receive signal vector at device  $k$  in (3) can be rewritten as

$$\mathbf{r}_k(n) = \frac{1}{K-1} \sum_{\ell=1, \ell \neq k}^K \mathbf{z}_\ell(n) + \Delta_k(n), \quad (17)$$

where the random variable  $\Delta_k(n)$  defined below represents the distortion caused by wireless propagation:

$$\Delta_k(n) = \frac{V(n)}{K-1} \sum_{\ell=1, \ell \neq k}^K \left( \frac{\mathbf{h}_{k\ell}^H(n) \mathbf{p}_k(n)}{\sqrt{\eta(n)}} - 1 \right) \mathbf{s}_\ell(n) + \mathbf{w}_k(n). \quad (18)$$

Note that the summation of  $\|\Delta_k(n)\|^2$  over devices gives the sum AirComp error in (6). Then based on (5) and (30), the state variable at device  $i$  is updated as

$$\mathbf{z}_k(n+1) = \sum_{\ell=1}^K W_{k\ell} \mathbf{z}_\ell(n) + \hat{\mathbf{g}}_k(n), \quad (19)$$

where  $\hat{\mathbf{g}}_k(n)$  represents the channel distorted version of the stochastic subgradient,  $\tilde{\mathbf{g}}_k(n)$ , namely  $\hat{\mathbf{g}}_k(n) = \tilde{\mathbf{g}}_k(n) + \beta \Delta_k(n)$ .

Next, the effects of AirComp error are reflected by the deviation of the noisy subgradient,  $\hat{\mathbf{g}}_k(n)$ , from its ground truth. Consider ZF beamforming designed in Section IV. From its definition, the channel distortion in (18),  $\Delta_k(n)$ , reduces to channel noise:  $\Delta_k(n) = \mathbf{w}_k(n)$ . It follows from this fact and Assumption 1 that the noisy subgradient  $\hat{\mathbf{g}}_k(n)$  yields an *unbiased estimate* of the ground truth:

$$\mathbb{E} [\hat{\mathbf{g}}_k(n)] = \mathbb{E} [\tilde{\mathbf{g}}_k(n) + \beta \Delta_k(n)] = \mathbf{g}_k(n). \quad (20)$$

Consider MMSE beamforming designed in Section III. It follows from (18) and its definition, the corresponding noisy subgradient, however, is a *biased estimate* of the ground truth:

$$\mathbb{E} [\hat{\mathbf{g}}_k(n)] = \mathbb{E} [\tilde{\mathbf{g}}_k(n) + \beta \Delta_k(n)] = \mathbf{g}_k(n) + \frac{\beta V(n)}{K-1} \sum_{\ell=1, \ell \neq k}^K \left( \frac{\mathbf{h}_{k\ell}^H(n) \mathbf{p}_k(n)}{\sqrt{\eta(n)}} - 1 \right) \mathbb{E} [\mathbf{s}_\ell(n)], \quad (21)$$

where  $\mathbb{E} [\mathbf{s}_\ell(n)] \neq 0$  as discussed in Section II. Even though the MMSE beamforming achieves lower AirComp error as defined in (6) than the ZF counterpart, the former's biased subgradient estimation leads to worse convergence performance as elaborated in the next subsection.

Last, a result useful for convergence analysis is obtained by bounding the deviation of the state variable (i.e., dual variable),  $\mathbf{z}_k(n)$ , from its average  $\bar{\mathbf{z}}(n) = \frac{1}{K} \sum_k \mathbf{z}_k(n)$ , termed *dual-variable deviation*. Note that it also serves as a measure of the deviation of individual noisy subgradients. Mathematically, the deviation is defined as  $\mathbb{E} \|\bar{\mathbf{z}}(n) - \mathbf{z}_k(n)\|_*$  with  $\|\cdot\|_* = \sup_{\|u\|=1} \langle \cdot, u \rangle$  denoting the dual norm. To bound the dual-variable deviation, based on Assumption 2, the



second moment of the noisy sub-gradient,  $\hat{\mathbf{g}}_k(n)$ , can be bounded as:

$$\mathbb{E} [\|\hat{\mathbf{g}}_k(n)\|^2] = \mathbb{E} [\|\tilde{\mathbf{g}}_k(n)\|^2 + \beta^2 \|\Delta_k\|^2] \leq \Omega^2 + \frac{\beta^2 \text{MSE}(n)}{K}, \quad (22)$$

where the AirComp error  $\text{MSE}(n)$  defined in (6) applies to both the cases of ZF and MMSE beamforming. Using the bound, the desired result is obtained as follows.

**Lemma 6.** For distributed optimization using distributed AirComp (with either MMSE or ZF beamforming), the expected dual-variable deviation can be bounded as:

$$\mathbb{E} \|\bar{\mathbf{z}}(n) - \mathbf{z}_k(n)\|_* \leq \frac{2\xi}{\beta(1 - \lambda_2)} \log(N\sqrt{K}) + 3\xi, \quad (23)$$

where  $\xi = \sqrt{\Omega^2 + \frac{\beta^2 \max_n \text{MSE}(n)}{K}}$  and  $\lambda_2$  is the second largest eigenvalue of the edge-weight matrix  $\mathbf{P}$ .

*Proof.* See Appendix G. □

One can observe from the above result that dual-variable deviation grows with the numbers of rounds and users following  $O(\log(N\sqrt{K}))$ . Nevertheless, it is shown in the next section that with a properly chosen step size, the noisy primal variables,  $\{\hat{\mathbf{x}}_k(N)\}$ , can asymptotically approach the optimal point as the number of rounds,  $N$ , increases.

### B. Convergence Analysis

The convergence of the distributed-optimization algorithm (see Appendix A) as implemented using distributed AirComp is analyzed as follows. To begin with, given  $N$  rounds, the convergence is evaluated using the metric of expected suboptimality gap,  $\mathbb{E} [f(\hat{\mathbf{x}}_k(N)) - f(\mathbf{x}^*)]$ , where  $\hat{\mathbf{x}}_k(N)$  and  $\mathbf{x}^*$  represent the noisy state at device  $k$  and the optimal point, respectively.

**Theorem 1** (Convergence with Distributed AirComp). Given the optimal point  $\mathbf{x}^*$ , assume that the proximal function satisfying (26) in Appendix A is bounded by some constant  $R$  as  $\psi(\mathbf{x}^*) \leq R^2$ . Let the step size  $\alpha(n)$  be chosen as  $\alpha(n) = \frac{R\sqrt{1-\lambda_2}}{4\xi\sqrt{n}}$ . Then the expected suboptimality gap can be bounded for the ZF multicast beamforming as

$$\mathbb{E} [f(\hat{\mathbf{x}}_k(N)) - f(\mathbf{x}^*)] \leq \frac{20R \log(N\sqrt{K})}{\beta\sqrt{N}\sqrt{1-\lambda_2}} \left( \Omega^2 + \frac{\beta^2 \max_n \text{MSE}(n)}{K} \right)^{\frac{1}{2}}, \quad (24)$$

and for the MMSE design as

$$\mathbb{E}[f(\hat{\mathbf{x}}_k(N)) - f(\mathbf{x}^*)] \leq \frac{20R \log(N\sqrt{K})}{\beta\sqrt{N}\sqrt{1-\lambda_2}} \left( \Omega^2 + \frac{\beta^2 \max_n \text{MSE}(n)}{K} \right)^{\frac{1}{2}} + \frac{\|\mathbf{x}^*\|}{N} \sum_{n=1}^N \sqrt{\frac{\text{MSE}(n)}{K}}, \quad (25)$$

where  $\lambda_2$  is the second largest eigenvalue of the edge-weight matrix  $\mathbf{P}$ .

*Proof.* See Appendix H. □

In the above results, those terms comprising the AirComp error,  $\text{MSE}(n)$ , reflect the effects of distributed AirComp. For a sanity check, substituting  $\text{MSE}(n) = 0$  into the results in Theorem 1 yields the existing result for the ideal case of noiseless channels [29]. Next, comparing with (24), the last term at the right-hand side of (25) reveals slower convergence due to the MMSE beamforming as opposed to the ZF counterpart due to the former's bias in subgradient estimation as shown in the preceding subsection. More critically, the said term does not vanish as the number of rounds,  $N$ , grows, but instead converges to a constant,  $\|\mathbf{x}^*\| \cdot \mathbb{E}[\sqrt{\text{MSE}(n)/K}]$ . As a result, in terms of expected suboptimality gap, ZF beamforming can significantly outperform the MMSE counterpart in the regime of low-to-medium SNRs. Last, the AirComp-error term, introduced by  $\frac{\beta^2 \max_n \text{MSE}(n)}{K}$ , scales with the number of rounds,  $N$ , and devices,  $K$ , following  $O(\log(N\sqrt{K})/\sqrt{N})$ . However, it decays with the increasing transmit SNR following  $O(1/\sqrt{\text{SNR}})$ . This leads to a narrowing performance gap between the ZF and MMSE beamforming as the SNR grows.

## VI. SIMULATION RESULTS

A distributed system with a varying number of devices,  $K$ , is simulated. In the system, all channel gains are modelled as i.i.d. Rician fading with the power ratio between the direct and scatter paths being 0.6 and unit total power. Transmit SNR is defined as  $\text{SNR} = P_0/\sigma^2$  and is set equal for all devices. The bandwidth is  $B = 1$  MHz. Other case-dependent simulation settings are specified in the sequel. For distributed optimization, we consider the specific scenario of distributed FEEL that trains a classifier model for handwritten-digit recognition using the well-known MNIST dataset. This dataset contains 60,000 labeled training data samples in total. To distribute these samples in a *non i.i.d.* manner, they are first sorted by their digit label, then divided into 20 shards of size 3,000, with 2 shards allocated to each of the 10 devices. The classifier model is implemented using a 6-layer convolutional neural network (CNN) that consists

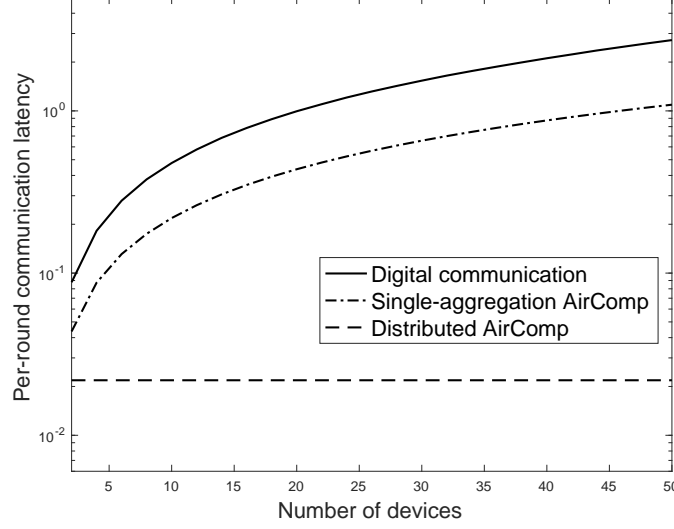


Fig. 2. Per-round communication latency with transmit SNR = 20 dB.

of two  $5 \times 5$  convolution layers with ReLU activation, each followed by a  $2 \times 2$  max pooling layer, a fully-connected layer with 512 units and ReLU activation, and a final softmax output layer. The total number of parameters is 21,840. The test accuracy is defined as the lowest test accuracy among all devices. In this scenario, the LSI of each device is a locally computed stochastic gradient for model updating.

Two benchmarking schemes are described below

- **Digital communication:** Each coefficient of a transmitted state variable is quantized into  $Q = 16$  bits, which are transmitted reliably at a capacity-achieving rate. In each round, devices take turn to broadcast their LSI to peers based on *time-division multiple access* (TDMA). ZF precoding is applied.
- **Single-aggregation AirComp:** In each round, the  $K$  AirComp processes in the system are orthogonalized or equivalently completed sequentially using TDMA, giving the name of the scheme. Consequently, the per-round latency is  $K$ -time higher than that of the proposed distributed AirComp although the AirComp error is smaller as demonstrated in the sequel. The traditional beamforming design for a multiple-input-single-output channel is applied together with effective channel inversion for receive-signal alignment, which is a special case of the AirComp design in [15].

### A. Performance of Distributed AirComp

In Fig. 2, the per-round communication latency for LSI exchange between  $K$  devices is compared between the proposed distributed AirComp and benchmarking schemes for a varying number of devices,  $K$ . The transmit SNR is 20 dB. To support the number of devices as many as 50, each device is provisioned with a large-scale array with  $N_t = 100$ . One can observe that the feature of simultaneous multicasting enables the proposed distributed AirComp to keep the latency constant instead of increasing with the number of devices as for the benchmarking schemes. Distributed AirComp is observed to achieve much lower latency than the latter with the gap increasing rapidly as  $K$  grows. When there are many devices (e.g., 50), the latency reduction of distributed AirComp is more than *two-order of magnitude* with respect to (w.r.t.) digital communication and about 50-time w.r.t. single-aggregation AriComp.

Next, the curves of AirComp error versus transmit SNR, number of transmit antennas  $N_t$ , and number of devices  $K$  are plotted in Fig. 3(a), 3(b), and 3(c), respectively. By default, the number of devices  $K = 5$ . All simulation results demonstrate decreasing AirComp error as any of the three parameters increases. Relatively small array sizes,  $N_t = 4$  and  $N_t = 18$ , are considered in Fig. 3(a) and 3(c), respectively, to investigate the limitations of distributed AirComp. One can observe from Fig. 3(a) that distributed AirComp, which strives to support  $K$  AirComp processes simultaneously despite a small array size, incurs about 10-time larger AirComp error than single-aggregation AirComp. Furthermore, as shown in Fig. 3(c), the former sees that the AirComp error saturates as  $K$  increases, indicating the cancellation of the opposite effects of aggregation gain in error suppression and severer receive-signal misalignment. The above disadvantage of distributed AirComp as a price for dramatic latency reduction is alleviated when the array size increases as shown in Fig. 3(b). For instance, for  $N_t = 40$ , the error ratio between distributed AirComp and single-aggregation counterpart reduces to below 4 times. In addition, it can be observed that for distributed AirComp, MMSE multicast beamforming outperforms the ZF design in the regimes of low-to-medium SNRs or array size or as the number of devices grows.

### B. Performance of Distributed Optimization

Consider FEEL, a typical scenario of distributed optimization, where the number of devices  $K = 10$  and each is equipped with  $N_t = 18$  antennas. The distributed FEEL algorithm is implemented using either distributed AirComp or one of the benchmarking schemes. Define the

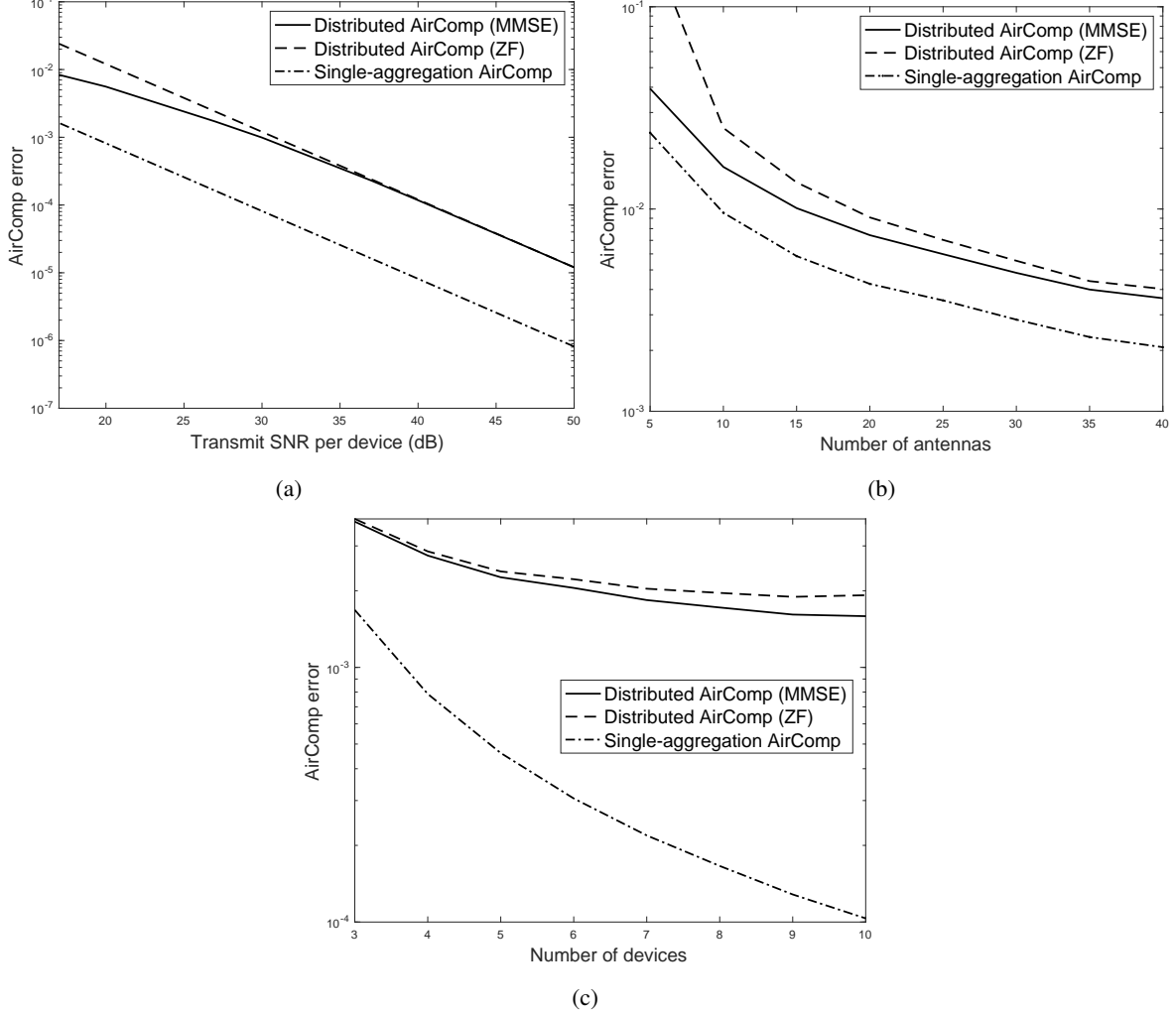


Fig. 3. Comparison of AirComp error between distributed AirComp with MMSE and ZF beamforming and single-aggregation AirComp for (a) varying transmit SNR, (b) a varying number of transmit antennas  $N_t$ , and (c) a varying number of devices  $K$ . The default value of  $K = 5$  is used.

communication latency given a number of rounds as the accumulated latency from the start of the task to the current round. The test accuracy of the considered schemes are compared in Fig. 4 as a function of communication latency. First, by comparing sub-figures in the same row, the dramatic convergence-speed acceleration achieved by distributed AirComp is aligned with the latency comparison in Fig. 2. Second, in terms of converged test accuracy, distributed AirComp with ZF multicast beamforming performs similarly as the benchmarking schemes. Third, in the context of distributed AirComp, MMSE beamforming design is observed from the left-most subfigure of Fig. 4(a) to suffer significant accuracy loss (i.e., 15%) w.r.t. the ZF counterpart due to the bias in subgradient estimation as discussed in Section V. The loss is significant in

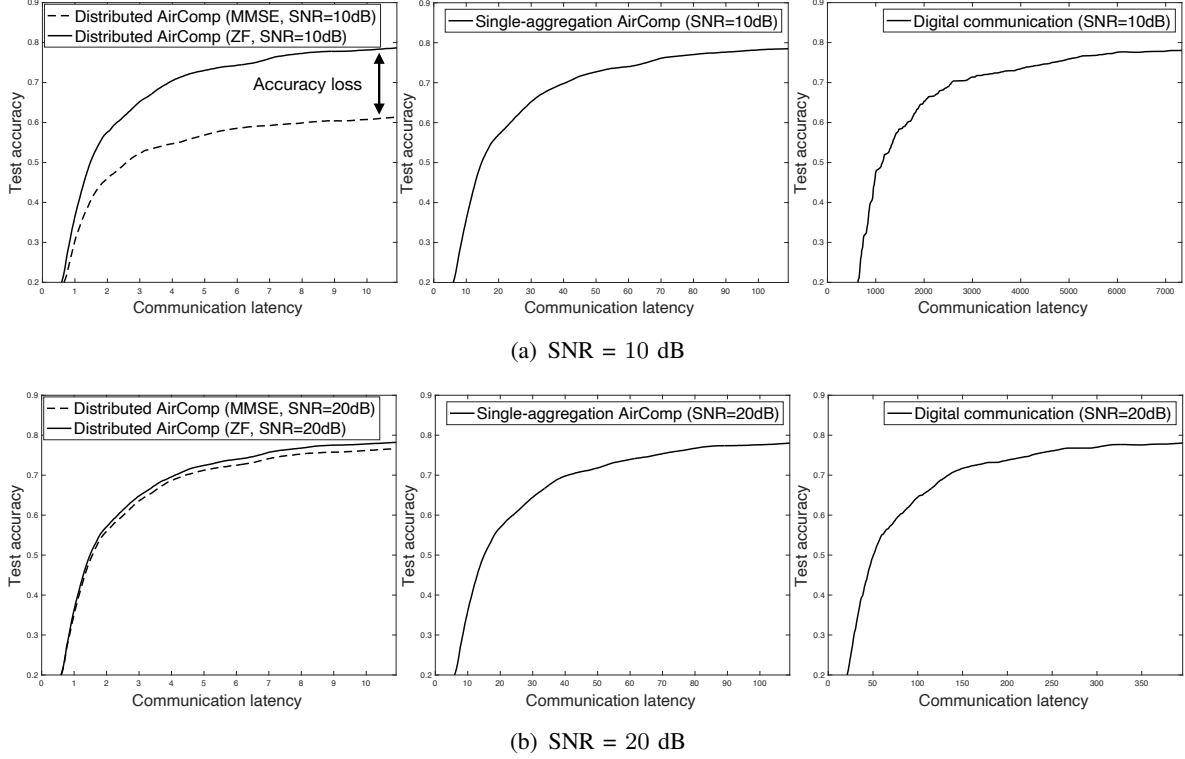


Fig. 4. Test-accuracy comparison between (from left to right) distributed AirComp, single-aggregation AirComp, and digital communication for varying communication latency for (a) a medium SNR (SNR = 10 dB) or (b) a high SNR (SNR = 20 dB).

the regime of low-to-medium SNRs but vanishes at high SNRs as observed from the left-most subfigure of Fig. 4(b).

## VII. CONCLUDING REMARKS

To overcome the communication bottleneck in the deployment of distributed optimization in a distributed network, we have proposed the framework of distributed AirComp that realizes a one-step distributed aggregation of the local states of all devices. The framework features the seamless integration of simultaneous multicast beamforming of all devices and their full-duplex communication, which makes it possible to support multiple concurrent over-the-air aggregation processes at devices. This has led to dramatic communication-latency reduction when there are many devices and thus provides a promising solution to data intensive applications such as distributed machine learning and high-mobility applications, such as drone swarm or vehicle platooning.

This work points to a number of directions warranting follow-up investigations. It is interesting to extend the current single-stream transmission to distributed AirComp with spatial multiplex-

ing, which can further shorten communication latency. On the other hand, the performance of distributed AirComp can be enhanced by distributed resource allocation such as adaptive power control and broadband transmission. Last, particularization of the current design to the area of distributed machine learning provides abundance of cross-disciplinary research opportunities, for example, distributed reinforcement learning with distributed AirComp.

## APPENDIX

### A. Distributed Optimization Algorithm

Consider distributed optimization in a distributed network described by an undirected graph denoted as  $\mathbb{G} = (\mathbb{V}, \mathbb{E})$ , where  $\mathbb{V}$  represents the set of  $K$  nodes (e.g., edge devices) and  $\mathbb{E}$  represents the set of edges. Each edge, say the one connecting the  $k$ th and  $\ell$ th node, is assigned a non-negative weight denoted as  $P_{k\ell}$ . Then the graph structure can be specified by the  $K \times K$  weight matrix  $\mathbf{P}$  with  $(k, \ell)$ th element being  $P_{k\ell}$ . In particular, two nodes  $k \neq \ell$  are connected (i.e.,  $(k, \ell) \in \mathbb{E}$ ) if and only if  $P_{k\ell} > 0$  or otherwise  $P_{k\ell} = 0$ . The matrix  $\mathbf{P}$  is constrained to be doubly stochastic, namely that  $\sum_{\ell=1}^K P_{k\ell} = \sum_{k=1}^K P_{k\ell} = 1$ .

Given the distributed network, the problem of distributed optimization in (4) can be solved using the classic iterative algorithm of *distributed dual averaging* described as follows [29], [31]. To begin with, let each device, say device  $k$ , maintain a two-variable state,  $(\mathbf{z}_k(n), \mathbf{x}_k(n))$ , each being a  $D \times 1$  row vector with real elements and  $\mathbf{x}_k(n) \in \mathcal{X}$ , which is the support of the objective function. Moreover, a key component of the algorithm is a proximal function  $\psi : \mathbb{R}^D \rightarrow \mathbb{R}$  that is assumed to be 1-strongly convex with respect to some norm  $\|\cdot\|$ :

$$\psi(\mathbf{y}) \geq \psi(\mathbf{x}) + \langle \nabla \psi(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{1}{2} \|\mathbf{x} - \mathbf{y}\|^2 \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{X}. \quad (26)$$

One example of such a function is  $\psi(\mathbf{x}) = \frac{1}{2} \|\mathbf{x}\|^2$ . For device  $k$ , the state variable  $\mathbf{x}_k(n) \in \mathcal{X}$  represents the local estimate of the optimal solution for (4) at iteration  $n$ . Recall that  $\tilde{\mathbf{g}}_k(n)$  represents the stochastic subgradient of the local loss function as computed at device  $k$  using local data. Then given non-increasing step size  $\{\alpha(n)\}$ , all devices perform simultaneous updating of their states [31]: for all  $k$  and the  $n$  iteration,

$$\mathbf{z}_k(n+1) = (1 - \beta) \mathbf{z}_k(n) + \beta \sum_{\ell \in N(k)} P_{k\ell} \mathbf{z}_\ell(n) + \tilde{\mathbf{g}}_k(n), \quad (27)$$

$$\mathbf{x}_k(n+1) = \Pi_{\mathcal{X}}^{\psi}(\mathbf{z}_k(n+1), \alpha(n)), \quad (28)$$

where the constant  $\beta \in (0, 1)$ ,  $N(k)$  represents the neighbourhood of node  $k$  on the graph  $\mathbb{G}$ , and  $\Pi_{\mathcal{X}}^{\psi}$  the projection function defined as

$$\Pi_{\mathcal{X}}^{\psi}(\mathbf{z}_k(n+1), \alpha(n)) := \operatorname{argmin}_{\mathbf{x} \in \mathcal{X}} \left\{ \langle \mathbf{z}_k(n+1), \mathbf{x} \rangle + \frac{1}{\alpha(n)} \psi(\mathbf{x}) \right\} \quad (29)$$

with  $\psi(\cdot)$  being the proximal function that satisfies (26). For ease of notation, define the weight matrix  $\mathbf{W} = (1 - \beta)\mathbf{I} + \beta\mathbf{P}$  with the  $(k, \ell)$ th element denoted as  $W_{k\ell}$ . Then (27) is rewritten as

$$\mathbf{z}_k(n+1) = \sum_{\ell=1}^K W_{k\ell} \mathbf{z}_{\ell}(n) + \tilde{\mathbf{g}}_k(n), \quad (30)$$

Note that the updating in (30) is for node  $k$  to compute a consensus based on the states of the peers. On the other hand, the projection function in (29) yields  $\mathbf{x}_k(n+1)$  that minimizes an averaged first-order approximation of the objective function. It is assumed that  $\psi(\mathbf{x}) \geq 0 \forall \mathbf{x} \in \mathcal{X}$  and  $\psi(\mathbf{0}) = 0$  without loss of generality.

The above updating procedure is iterated until a consensus on the minimum of the objective is reached, as the changes of the states of all devices fall below a given threshold. Specifically, given a threshold  $\varepsilon$ , the convergence criterion can be specified using the expected suboptimality gap [29]:

$$\max_k \mathbb{E} [f(\hat{\mathbf{x}}_k(N)) - f(\mathbf{x}^*)] \leq \varepsilon, \quad (31)$$

where  $\varepsilon$  is a given constant and the expectation in (31) taken over all sources of randomness.

### B. Proof of Lemma 1

First, given  $\sum_{\ell=1}^K \sum_{k \neq \ell}^K (\mathbf{h}_{k\ell}^H \mathbf{p}_k + \mathbf{p}_k^H \mathbf{h}_{k\ell}) \geq 0$ , it is direct to see that the objective function of P1.2 is positive. Denote  $x_{k\ell} = \operatorname{Re}(\mathbf{h}_{k\ell}^H \mathbf{p}_k)$ ,  $y_{k\ell} = \operatorname{Im}(\mathbf{h}_{k\ell}^H \mathbf{p}_k)$  as the real and imaginary parts of  $\mathbf{h}_{k\ell}^H \mathbf{p}_k$ , respectively. Then the numerator and denominator of objective function of Problem (P1.2) can be written as  $A(\mathbf{x}) = 2 \sum_{\ell=1}^K \sum_{k \neq \ell}^K x_{k\ell}$  and  $B(\mathbf{x}, \mathbf{y}) = 2 \sqrt{K\sigma^2 + \sum_{\ell=1}^K \sum_{k \neq \ell}^K (x_{k\ell}^2 + y_{k\ell}^2)}$  respectively. Notice that the eigenvalues of the Hessian matrix of function  $B(\mathbf{x}, \mathbf{y})$  :  $e_1 = \frac{2K\sigma^2}{(K\sigma^2 + \sum_{k \neq \ell}^K (x_{k\ell}^2 + y_{k\ell}^2))^{3/2}}$ ,  $e_2 = e_3 = \dots = e_{2K(K-1)} = \frac{2}{(K\sigma^2 + \sum_{k \neq \ell}^K (x_{k\ell}^2 + y_{k\ell}^2))^{1/2}}$  are all positive numbers, hence the  $B(\mathbf{x}, \mathbf{y})$  is a convex function. Then, since  $A(\mathbf{x})$  is a linear function, one can obtain that the the Problem (P1.2) is a concave-convex fractional program, with its objective function proved to be strictly quasi-concave and pseudo-concave [30]. Therefore, the objective function



of Problem (P1.2) is a positive unimodal function and Problem (P1.2) has an unique optimal solution, and the proof is completed.

### C. Proof of Lemma 2

It is first to prove that for any  $\alpha_1 \leq \alpha_2$ ,  $p_k^*(\alpha_1) \leq p_k^*(\alpha_2)$  by contradiction. Assume there exists  $\alpha_1 \leq \alpha_2$  such that  $p_k^*(\alpha_1) > p_k^*(\alpha_2)$ . Since Problem P1.2 has been proved to be an strictly quasi-concave function, then for  $\alpha_1 \leq \alpha_2$ , the size of feasible region  $\mathcal{P}_1$  for  $\alpha = \alpha_1$  is larger than or equal to the size of  $\mathcal{P}_2$  for  $\alpha = \alpha_2$ , which means we can at least find a set of  $\{\mathbf{p}_k\}$  in  $\mathcal{P}_1$  that makes  $p_k^*(\alpha_1) = p_k^*(\alpha_2)$ . This contradicts the previous assumption and thus for any  $\alpha_1 \leq \alpha_2$ ,  $p_k^*(\alpha_1) \leq p_k^*(\alpha_2)$ .

Moreover, for  $\alpha = 0$ , we have  $\mathbf{p}_k = \mathbf{0}$ ,  $\forall k$ , and thus  $p_k^*(\alpha) = 0$ . Then, for any  $\alpha > 0$ , there must  $\exists \mathbf{p}_k$  such that  $\|\mathbf{p}_k\|^2 > 0$ , which means  $p_k^*(\alpha) > 0$ . To this end,  $p_k^*(\alpha)$  cannot be a constant function for all  $\alpha$ . Then the proof is finished.

### D. Proof of Lemma 3

First, the objective of Problem (P1.4) is linear and the second constraint is a convex set. Then given any  $\alpha$ , left-hand-side of the first constraint can be regarded as a 2-norm function and hence convex. Then the right-hand-side of the first constraint is a linear function. Therefore, the proof is completed.

### E. Proof of Lemma 4 and Lemma 5

We first give the corresponding Lagrange function of Problem (P1.4) as below

$$\begin{aligned} \mathcal{L}(\{\mathbf{p}_k\}, p_{\max}, \lambda, \{v_k\}) = & p_{\max} + \\ & \lambda \left( 2\alpha \sqrt{(K\sigma^2 + \sum_{\ell=1}^K \sum_{k \neq \ell}^K \mathbf{h}_{k\ell}^H \mathbf{p}_k \mathbf{p}_k^H \mathbf{h}_{k\ell})} - \left( \sum_{\ell=1}^K \sum_{k \neq \ell}^K (\mathbf{h}_{k\ell}^H \mathbf{p}_k + \mathbf{p}_k^H \mathbf{h}_{k\ell}) \right) \right) + \sum_{k=1}^K v_k (\mathbf{p}_k^H \mathbf{p}_k - p_{\max}) \end{aligned} \quad (32)$$

where  $\lambda$  and  $\{v_k\}$  are the Lagrangian multipliers. The KKT conditions are given by

$$\left\{ \begin{array}{l} \frac{\partial \mathcal{L}(\{\mathbf{p}_k\}, p_{\max}, \lambda, \{v_k\})}{\partial p_{\max}} = 1 - \sum_{k=1}^K v_k = 0, \\ \frac{\partial \mathcal{L}(\{\mathbf{p}_k\}, p_{\max}, \lambda, \{v_k\})}{\partial \mathbf{p}_k} = \lambda \left( \frac{\alpha \left( \sum_{\ell \neq k}^K \mathbf{h}_{k\ell} \mathbf{h}_{k\ell}^H \right) \mathbf{p}_k}{\sqrt{K\sigma^2 + \sum_{\ell=1}^K \sum_{k \neq \ell}^K \|\mathbf{h}_{k\ell}^H \mathbf{p}_k\|^2}} - \sum_{\ell \neq k}^K \mathbf{h}_{k\ell} \right) + 2v_k \mathbf{p}_k = \mathbf{0}, \forall k, \\ \lambda \left( \alpha \sqrt{K\sigma^2 + \sum_{\ell=1}^K \sum_{k \neq \ell}^K \|\mathbf{h}_{k\ell}^H \mathbf{p}_k\|^2} - \left( \sum_{\ell=1}^K \sum_{k \neq \ell}^K (\mathbf{h}_{k\ell}^H \mathbf{p}_k + \mathbf{p}_k^H \mathbf{h}_{k\ell}) \right) \right) = 0, \\ v_k (\mathbf{p}_k^H \mathbf{p}_k - p_{\max}) = 0, \forall k. \end{array} \right. \quad (33)$$

In (33), the first condition indicates  $\exists k, v_k \neq 0$ . This reveals that for any  $\alpha$ , at least one device transmits with power  $p^*(\alpha)$ . Then by Lemma 2, one can conclude that for the maximum aligned fraction  $\alpha$ , at least one device transmits with the maximum power  $P_0$ .

Next, together with the second and the forth conditions, one can have  $\lambda \neq 0$ . By the second condition, let a constant  $c_0 = \frac{\alpha}{\sqrt{K\sigma^2 + \sum_{\ell=1}^K \sum_{k \neq \ell}^K \|\mathbf{h}_{k\ell}^H \mathbf{p}_k\|^2}}$ , we have

$$[c_0 \lambda (\mathbf{H}_k \mathbf{H}_k^H) + 2v_k \mathbf{I}] \mathbf{p}_k = \mathbf{H}_k \mathbf{1}_{(K-1)}. \quad (34)$$

By denoting  $\mu_k = \frac{2v_k}{c_0 \lambda}$ , then the desired results are obtained.

#### F. Proof of Proposition 1

The problem of finding the smallest norm solution for  $\|\mathbf{H}_k^H \mathbf{p}_k - \sqrt{\eta} \mathbf{1}_{(K-1)}\|^2 = 0$  is equivalent to the following optimization problem:

$$\begin{array}{ll} \min_{\mathbf{p}_k} & \|\mathbf{p}_k\|^2 \\ \text{s.t.} & \mathbf{H}_k^H \mathbf{p}_k = \sqrt{\eta} \mathbf{1}_{(K-1)}. \end{array}$$

This problem is convex since both the objective and feasible region are convex. By reusing Lagrange multiplier  $\lambda$ , the Lagrange function can be written as

$$\mathcal{L}(\mathbf{p}_k, \lambda) = \|\mathbf{p}_k\|^2 + \lambda (\mathbf{H}_k^H \mathbf{p}_k - \sqrt{\eta} \mathbf{1}_{(K-1)}) . \quad (35)$$

The KKT conditions are given by

$$\begin{cases} \frac{\partial \mathcal{L}(\mathbf{p}_k, \lambda)}{\partial \mathbf{p}_k} = 2\mathbf{p}_k + \mathbf{H}_k \lambda^H = \mathbf{0}, \\ (\mathbf{H}_k^H \mathbf{p}_k - \sqrt{\eta} \mathbf{1}_{(K-1)}) = \mathbf{0}. \end{cases} \quad (36)$$

Take the first condition into the second one, one can get the expression of  $\lambda$ . Then take  $\lambda$  back to the first condition, the desired result is obtained.

### G. Proof of Lemma 6

From (19), the average state in round  $n$  is given as

$$\begin{aligned} \bar{\mathbf{z}}(n+1) &= \frac{1}{K} \sum_{k=1}^K \mathbf{z}_k(n+1) = \frac{1}{K} \sum_{k=1}^K \left( \sum_{\ell=1}^K W_{k\ell}(n) \mathbf{z}_\ell(n) + \hat{\mathbf{g}}_k(n) \right), \\ &= \bar{\mathbf{z}}(n) + \frac{1}{K} \sum_{k=1}^K \hat{\mathbf{g}}_k(n). \end{aligned} \quad (37)$$

Define the matrix  $\Phi(n, s) = \mathbf{W}^{n-s+1}$ . Then, for the state update  $\mathbf{z}_k(n+1)$  at device  $k$ , it is expanded as follows

$$\mathbf{z}_k(n+1) = \sum_{\ell=1}^K [\Phi(n, s)]_{k\ell} \mathbf{z}_\ell(s) + \sum_{r=s+1}^n \left( \sum_{\ell=1}^K [\Phi(n, r)]_{k\ell} \hat{\mathbf{g}}_\ell(r-1) \right) + \hat{\mathbf{g}}_k(n), \quad (38)$$

where  $[\Phi(n, s)]_{k\ell}$  is the  $\ell$ -th entry of the  $k$ -th column of  $\Phi(n, s)$ . Since the initial state  $\mathbf{z}_k(0) = \mathbf{0}$ , using (37) and (38) yields

$$\bar{\mathbf{z}}(n) - \mathbf{z}_k(n) = \sum_{s=1}^{n-1} \sum_{\ell=1}^K \left( \frac{1}{K} - [\Phi(n-1, s)]_{k\ell} \right) \hat{\mathbf{g}}_\ell(s-1) + \left( \frac{1}{K} \sum_{\ell=1}^K (\hat{\mathbf{g}}_\ell(n-1) - \hat{\mathbf{g}}_k(n-1)) \right). \quad (39)$$

Based on (22),  $\xi^2 = \Omega^2 + \frac{\beta^2 \max_n \text{MSE}(n)}{K}$  denotes an upper bound on the second moment of  $\hat{\mathbf{g}}_k(n)$ .

Then by Jensen's inequality, one has  $(\mathbb{E} [\|\hat{\mathbf{g}}_k(n)\|])^2 \leq \mathbb{E} [\|\hat{\mathbf{g}}_k(n)\|^2] \leq \xi^2$  for all  $k$ . Hence

$$\mathbb{E} \|\bar{\mathbf{z}}(n) - \mathbf{z}_k(n)\|_* \leq \sum_{s=1}^{n-1} \xi \left\| [\Phi(n-1, s)]_k - \frac{\mathbf{1}}{K} \right\|_1 + 2\xi. \quad (40)$$

Following steps similar to [31], we separate the sum in (40) into two terms by a cutoff point  $\hat{n} = n - \frac{\log N \sqrt{K}}{\beta \log \lambda^{-1}}$ , where  $\lambda_2 = \max \{\lambda_2(\mathbf{P}), -\lambda_K(\mathbf{P})\}$  is the second-largest magnitude of eigenvalues

of  $\mathbf{P}$ ,

$$\mathbb{E} \|\bar{\mathbf{z}}(n) - \mathbf{z}_k(n)\|_* \leq \sum_{s=1}^{\hat{n}} \xi \left\| [\Phi(n-1, s)]_k - \frac{\mathbf{1}}{K} \right\|_1 + \sum_{s=\hat{n}+1}^{n-1} \xi \left\| [\Phi(n-1, s)]_k - \frac{\mathbf{1}}{K} \right\|_1 + 2\xi. \quad (41)$$

Then for  $s \leq \hat{n}$ , we have  $\|[\Phi(n-1, s)]_k - \mathbf{1}/K\|_1 \leq 1/K$  and for larger  $s$ , we relax this term via a more loose bound  $\|[\Phi(n-1, s)]_k - \mathbf{1}/K\|_1 \leq 2$ . Taking these two bounds into (41) together with the fact  $\log \lambda_2^{-1} \geq 1 - \lambda_2$ , one can obtain the desired result.

#### H. Proof of Theorem 1

The proof is close to the convergence proof for distributed dual averaging [31] with some modifications. First, an useful lemma is introduced as follows.

**Lemma 7.** [31, lemma 8] Let  $\{\mathbf{x}_k(n)\}$  and  $\{\mathbf{z}_k(n)\}$  be the primal and dual variables, then the expected suboptimality gap for both ZF and MMSE beamforming cases is bounded as

$$\begin{aligned} \mathbb{E} [f_k(\hat{\mathbf{x}}_k(N)) - f(\mathbf{x}^*)] &\leq \frac{1}{N\alpha(N)} \psi(\mathbf{x}^*) + \frac{\xi^2}{2N} \sum_{n=1}^N \alpha(n-1) \\ &+ \frac{L + \xi}{NK} \sum_{n=1}^N \sum_{k=1}^K \alpha(n) \mathbb{E} [\|\bar{\mathbf{z}}(n) - \mathbf{z}_k(n)\|] + \frac{L}{N} \sum_{n=1}^N \alpha(n) \mathbb{E} [\|\bar{\mathbf{z}}(n) - \mathbf{z}_k(n)\|] \\ &+ \mathbb{E} \left[ \frac{1}{NK} \sum_{n=1}^N \sum_{k=1}^K \langle \mathbf{g}_k(n) - \hat{\mathbf{g}}_k(n), \mathbf{x}_k(n) - \mathbf{x}^* \rangle \right] \end{aligned} \quad (42)$$

where  $\xi = \sqrt{\Omega^2 + \frac{\beta^2 \max_n \text{MSE}(n)}{K}}$ , and  $\|\cdot\|$  represents the  $\ell_2$ -norm that is its own dual.

For the ZF beamforming case, by (20), one can find that the last term of (42) is equal to zero. By taking the upper bound in Lemma 6 into (42) and with  $\psi(\mathbf{x}^*) \leq R^2$ ,  $\alpha(n) = \frac{R\sqrt{1-\lambda_2}}{4\xi\sqrt{n}}$ , the convergence for ZF beamforming in Theorem 1 can be obtained. However, for the MMSE beamforming case, by (21), the last term of (42) is nonzero due to the biased gradient.

Nevertheless, this term can be bounded as

$$\begin{aligned}
\mathbb{E} \left[ \frac{1}{NK} \sum_{n=1}^N \sum_{k=1}^K \langle \mathbf{g}_k(n) - \hat{\mathbf{g}}_k(n), \mathbf{x}_k(n) - \mathbf{x}^* \rangle \right] &\leq \mathbb{E} \left[ \frac{1}{NK} \sum_{n=1}^N \sum_{k=1}^K \|\mathbf{g}_k(n) - \hat{\mathbf{g}}_k(n)\| \|\mathbf{x}_k(n) - \mathbf{x}^*\| \right] \\
&\leq \frac{1}{NK} \sum_{n=1}^N \sum_{k=1}^K \mathbb{E} [\|\Delta_k(n)\|] \mathbb{E} [\|\mathbf{x}_k(n) - \mathbf{x}^*\|] \\
&\leq \frac{\|\mathbf{x}^*\|}{N} \sum_{n=1}^N \sqrt{\text{MSE}(n)/K}.
\end{aligned} \tag{43}$$

The last inequality follows from  $\mathbb{E} [\|\mathbf{x}_k(n) - \mathbf{x}^*\|] \leq \mathbb{E} [\|\mathbf{x}_k(0) - \mathbf{x}^*\|]$  and  $\mathbf{x}_k(0) = 0, \forall k$ . Taking this into (42), then Theorem 1 is proved.

## REFERENCES

- [1] M. G. Rabbat and R. D. Nowak, “Quantized incremental algorithms for distributed optimization,” *IEEE J. Sel. Areas Commun.*, vol. 23, no. 4, pp. 798–808, 2005.
- [2] 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,” *Proc. IEEE*, vol. 107, no. 8, pp. 1738–1762, 2019.
- [3] 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, 2021.
- [4] W. Y. B. Lim, N. C. Luong, D. T. Hoang, Y. Jiao, Y.-C. Liang, Q. Yang, D. Niyato, and C. Miao, “Federated learning in mobile edge networks: A comprehensive survey,” *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 2031–2063, 2020.
- [5] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, “A joint learning and communications framework for federated learning over wireless networks,” *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 269–283, 2020.
- [6] Q. Zeng, Y. Du, K. Huang, and K. K. Leung, “Energy-efficient radio resource allocation for federated edge learning,” in *Proc. IEEE Int. Conf. Commun. Workshops (ICC WKSHPs)*, Dublin, Ireland, Jun. 7–11, 2020.
- [7] G. Zhu, Y. Du, D. Gündüz, and K. Huang, “One-bit over-the-air aggregation for communication-efficient federated edge learning: Design and convergence analysis,” *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 2120–2135, 2020.
- [8] Y. Du, S. Yang, and K. Huang, “High-dimensional stochastic gradient quantization for communication-efficient edge learning,” *IEEE Trans. Signal Process.*, vol. 68, pp. 2128–2142, 2020.
- [9] H. H. Yang, Z. Liu, T. Q. Quek, and H. V. Poor, “Scheduling policies for federated learning in wireless networks,” *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 317–333, 2019.
- [10] J. Ren, Y. He, D. Wen, G. Yu, K. Huang, and D. Guo, “Scheduling for cellular federated edge learning with importance and channel awareness,” *IEEE Trans. Wireless Commun.*, vol. 19, no. 11, pp. 7690–7703, 2020.
- [11] M. Chen, D. Gündüz, K. Huang, W. Saad, M. Bennis, A. V. Feljan, and H. V. Poor, “Distributed learning in wireless networks: Recent progress and future challenges,” *IEEE J. Sel. Areas Commun.*, 2021.
- [12] 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.
- [13] M. M. Amiri and D. Gündüz, “Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air,” *IEEE Trans. Signal Process.*, vol. 68, pp. 2155–2169, 2020.

- [14] 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.
- [15] 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, 2019.
- [16] H. Xing, O. Simeone, and S. Bi, "Federated learning over wireless device-to-device networks: Algorithms and convergence analysis," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 12, pp. 3723–3741, 2021.
- [17] Y. Shi, Y. Zhou, and Y. Shi, "Over-the-air decentralized federated learning," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Melbourne, Australia, Jul. 12-20, 2021.
- [18] E. Ozfatura, S. Rini, and D. Gündüz, "Decentralized sgd with over-the-air computation," in *2020 IEEE Global Communications Conference (GLOBECOM)*, Taipei, Taiwan, Dec. 7-11, 2020.
- [19] S. Savazzi, S. Kianoush, V. Rampa, and M. Bennis, "A joint decentralized federated learning and communications framework for industrial networks," in *IEEE Int. Workshop Comput. Aided Model. Des. Commun. Links Netw. (CAMAD)*, Pisa, Italy, Sept. 14-16, 2020.
- [20] A. C. Cirik, Y. Rong, and Y. Hua, "Achievable rates of full-duplex mimo radios in fast fading channels with imperfect channel estimation," *IEEE Trans. Signal Process.*, vol. 62, no. 15, pp. 3874–3886, 2014.
- [21] N. D. Sidiropoulos, T. N. Davidson, and Z.-Q. Luo, "Transmit beamforming for physical-layer multicasting," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 2239–2251, 2006.
- [22] O. Mehanna, N. D. Sidiropoulos, and G. B. Giannakis, "Joint multicast beamforming and antenna selection," *IEEE Trans. Signal Process.*, vol. 61, no. 10, pp. 2660–2674, 2013.
- [23] Y. Sun and K. R. Liu, "Transmit diversity techniques for multicasting over wireless networks," in *IEEE Wirel. Commun. Netw. Conf. (WCNC)*, Atlanta, GA, USA, March 21-25, 2004.
- [24] W. Lee, H. Park, H.-B. Kong, J. S. Kwak, and I. Lee, "A new beamforming design for multicast systems," *IEEE Trans. Veh. Technol.*, vol. 62, no. 8, pp. 4093–4097, 2012.
- [25] X. Cao, G. Zhu, J. Xu, and K. Huang, "Optimized power control for over-the-air computation in fading channels," *IEEE Trans. Wireless Commun.*, vol. 19, no. 11, pp. 7498–7513, 2020.
- [26] 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.
- [27] Z. Zhang, X. Chai, K. Long, A. V. Vasilakos, and L. Hanzo, "Full duplex techniques for 5g networks: self-interference cancellation, protocol design, and relay selection," *IEEE Commun. Mag.*, vol. 53, no. 5, pp. 128–137, 2015.
- [28] N. Zhang and M. Tao, "Gradient statistics aware power control for over-the-air federated learning," *IEEE Trans. Wireless Commun.*, vol. 20, no. 8, pp. 5115–5128, 2021.
- [29] J. Duchi, A. Agarwal, and M. Wainwright, "Dual Averaging for Distributed Optimization: Convergence Analysis and Network Scaling," *IEEE Trans. Automat. Contr.*, vol. 57, no. 3, pp. 592–606, Mar. 2012.
- [30] S. Schaible, "Fractional programming. I, duality," *Manag. Sci.*, vol. 22, no. 8, pp. 858–867, 1976.
- [31] R. Saha, S. Rini, M. Rao, and A. Goldsmith, "Decentralized optimization over noisy, rate-constrained networks: Achieving consensus by communicating differences," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 2, pp. 449–467, 2022.