Over-the-air Clustered Wireless Federated Learning

Ayush Madhan-Sohini*
ECE, IIITD
New Delhi, India
ayush19156@iiitd.ac.in

Divin Dominic*
ECE, IIITD
New Delhi, India
divin19163@iiitd.ac.in

Nazreen Shah
ECE, IIITD
New Delhi, India
nazreens@iiitd.ac.in

Ranjitha Prasad
ECE, IIITD
New Delhi, India
ranjitha@iiitd.ac.in

Abstract—Privacy and bandwidth constraints have led to the use of federated learning (FL) in wireless systems, where training a machine learning (ML) model is accomplished collaboratively without sharing raw data. While using bandwidth-constrained uplink wireless channels, over-the-air (OTA) FL is preferred since the clients can transmit parameter updates simultaneously to a server. A powerful server may not be available for parameter aggregation due to increased latency and server failures. In the absence of a powerful server, decentralised strategy is employed where clients communicate with their neighbors to obtain a consensus ML model while incurring huge communication cost. In this work, we propose the OTA semi-decentralised clustered wireless FL (CWFL) and CWFL-Prox algorithms, which is communication efficient as compared to the decentralised FL strategy, while the parameter updates converge to global minima as $O(1/T)$ for each cluster. Using the MNIST and CIFAR10 datasets, we demonstrate the accuracy performance of CWFL is comparable to the central-server based COTAF and proximal constraint based methods, while beating single-client based ML model by vast margins in accuracy.

Index Terms—Federated Learning, Over-the-air, Wireless, Clustering, Decentralised Learning

I. INTRODUCTION

The proliferation of wireless devices in our daily lives has led to the rapid evolution of capabilities of wireless technologies towards greater network coverage, higher throughput, and lower latency while supporting high user densities. Naturally, an array of wireless protocols and standards cater to coexistence and interoperability, further leading to data-driven machine learning (ML) solutions in wireless systems.

Conventional ML approaches learn the model at a central entity (referred to as a server) irrespective of the source of data, i.e., if the source is an edge device (referred to as a client), then the data at the client is transmitted to the server for model training. Transmission of raw data is not always feasible in wireless communications due to bandwidth and privacy constraints. This naturally triggers the idea of distributed learning approaches that retain data at clients. Federated learning (FL) is one such distributed approach that is particularly suited to tackle the aforementioned challenges in wireless communications. In conventional FL, clients train a parametric ML model, while the server periodically collects these models and aggregates the parameters to form a global model. Subsequently, these models are broadcasted back to the clients for downstream tasks.

We acknowledge research grants from MeITY, Power grant from DST SERB, and PMRF research grant from the govt. of India.

While employing FL in wireless communications, the base station plays the role of the server [1], which implies that the parameters from the clients are sent over a resource-constrained uplink channel. Furthermore, the multi-user nature of FL also necessitates the use of orthogonal time and frequency resources. For example, in frequency division multiplexing (FDM), each user is assigned a dedicated bandwidth, leading to a bandwidth deficit and an increase in energy consumption if there are numerous participating clients. Hence, repeated communication of the local models to the server entails a severe load on this uplink channel. Several strategies, such as sparsification and quantization, are adopted for efficiently communicating local models [2].

A popular paradigm for efficient communication over a common uplink multiple access channel (MAC) is over-the-air FL (OTA-FL) strategy. OTA-FL allows clients to simultaneously transmit updates using analog signaling over the uplink channel in a non-orthogonal manner, hence optimizing available temporal and spectral resources [3]. Several variations of the OTA-FL algorithm have been proposed [1], [4]. In [3], the authors propose the COTAF algorithm, which uses a novel pre-coding technique to facilitate high throughput OTA-FL over wireless channels. This is the first-of-its-kind technique that achieves $O(1/T)$ convergence in parameter updates in the presence of noise.

Since its inception, centralised FL where the server creates the global model has been popular in the literature owing to superior performance and convergence guarantees. However, such a centralised approach is vulnerable to increased latency due to bottlenecks and server failures. For example, in D2D communications, a base station-centric architecture is not feasible due to connectivity and computational constraints.

In such scenarios, a decentralised wireless FL architecture is lucrative [5] since the clients can communicate with each other directly. In decentralised learning, each client performs local SGD steps using a mini-batch of data drawn from its local dataset and communicates the updated model parameters to its neighboring clients. Each client incorporates the information received from the neighbors with its local information using a mixing matrix to achieve consensus. However, using the mixing matrix forces decentralised techniques to rely on communication over orthogonal uplink channels [5], i.e., without a server, communication efficiency is poor and scales as $O(K^2)$. Hence, it is of utmost importance to devise decentralised FL...
schemes that are communication efficient while employing wireless uplink channels.

**Contributions:** We propose a novel communication-efficient semi-decentralised OTA-FL strategy, which we refer to as clustered wireless federated learning (CWFL). CWFL involves decentralised data-agnostic clustering of clients using the following steps: (a) Representative members (mid-tier computational devices) of each cluster called cluster-heads build a cluster-level ML model using OTA communication, (b) Cluster-heads exchange model updates among each other in a decentralised fashion to arrive at a consensus ML model. Briefly, the contributions are as follows:

- CWFL has an improved communication complexity of \( \mathcal{O}(C^2) \), where \( C \) is the number of cluster-heads.
- CWFL with a proximal constraint (CWFL-Prox) has improved robustness to statistical heterogeneity in the presence of noise.
- In the presence of statistically heterogeneous data, CWFL achieves \( \mathcal{O}(1/T) \) convergence, similar to counterparts [3] where the server is present.

There are several real-world use-cases where the decentralised set-up is crucial. Examples include D2D wireless networks where the agents are connected in a decentralized topology, D2D relay clustering system [6], wireless fog networks where the distributed computing paradigm encapsulates communication among edge devices such as local area servers, UAVs and cloud servers, and the wireless sensor network-inspired-IoT networks [7]. Using the MNIST and CIFAR10 datasets, we demonstrate that CWFL and CWFL-Prox frameworks perform similar (with respect to model accuracy) to the centralized FL technique [3]. We also argue that CWFL-type architecture is more preferred to COTAF (in terms of accuracy) under certain conditions. By design, our scheme has a lower communication complexity than the decentralised framework [5]. A non-trivial aspect of novelty is that of convergence, which is achieved using the proposed novel power control mechanism.

## II. SYSTEM MODEL

We consider a wireless multi-user system with \( K \) clients. Each participating client has access to a data set \( D_k \), which consists of \( N_k \) instances, i.e., the total number of instances is given by \( N = \sum_{k=1}^{K} N_k \). In supervised learning, the dataset at the \( k \)-th client, \( D_k \), consists of data samples given as a set of input-output pairs \( \{z_i, y_i\} \in D_k \) for \( i = [N_k] \), where \( z_i \in \mathbb{R}^m \) and \( y_i \in \mathbb{R} \) is the label for the sample \( z_i \). This data may be generated at the clients via onboard sensors or interactions with mobile apps. A typical learning problem is to find the model parameter \( \theta_k \in \mathbb{R}^d \) by optimizing the empirical loss function on, \( D_k \), given by

\[
    f_k(\theta_k) \triangleq \frac{1}{N_k} \sum_{i \in D_k} l(z_i; \theta_k),
\]

where \( l(z_i; \theta_k) \) is the loss per instance. In the conventional server-based FL framework, the global ML model is assumed to be parameterised by \( \theta_k \in \mathbb{R}^d \) itself, which leads to a global objective function given as

\[
    \min_{\theta_k} F(\theta_k) \triangleq \frac{1}{K} \sum_{k=1}^{K} f_k(\theta_k).
\]

### A. Wireless Federated Learning

In wireless communications, a base-station based server communicates with \( K \) wireless clients. The parameter updates from the clients to the server, and global synchronization from server to client happens over the resource constrained uplink and downlink channels, respectively. Typically, downlink transmission is assumed to be error-free [3] owing to availability of sophisticated error-control methods. Further, it is assumed that communication takes place over \( T \) rounds, where the number of channel uses in each round depends upon the FL strategy. For \( t \in [T] \) and \( k \in [K] \), the signal received at the server via the uplink AWGN channel is given by

\[
    y^t = \sum_{k=1}^{K} x^t_k + w^t, \quad 0 < E[\|x^t_k\|^2] \leq P.
\]

The additive noise at the server is modeled as \( w^t \sim \mathcal{N}(0, \sigma^2 I_d) \), where \( \sigma^2 \) is the noise variance. We assume that the input is power constrained and \( P \) represents the available transmission power, i.e., \( x^t_k \) is given by

\[
    x^t_k = \sqrt{p^t(\theta^t_k - \theta^t)}, \quad p^t \leq \frac{E}{\max_k E[\|\theta^t_k - \theta^t\|^2]},
\]

where \( \theta^t_k \) is the parameter update at the \( k \)-th client for the \( t \)-th communication round, and \( \theta^t \) is the global parameter update upto \( t \). Here, \( p^t \) is a pre-coding factor that scales the model parameters as \( t \) progresses such that the power constraint stated in (3) is satisfied in an expected sense.

### B. Over-the-Air Noisy Federated Averaging

Federated averaging is one of the most widely adopted algorithms [8], where the global model is learnt at the server by aggregating the client parameters. First, the server shares \( F(\theta) \) in (3) is satisfied in an expected sense. From (5), it is evident that after decoding, OTA leads to a noisy version of the global update is shared with the clients. The above steps conclude one communication round. The \( K \) clients convey their parameter updates to the server during pre-designated aggregation time-slots \( t \in \mathcal{H} \) using incremental updates, as given in (3). These updates are aggregated at the server to obtain the global update, and this global update is shared with the clients. The above steps conclude one communication round, and the steps are repeated for many communication rounds until the model converges.

In the context of OTA, for \( t \in \mathcal{H} \), the received signal is given by (3), and the transmitted signal is as given in (4) for \( t \in [T] \). The decoding rule used at the server is given by

\[
    \theta^t = \frac{y^t}{K \sqrt{p^t}} + \theta^{t-E} = \frac{1}{K} \sum_{k=1}^{K} \theta^t_k + \tilde{w}^t,
\]

where \( \tilde{w}^t \sim \mathcal{N}(0, \frac{\sigma^2}{K_2} I_d) \). Note that \( \theta^{t-E} \) is the previous global update referred to as \( \theta^t \) in (4). From (5), it is evident that after decoding, OTA leads to a noisy version of the federated averaging-based update at the server.

In the presence of statistical heterogeneity at the clients, model learnt at the server tends to vary drastically in each
round. In order to render FL robust to heterogeneity, Fedprox [9] introduces the proximal term as a constraint on the local objective, \( f_k(\theta) \), i.e., distributed optimization in (2), \( f_k(\theta) \) is replaced by a proximal term \( f_k^\ell(\theta) \triangleq f_k(\theta) + \frac{\nu}{2} \| \theta - \bar{\theta}_t \|^2 \). The constraint ensures that the local update lies close (in a sphere) to the previous global update \( \bar{\theta}_t \).

### III. Clustered Wireless Federated Learning

We present the novel CWFL strategy to address the issue of communication complexity in the decentralised setting. The fundamental idea of CWFL is to cluster the clients to distribute the tasks of a server to a subset of clients, which we refer to as cluster-heads. Cluster-heads are the clients that are mid-tier computational devices. The clustering process is initiated by choosing \( C \) cluster-heads, and clients are randomly assigned to each cluster-head such that clusters consist of non-overlapping set of clients. The clients of the \( c \)-th cluster are indexed by \( k_c \in [K_c] \), such that \( \sum_{c=1}^C K_c = K \). CWFL is implemented in a hierarchical fashion, where the first phase of a given communication round is the uplink transmission, i.e., from the clients in each cluster to their respective cluster-heads. This phase is OTA; hence, each cluster-head takes one channel use for communication, resulting in a total of \( C \) channel uses. In the second phase, the cluster-heads communicate among each other in a decentralised manner. CWFL requires \( C(C-1) \) channel uses to obtain the consensus update in each communication round \( t \) since every cluster-head needs one channel-use each for communication with \( C-1 \) possible neighbors, in the worst case. Hence, the total number of channel uses per communication round scales as \( O(C^2) \) unlike the decentralised strategy in [5] where the communication complexity scales as \( O(K^2) \).

In phase 1 the input signal at each client is given by (4), where \( \theta_t \) is the cluster-level global parameter update. The corresponding channel output at the \( c \)-th cluster-head is given by \( y_c^t \in \mathbb{R}^d \) for \( t \in \mathcal{H} \) is given by

\[
y_c^t = \frac{\nu}{\sqrt{s_c}} \sum_{k=1}^{K_c} x_k^t + w_c^t = \sqrt{\frac{\nu}{s_c}} \sum_{k=1}^{K_c} (\theta_{k}^t - \theta_c^{t-E}) + w_c^t, \tag{6}
\]

where \( w_c^t \sim \mathcal{N}(0, \sigma_c^2 I_d) \) is the additive noise at the cluster head \( c \). As compared to (5), the above expression performs OTA parameter update at the cluster level. Given \( \tilde{w}_c^t \sim \mathcal{N}(0, \frac{\nu}{\sqrt{s_c}} I_d) \), the resulting parameter update at the \( c \)-th cluster-head is given by

\[
\bar{\theta}_c^t = \frac{1}{K_c \sqrt{s_c}} y_c^t + \theta_c^{t-E} = \frac{1}{K_c} \sum_{k=1}^{K_c} \theta_k^t + \tilde{w}_c^t. \tag{7}
\]

Since direct communication with the server is hindered, a decentralised learning architecture is adopted. We assume a symmetric doubly stochastic mixing matrix \( W \in \mathbb{R}^{C \times C} \) where \( W(c, j) = W(j, c) \) and \( W(c, c) = 0 \) for \( 1 \leq c, j \leq C \). The entries of \( W \) encode the extent to which cluster-head \( c \) can affect cluster-head \( j \), while \( W(c, j) = 0 \) implies that cluster-heads \( c \) and \( j \) are disconnected. The \( c \)-th cluster-head transmits \( s_c^t \in \mathbb{R}^d \) given by

\[
s_c^t = \sqrt{q_c^t} \bar{\theta}_c^t, \quad \text{where } q_c^t = \max_{\mathcal{H}} \frac{P_s}{\lambda_c} \| \bar{\theta}_c^t \|^2. \tag{8}
\]

Here, \( q_c^t \) is a pre-coding factor that scales the model parameters so that the power constraint given by \( 0 \leq \mathbb{E}[\| s_c^t \|^2] \leq P_s \) is satisfied for every cluster-head. The power constraint is applied on the expected value of the transmit signal, as given in (8). In the decentralised setting, we set the parameter update shared by the neighbors of the \( c \)-th cluster-head, \( r_c^t \in \mathbb{R}^d \) as

\[
r_c^t = \sum_{j=1}^C W(c, j) s_j^t + v_c^t, \tag{9}
\]

for all \( c, j \in [C] \) and \( W(c, c) = 0 \). Note that computing (9) at the \( c \)-th cluster-head requires \( C-1 \) channel uses. The received signal (9) at every cluster-head is corrupted by noise, which is cumulatively represented as \( v_c^t \sim \mathcal{N}(0, \kappa_c^2 I_d) \) where \( \kappa_c^2 \) is the variance of the additive noise as derived in Lemma 1.

#### Algorithm 1 Clustered Wireless FL

**Input:** Dataset \( \mathcal{D}_k \) at the \( k \)-th client  
**Output:** Consensus parameter \( \bar{\theta}_c^t \) for all \( c \)

1. for \( t \in [T] \) do  
2. if \( t \in \mathcal{H} \) then  
3. Obtain \( \bar{\theta}_c^t \) and \( \theta_c^{t-E} \) (7) and (10), for all \( c \)  
4. Cluster-head broadcasts \( \theta_c^t \) to \( \forall k_c \in [K_c] \), for all \( c \)  
5. if \( t \notin \mathcal{H} \) then  
6. Perform local SGD on \( \theta_c^t \) for all \( c \)  
return Consensus parameter \( \bar{\theta}_c^t \) for all \( c \).

#### Lemma 1. The effective distribution of \( \bar{\theta}_c^t \) is given by \( \bar{\theta}_c^t \sim \mathcal{N}(0, \kappa_c^2 I_d) \), where \( \kappa_c^2 = \sum_{j=1}^C W(c, j) \sigma_j^2 I_d \).

In order to recover the consensus parameter update at the \( c \)-th cluster head given by \( \bar{\theta}_c^t \), the decoding rule given by:

\[
\tilde{\theta}_c^t = \bar{\theta}_c^t + \frac{r_c^t}{\sqrt{q_c^t}} = \bar{\theta}_c^t + \sum_{j=1}^C W(c, j) \bar{\theta}_j^t + \tilde{v}_c^t, \tag{10}
\]

where \( \tilde{v}_c^t \sim \mathcal{N}(0, \tilde{\kappa}_c^2 I_d) \) and \( \tilde{\kappa}_c^2 = \sum_{j=1}^C W(c, j) \sigma_j^2 I_d \) in the third phase, the cluster heads concurrently transmit the cluster-level updates to the clients in their respective clusters. The update process is summarized in Fig. 1 and in Algorithm 1.

### IV. Convergence of CWFL

We demonstrate the convergence of the CWFL parameters using standard assumptions of \( L \)-Lipschitz smoothness, \( \mu \)-strong convexity of \( f_k(\cdot) \), and \( G \)-boundedness and \( \alpha_k \)-bounded variance of stochastic gradients [3].
Theorem 1. Under the standard assumptions and given constants $L, \mu, \alpha_k, G$, choosing $\gamma = \max(E, \frac{1}{\mu \eta})$ and choosing the learning rate $\eta_t = \frac{1}{2(\gamma + 1)}$ for $t \in [T]$, each cluster in CWFL satisfies $O(1/T)$ convergence rate since

$$E[\|\theta^*_c - \theta^\ast\|^2] \leq \frac{2 \max(4Q_1, \mu^2 \gamma) \|\theta^0_c - \theta^\ast\|^2}{\mu^2 (T + 1 - \gamma)},$$

(11)

where $Q_1 = 3C \sum_{j=1}^{C} (W(c, j))^2 P_2 A + 8E^2 G^2 + 6 \Gamma + \frac{K_c}{2} \sum_{k=1}^{K_k} \alpha_k^2 + \frac{4K_e^2 E^2 G^2}{P_2 \mu}$, $d = \sum_{j=1}^{C} (W(c, j))^2$, and $A = \frac{8E^2 P_2 C}{P_2 \mu}$

Let $\bar{\gamma} = \max(6, \frac{2L}{\mu \eta})$, each cluster in $\mathcal{H}$, we obtain the following:

$$E[\|\theta^*_c - \theta^\ast\|^2] \leq (2 - \mu \eta^2) E[\|\theta^\ast_c - \theta^\ast\|^2]$$

+ $3(\mu \eta^2) \sum_{j=1}^{C} (W(c, j))^2 P_2 A$

+ $8(\mu \eta^2) E^2 G^2 + 6L(\mu \eta^2) \Gamma - \frac{5 \mu \eta}{3} \cdot E[\|F(\theta^\ast_c) - F^\ast\|^2]$ (15)

Taking the expectation of (14) and using Lemmas 2, 4, 5 and (15), we obtain the following:

$$E[(\theta^\ast_c - \theta^\ast)^2] \leq \frac{4E^2(\mu \eta^2) G^2}{\mu^2 (T + 1)},$$

where $\theta^\ast$ is the global optimum.

Proof. The virtual sequence from (7) for $t \in [T]$ is given as

$$\tilde{\theta}^t_c = \frac{1}{K_c} \sum_{k=1}^{K_c} \theta^t_k + \bar{w}^t_{c} \mathds{1}_{t \in \mathcal{H}},$$

(12)

where $\mathds{1}_{t}$ is the indicator function. We define

$$g^t_{c} \triangleq \frac{1}{K_c} \sum_{k=1}^{K_c} \nabla f(\theta^t_k, w_k), \quad g^t_{c} \triangleq \frac{1}{K_c} \sum_{k=1}^{K_c} \nabla f(\theta^t_k).$$

Let $\bar{w}^t_{c} \triangleq \bar{w}^t_{c} \mathds{1}_{t \in \mathcal{H}}$, so that $E[\|\bar{w}^t_{c}\|^2] = \frac{d^2}{K_c^2} E[\|w_t\|^2]$.

Since SGD is used at each client, we have

$$\tilde{\theta}^{t+1}_c = \tilde{\theta}^{t}_c - \eta^t g^t_{c} + \bar{w}^t_{c}.$$ (13)

The above equation differentiates our proof from previous works such as [3], [10]. Since previous works were server-based, $\tilde{\theta}^{t+1}_c$ would depend upon the previous update $\tilde{\theta}^{t}_c$. However, since CWFL assigns a consensus update as the global update, $\tilde{\theta}^{t}_c$ is replaced by $\tilde{\theta}^{t}_c$ in the virtual sequence. Note that $\tilde{\theta}^{t}_c = \theta^t_k \forall k \in [K_c], \theta^t_k$, when $t \in \mathcal{H}$.

Using (10) and (13) we have the following:

$$\|\tilde{\theta}^{t+1}_c - \theta^\ast\|^2 = \|\tilde{\theta}^{t}_c - \eta^t g^t_{c} - \theta^\ast + \bar{w}^t_{c}\|^2 =$$

$$\|\tilde{\theta}^{t}_c + \sum_{j=1}^{C} W(c, j) \theta^j_{c} - \eta^t g^t_{c} - \theta^\ast + \eta^t g^t_{c} - \eta^t g^t_{c} + \bar{w}^t_{c}\|^2$$

$$= \|\tilde{\theta}^{t}_c + \sum_{j=1}^{C} W(c, j) \theta^j_{c} - \eta^t g^t_{c} - \theta^\ast\|^2$$

$$+ (\eta^t)^2 \|g^t_{c} - \eta^t g^t_{c} + \bar{w}^t_{c} + \bar{w}^t_{c}\|^2$$

$$+ 2\eta^t \langle \tilde{\theta}^{t}_c + \sum_{j=1}^{C} W(c, j) \theta^j_{c} - \theta^\ast - \eta^t g^t_{c}, g^t_{c} - \bar{w}^t_{c} + \bar{w}^t_{c} \rangle.$$ (14)

where we add and subtract $\eta^t g^t_{c}$ to obtain the second step.

Following the proof in lemma A.3 of [3], where $N$ is replaced by the number of clients in the $c$-th cluster, $K_c$, when $\eta^t \leq 2(2+E)\eta^t$ for all $t \geq 0$ [10], [11] and when assumption 3 holds, we can derive the following bound:

$$\frac{1}{K_c} \sum_{k=1}^{K_c} E[\|\tilde{\theta}^{t}_k - \theta^t_k\|^2] \leq 4E^2(\eta^t)^2 G^2.$$ (15)
Lemma 3. If \( p^t \leq q^{t-E} \) and \( \eta^t \leq 2\eta^{t+E} \), we have:

\[
\frac{1}{q^t} \leq \frac{8E^2(\eta^t)^2G^2}{\eta^t} + d \max_c \sum_j W(c, j)\sigma_j^2 + \frac{d}{P_2} \sum_c \left( \sum_j W(c, j)\sigma_j^2 \right) + \frac{4E^2(\eta^t)^2G^2}{\eta^t} \mathbb{1}_{1 \in \mathcal{H}} (18)
\]

Lemma 4. (Bounding the noisy variance) : If \( \eta^t \leq 2\eta^{t+E} \), \( p^t \leq q^{t-E} \), the term \( T_2 \) is given as:

\[
\mathbb{E} \left[ \| \theta_c^t - \theta_c^t \| ^2 + \frac{\omega_c}{\eta^t} + \frac{\sigma_c^2}{\eta^t} \right] \leq \frac{1}{\eta^t} \sum_k \alpha_k^2 + \frac{4d\sigma^2G^2}{\eta^tP_1K_2} \mathbb{1}_{1 \in \mathcal{H}} + d \sum_c \sum_j W(c, j)\sigma_j^2A (19)
\]

where,

\[ A = \frac{8\rho^2G^2}{\eta^t} \left( \sum_c \sum_j (W(c, j))^2 + \frac{d}{P_2} \max_c \sum_j W(c, j)\sigma_j^2 \right) + \frac{8E^2G^2}{\eta^t} + \frac{4E^2G^2}{\eta^t} \mathbb{1}_{1 \in \mathcal{H}} \]

Lemma 5. (Bounding the divergence in non-noisy consensus update) : If \( \eta^t \leq 2\eta^{t+E} \), and using Lemma 3 we have:

\[
\mathbb{E} \left[ \sum_j W(c, j)\hat{\theta}_j^t \right]^2 \leq C(\eta^t)^2 \sum_j (W(c, j))^2P_2A (20)
\]

V. EXPERIMENTAL RESULTS

In this section, we demonstrate the performance of the proposed CWFL and CWFL-Prox algorithms, focusing on the following: (a) Accuracy across communication rounds, (b) Robustness across different number of clusters and statistical heterogeneity. We compare the performance of the proposed algorithms with COTAF [3], FedProx [9] implemented in the wireless framework, which we refer to as COTAF-Prox and single-client training where we assume that all the models train on local dataset and no FL strategy is employed. Datasets and ML models: We consider the image classification task, where the data is distributed among \( K = 25 \) clients. We use the popular MNIST and CIFAR10 datasets for benchmarking. In the case of MNIST, the ML model is a convolutional neural network (CNN) consisting of 4 layers including 2 convolutional layers, and batch size \( |\omega_k| = 64 \). In the case of CIFAR10, the ML model is a CNN consisting of 6 layers, including 3 convolutional layers and \( |\omega_k| = 32 \). Both the architectures use ReLU activation, a learning rate of \( \eta = 0.001 \). Unless mentioned otherwise, each client has instances pertaining to any 4 classes in the dataset with \( E = 3 \).

A. Accuracy Across Communication Rounds

In Fig. 2 and Fig. 3, we depict the evolution of accuracy across 50 communication rounds, in the presence of statistical heterogeneity. We observe that the algorithms that use the proximal constraint consistently performs better than its counterparts without the proximal constraint. In Fig. 2, we observe that CWFL-Prox using 4 clusters (CWFL-4-Prox) performs similar to COTAF-Prox, and outperforms COTAF. However, we see that CWFL with 3 clusters (CWFL-3-Prox) has slower convergence, but a higher average accuracy as compared to CWFL-4-Prox. This is also evident in Fig. 4 (right). On the other hand, in Fig. 3 and Fig. 4 (left) for MNIST dataset, we observe that CWFL-3-Prox has a similar performance as compared to COTAF, but COTAF-Prox has the best performance. Here, using 3 clusters is optimal for CWFL framework. In both the datasets, we see that we outperform the single client training method, where we train each client on the local dataset with no FL strategy being used. In summary, when a strong server is absent, CWFL is a good replacement of COTAF as there is little loss in accuracy.
Fig. 5: Accuracy of CWFL and CWFL-Prox for different number of output classes (MNIST and CIFAR10).

B. Benefits of CWFL

There are several scenarios where a scheme like CWFL may be preferable as compared to COTAF. For instance, consider the case when server experiences poor SNR conditions as compared to cluster-heads. In order to simulate this scenario, we have set the SNR at the clients to be 1dB lower than at the server. From Fig. 6 (left), we see that CWFL performs better in such scenarios by 7-8 dB before finally converging to the same accuracy. This occurs mainly because collectively, at all cluster-heads, the signals experience lower distortion leading to faster consensus. The benefit of using CWFL as compared to the DSGD as employed in [5] is illustrated in Fig. 6 (right), and the impact is high as the number of clusters increase. By design, $C << K$ and hence, communication complexity of CWFL is lower than DSGD.

C. Results on Attributes of Federated Learning

In this section, we study the behavior of the proposed approach with varying number of clusters and varying degree of statistical heterogeneity. Here, the data is partitioned such that each client has access to instances pertaining to any 2, 4 or 8 classes of MNIST and CIFAR10 datasets. The accuracy obtained using CWFL and CWFL-Prox in these cases are depicted in Fig. 5. As expected, these algorithms perform well when there is information pertaining to 8 classes. In challenging datasets such as CIFAR10, the accuracy drops sharply for the pathological 2 class scenario. For both the datasets, CWFL performs well in the presence of 4 classes. The performance of CWFL-Prox with 4 classes is almost as good as CWFL and CWFL-Prox with 8 classes.

The accuracy performance of CWFL and CWFL-Prox as compared to COTAF for different number of clusters was investigated, and we observe that optimal number of clusters in this setting is 3 or 4. While CWFL-3-Prox outperforms CWFL-4-Prox and COTAF for the CIFAR10 dataset, in the MNIST dataset, the performance of CWFL-3-Prox is similar to COTAF with respect to average accuracy. In both the cases, it is optimal to choose 3 clusters instead of 4.

VI. CONCLUSIONS

In scenarios where a powerful server is absent, we proposed a semi-decentralised CWFL and CWFL-Prox frameworks. This framework is based on data-agnostic random clustering of clients, where FL is accomplished using OTA power controlled transmission over wireless uplink channels. We showed that CWFL is a convergent FL scheme ($O(1/T)$) that requires fewer channel uses as compared to decentralised FL. Using the MNIST and CIFAR10 datasets, we demonstrated that the accuracy performance of CWFL is as good as COTAF, while CWFL-Prox outperforms COTAF. The proximal constraint forces the per-client parameter update to lie in the vicinity of the previous global update, and this mitigates the combined impact of the noise and statistical heterogeneity.

REFERENCES

[1] M. M. Amiri, Deniz G., Sanjeev R. K., and H Vincent P., “Convergence of federated learning over a noisy downlink,” IEEE Trans. Wireless Comm., 2021.
[2] Amirhossein R., Aryan M., Hamed H., Ali J., and Ramtin P., “Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization,” in International Conference on Artificial Intelligence and Statistics. PMLR, 2020.
[3] Tomer S., Nir S., Kobi C., and Yonina C. E., “Cotaf: Convergent over-the-air federated learning,” in GLOBECOM 2020, 2020, pp. 1–6.
[4] Kai Y., Tao J., Yuanming S., and Zhi D., “Federated learning via over-the-air computation,” IEEE Transactions on Wireless Communications, vol. 19, no. 3, pp. 2022–2035, 2020.
[5] Hong X., Osvaldo S., and Suzhi B., “Federated learning over wireless device-to-device networks: Algorithms and convergence analysis,” IEEE Journal on Selected Areas in Communications, vol. 39, no. 12, pp. 3723–3741, 2021.
[6] Zhenyu Wang, Ting Zhou, Tianheng Xu, and Honglin Hu, “An iterative greedy user clustering algorithm for d2d-relay in vehicular communication systems,” IET Microwaves, Antennas and Propagation, vol. 13, 07 2019.
[7] Yanwei Li, Zhenyu Zhou, and Takuro Sato, “A cluster-based consensus algorithm in a wireless sensor network,” International Journal of Distributed Sensor Networks, vol. 9, no. 3, pp. 547124, 2013.
[8] Brendan M., Eider M., Daniel R., Seth H., and Blaise A. y A., “Communication-efficient learning of deep networks from decentralized data,” in Artificial Intelligence and statistics. PMLR, 2017.
[9] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith, “Federated optimization in heterogeneous networks,” Proc. of MLSys, vol. 2, 2020.
[10] Xiang L., Kaixuan H., Wenhao Y ., Shusen W., and Zhihua Z., “On the communication of fedavg on non-iid data,” in ICLR, 2019.
[11] Sebastian Urban Stich, “Local sgd converges fast and communicates little,” in ICLR 2019-International Conference on Learning Representations, 2019, number CONF.