Over-the-Air Federated Learning From Heterogeneous Data

Tomer Sery, Nir Shlezinger, *Member, IEEE*, Kobi Cohen, *Senior Member, IEEE*, and Yonina C. Eldar, *Fellow, IEEE*

***Abstract*—We focus on over-the-air (OTA) Federated Learning (FL), which has been suggested recently to reduce the communication overhead of FL due to the repeated transmissions of the model updates by a large number of users over the wireless channel. In OTA FL, all users simultaneously transmit their updates as analog signals over a multiple access channel, and the server receives a superposition of the analog transmitted signals. However, this approach results in the channel noise directly affecting the optimization procedure, which may degrade the accuracy of the trained model. We develop a Convergent OTA FL (COTAF) algorithm which enhances the common local stochastic gradient descent (SGD) FL algorithm, introducing precoding at the users and scaling at the server, which gradually mitigates the effect of noise. We analyze the convergence of COTAF to the loss minimizing model and quantify the effect of a statistically heterogeneous setup, i.e. when the training data of each user obeys a different distribution. Our analysis reveals the ability of COTAF to achieve a convergence rate similar to that achievable over error-free channels. Our simulations demonstrate the improved convergence of COTAF over vanilla OTA local SGD for training using non-synthetic datasets. Furthermore, we numerically show that the precoding induced by COTAF notably improves the convergence rate and the accuracy of models trained via OTA FL.**

***Index Terms*—Machine learning, optimization, gradient methods, wireless communication.**

I. Introduction

RECENT years have witnessed unprecedented success of machine learning methods in a broad range of applications [2]. These systems utilize highly parameterized models, such as deep neural networks (DNNs), trained using massive data sets. In many applications, samples are available at remote users, e.g. smartphones, and the common strategy is to gather these samples at a computationally powerful server, where the model is trained [3]. Often, data sets contain private information, and thus the user may not be willing to share them with the server. Furthermore, sharing massive data sets can result in a substantial burden on the communication links between the users and the server. To allow centralized training without data sharing, federated learning (FL) was proposed in [4] as a method combining distributed training with central aggregation, and is the focus of growing research attention [5], [6]. FL exploits the increased computational capabilities of modern edge devices to train a model on the users' side, having the server periodically synchronize these local models into a global one.

Two of the main challenges associated with FL are the heterogeneous nature of the data and the communication overhead induced by its training procedure [5]. Statistical heterogeneity arises when the data generating distributions vary between different sets of users [7]. This is typically the case in FL, as the data available at each user device is likely to be personalized towards the specific user. As an example, consider the task of sentence completion from text messages. Since users of different ages and backgrounds are likely to use different wording and sentence structures, the data sets from different users will be imbalanced. When training several instances of a model on multiple edge devices using heterogeneous data, each instance can be adapted to operate under a different statistical relationship, which may limit the inference accuracy of the global model [7]–[9].

The communication load of FL stems from the need to repeatedly convey a massive amount of model parameters between the server and a large number of users over wireless channels [9]. This is particularly relevant in uplink communications, which are typically more limited as compared to their downlink counterparts [10]. A common strategy to tackle this challenge is to reduce the amount of data exchanges between the users and the server, either by reducing the number of participating users [11], [12], or by compressing the model parameters via quantization [13], [14] or sparsification [15], [16]. All these methods treat the wireless channel as a set of independent error-free bit-limited links between the users and the server. As wireless channels are shared and noisy [17], a common way to achieve such communications is to divide the channel resources among users, e.g., by using frequency division multiplexing (FDM), and have the users utilize channel codes to overcome the noise. Such protocols are often utilized in wireless communication

Manuscript received September 25, 2020; revised February 28, 2021 and May 23, 2021; accepted June 9, 2021. Date of publication June 17, 2021; date of current version July 23, 2021. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Byonghyo Shim. This work was supported in part by the Benoziyo Endowment Fund for the Advancement of Science, the Estate of Olga Klein – Astrachan, in part by the European Union's Horizon 2020 research, and innovation program under Grant 646804-ERC-COG-BNYQ, in part by the Israel Science Foundation under Grant 0100101, in part by the Israel Science Foundation under Grant 2640/20, and in part by the U.S.-Israel Binational Science Foundation (BSF) under Grant 2017723. A short version of this paper that introduces the algorithm for i.i.d. data and preliminary simulation results was accepted for presentation in the 2020 IEEE Global Communications Conference (GLOBECOM) [1]. *(Corresponding author: Tomer Sery.)*

Tomer Sery, Nir Shlezinger, and Kobi Cohen are with the School of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer-Sheva 4486200, Israel (e-mail: seryt@post.bgu.ac.il; nirshlezinger1@gmail.com; kobi.cohen10@gmail.com).

Yonina C. Eldar is with the Math and CS Faculty, Weizmann Institute of Science, Rehovot 761001, Israel (e-mail: yonina.eldar@weizmann.ac.il).

Digital Object Identifier 10.1109/TSP.2021.3090323

1053-587X © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on July 29,2021 at 09:55:20 UTC from IEEE Xplore. Restrictions apply.

standards, as they mitigate the interference between the users and thus facilitate recovery of the individual messages sent by each user. This, however, results in each user being assigned a dedicated band whose width decreases with the number of users, which in turn increases the energy consumption required to meet a desirable communication rate and decreases the overall throughput and training speed.

An alternative FL approach is to allow the users to simultaneously utilize the complete temporal and spectral resources of the uplink channel in a non-orthogonal manner. This method, referred to as over-the-air (OTA) FL [18]–[22], exploits the fact that the server in FL requires the individual model updates as an intermediate step in aggregating them into a global model. Therefore, while the presence of interference makes recovery of each individual message more challenging compared to orthogonal transmissions, OTA FL directly recovers the federated averaged model, exploiting the inherent aggregation carried out by the shared channel as a form of OTA computation [23]. OTA FL builds upon the fact that when the participating users operate over the same wireless network, uplink transmissions are carried out over a mulitple access channel (MAC). Model-dependent inference over MACs is relatively well-studied in the sensor network literature, where methods for model-dependent inference over MACs and theoretical performance guarantees have been established under a wide class of problem settings (see [24], [25] and references therein). These studies focused on statistical-model-based parameter estimation or detection, in which the task is to recover a parameter of interest based on knowledge of an underlying statistical model by utilizing observations from different users in a wireless network. Such model-based inference is fundamentally different from machine learning paradigms, such as FL.

In the context of FL with OTA computations, the works [18], [19] considered scenarios where the model updates are sparse with an identical sparsity pattern, using this pattern to reduce communication overhead. However, these assumptions are not likely to hold when the data is heterogeneous. Additional related recent works on OTA FL, including [20], [22], [26], considered the distributed application of full gradient descent optimization over noisy channels. Energy management in OTA FL with gradient transmissions was studied in [27], [28], and OTA FL in MACs aided by reconfigurable intelligent surfaces was considered in [29]. Distributed learning based on full gradient descent admits a simplified and analytically tractable analysis [20], [22]. However, it requires each user to compute and repeatedly transmit the gradients over the complete data set. Consequently, it is less communication and computation efficient compared to methods which involve gradient computation over subsets of the data and multiple local iterations prior to transmission, such as local stochastic gradient descent (SGD) which is the dominant optimization scheme used in FL [4], [5]. Consequently, OTA FL schemes proposed in these previous works and the corresponding convergence analysis may not reflect the common application of FL, i.e., distributed training with heterogeneous data via local SGD.

The main advantage of OTA FL is that it enables users to transmit at increased throughput, being allowed to utilize the complete available bandwidth regardless of the number of participating users. However, a major drawback of existing OTA FL methods, e.g., [18]–[22], follows from the fact that using uncoded analog signalling results in the noise induced by the channels not being handled by channel coding. Consequently, the presence of noise affects the training procedure. In particular, the accuracy of learning algorithms such as SGD is known to be sensitive to noisy observations. For instance, while gradient-based optimization converges to the optimal solution for convex loss measures, in the presence of noise with sub-Gaussian tails, the model can only be shown to converge to some environment of the optimal solution [30]. This sensitivity of gradient-based optimization to noisy observations adds to the limited accuracy due to statistical heterogeneity of FL [31]. This implies that conventional FL algorithms, such as local SGD [32], exhibit degraded performance when combined with noise-inducing OTA computations. As a result, OTA FL is typically unable to converge to the optimal weights for convex loss measures. In this paper we overcome this drawback by introducing time-varying precoding that gradually mitigates the contribution of the channel noise over time.

A. Main Contributions

Specifically, we study OTA FL while accounting for the effect of channel noise as well as the heterogeneity of the data. Our main contributions are summarized as follows:

1) Algorithm development: We develop a joint computation and transmission scheme, named convergent OTA FL (COTAF). We introduce time-varying precoding to the transmitted signals, which accounts for the fact that the expected difference in each set of SGD iterations is gradually decreasing over time, while guaranteeing energy-bounded transmissions. COTAF results in an equivalent model where the effect of the noise induced by the channel is mitigated over time, thus facilitating high throughput FL over wireless channels, while preserving the accuracy and convergence properties of local SGD for distributed learning. By conveying the model updates over the uplink MAC, COTAF overcomes the need to divide the channel resources among users. In contrast to previous OTA FL works, e.g., [18]–[20], [22], [26], COTAF does not assume that the users compute a single full gradient step or that the models share a similar sparsity pattern, making it applicable to FL based on local SGD optimization with heterogeneous data.
2) Performance analysis: We analytically show that machine learning models trained by COTAF converge to the minimal achievable loss function in the presence of heterogeneous data. We provide three convergence bounds: The first two bounds consider FL over non-fading channels, characterizing the convergence of a weighted average of past models as in [32] and the convergence of the instantaneous model [12]. We then extend COTAF to fading channels and characterize the corresponding convergence of the instantaneous model.

Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on July 29,2021 at 09:55:20 UTC from IEEE Xplore. Restrictions apply.

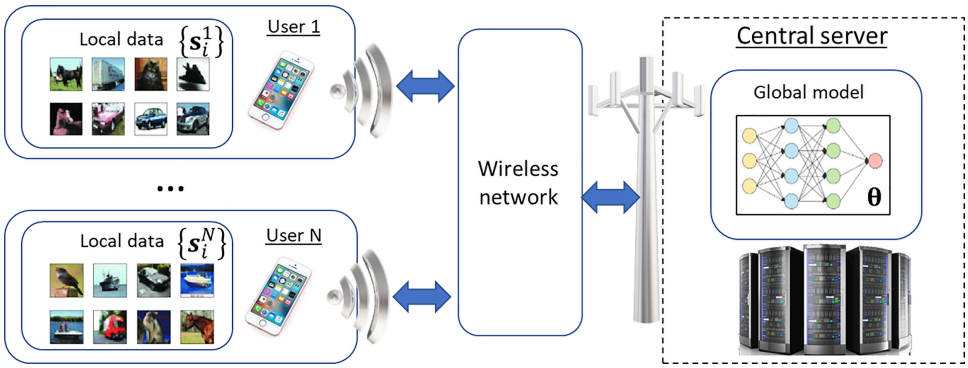

Fig. 1. An illustration of the distributed optimization setup. In this example, the data consists of images, where those of user 1 are biased towards car images, while those of user N contain a large portion of ship images, resulting in a heterogeneous setup.

3) Experimental study: We evaluate COTAF in two scenarios involving non-synthetic data sets: First, we train a linear estimator, for which the objective function is strongly convex, with the Million Song Dataset [33]. We demonstrate that COTAF approaches the accuracy of noise-free local SGD, while notably outperforming previous OTA FL strategies. Then, we train a convolutional neural network (CNN) over the CIFAR-10 dataset, representing a deep FL setup with a non-convex objective, for which a minor level of noise is known to contribute to convergence as means of avoiding local minima [34]. We demonstrate that COTAF improves the accuracy of trained models when using both i.i.d and heterogeneous data, outperforming not only conventional OTA FL, but also noise-free local SGD.

The rest of this paper is organized as follows: Section II briefly reviews the local SGD algorithm and presents the system model of OTA FL. Section III presents the COTAF scheme along with its theoretical convergence analysis. Numerical results are detailed in Section IV. Finally, Section V provides concluding remarks. Detailed proofs of our main results are given in the appendix.

Throughout the paper, we use boldface lower-case letters for vectors, e.g., $\boldsymbol{x}$. The ℓ_2 norm, stochastic expectation, and Gaussian distribution are denoted by $\|\cdot\|$, $\mathbb{E}[\cdot]$, and $\mathcal{N}(\cdot,\cdot)$ respectively. Finally, $\boldsymbol{I}_n$ is the $n\times n$ identity matrix, and $\mathbb{R}$ is the set of real numbers.

II. System Model

In this section we detail the system model for which COTAF is derived in the following section. We first formulate the objective of FL in Subsection II-A. Then, Subsection II-B presents the communication channel model over which FL is carried out. We briefly discuss the local SGD method, which is the common FL algorithm, in Subsection II-C, and formulate the problem in Subsection II-D.

A. Federated Learning

We consider a central server which trains a model consisting of d parameters, represented by the vector $\theta\in\Theta\subset\mathbb{R}^d$, using data available at N users, indexed by the set $\mathcal{N}=\{1,2,\ldots,N\}$, as illustrated in Fig. 1. Each user of index $n\in\mathcal{N}$ has access to a data set of D_n entities, denoted by $\{\boldsymbol{s}_i^n\}_{i=1}^{D_n}$, sampled in an i.i.d. fashion from a local distribution $\mathcal{X}_n$. The users can communicate with the central server over a wireless channel formulated in Subsection II-B, but are not allowed to share their data with the server.

To define the learning objective, we use $l(\cdot,\theta)$ to denote the loss function of a model parameterized by θ. The empirical loss of the nth user is defined by

$$f_n(\theta)\triangleq\frac{1}{D_n}\sum_{i=1}^{D_n}l(\boldsymbol{s}_i^n;\theta). \tag{1}$$

FL aims at minimizing the average loss:

$$\theta^\star\triangleq\arg\min_{\theta\in\Theta}\ F(\theta),\quad F(\theta)\triangleq\frac{1}{N}\sum_{n=1}^{N}f_n(\theta). \tag{2}$$

When the data is homogeneous, i.e., the local distributions $\{\mathcal{X}_n\}$ are identical, the local loss functions converge to the same expected loss measure on the horizon of a large number of samples $D_n\to\infty$. However, the statistical heterogeneity of FL, i.e., the fact that each user observes data from a different distribution, implies that the parameter vectors which minimize the local loss vary between different users. This property generally affects the behavior of the learning method used in FL, such as the common local SGD algorithm, detailed in Subsection II-C.

B. Communication Channel Model

FL is often carried out over wireless channels. We consider FL setups in which the N users communicate with the server using the same wireless network, either directly or via some wireless access point. As uplink communications, i.e., from the users to the server, is typically notably more constrained as compared to its downlink counterpart in terms of throughput [10], we focus on uplink transmissions over MAC. The downlink channel is modeled as supporting reliable communications at arbitrary

Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on July 29,2021 at 09:55:20 UTC from IEEE Xplore. Restrictions apply.

rates, as commonly assumed in FL studies [13]–[16], [18]–[20], [35].

We next formulate the uplink channel model. Wireless channels are inherently a shared and noisy media, hence the channel output received by the server at time instance t when each user transmits a $d \times 1$ vector $\boldsymbol{x}_t^n$ is given by

$$\boldsymbol{y}_t = \sum_{n=1}^{N} \boldsymbol{x}_t^n + \tilde{\boldsymbol{w}}_t, \tag{3}$$

where $\tilde{\boldsymbol{w}}_t \sim \mathcal{N}(0, \sigma_w^2 \boldsymbol{I}_d)$ is $d \times 1$ vector of additive noise. While we model the noise as Gaussian, being the common noise model in wireless communications, our analysis of FL in the sequel is invariant to the marginal distribution of the noise signal. The channel input is subject to an average power constraint

$$\mathbb{E}\left[\|\boldsymbol{x}_t^n\|^2\right] \leq P, \tag{4}$$

where $P > 0$ represents the available transmission power. The channel in (3) represents an additive noise MAC, whose main resources are its spectral band, denoted B, and its temporal blocklength τ, namely, $\boldsymbol{y}_t$ is obtained by observing the channel output over the bandwidth B for a duration of τ time instances.

The common approach in wireless communication protocols and in FL research is to overcome the mutual interference induced in the shared wireless channels by dividing the bandwidth into multiple orthogonal channels. This can be achieved by, e.g., FDM, where the bandwidth is divided into N distinct bands, or via time division multiplexing (TDM), in which the temporal block is divided into N slots which are allocated among the users. In such cases, the server has access to a distinct channel output for each user, of the form

$$\boldsymbol{y}_t^n = \boldsymbol{x}_t^n + \tilde{\boldsymbol{w}}_t^n, \quad n \in \mathcal{N}. \tag{5}$$

Orthogonalization of the channels in (5) facilitates recovery of each $\boldsymbol{x}_t^n$ individually. However, the fact that each user has access only to $1/N$ of the channel resources implies that its throughput, i.e., the volume of data that can be conveyed reliably, is reduced accordingly [17, Ch. 4]. In order to facilitate high throughput FL, we do not restrict the users to orthogonal communication channels, and thus the server has access to the shared channel output (3) rather than the set of individual channel outputs in (5).

We derive our OTA FL scheme and analyze its performance assuming that the users communicate with the server of the noisy MAC (3). However, in practice wireless channels often induce fading in addition to noise. Each user of index n experiences at time t a block fading channel $\tilde{h}_t^n = h_t^n e^{j\phi_t^n}$, where $h_t^n > 0$ and $\phi_t^n \in [-\pi, \pi]$ are its magnitude and phase, respectively. In such cases, the channel input-output relationship is given by

$$\boldsymbol{y}_t = \sum_{n=1}^{N} \tilde{h}_t^n \boldsymbol{x}_t^n + \tilde{\boldsymbol{w}}_t. \tag{6}$$

Therefore, we show how the proposed COTAF algorithm can be extended to fading MACs of the form (6). In our extension, we assume that the participating entities have channel state information (CSI), i.e., knowledge of the fading coefficients. Such knowledge can be obtained by letting the users sense their channels, or alternatively by having the access point/server periodically estimate these coefficients and convey them to the users.

C. Local SGD

Local SGD, also referred to as *federated averaging* [4], is a distributed learning algorithm aimed at recovering (2), without having the users share their local data. This is achieved by carrying out multiple training rounds, each consisting of the following three phases:

1) The server shares its current model at time instance t, denoted by θ_t, with the users.
2) Each user sets its local model $\boldsymbol{\theta}_t^n$ to θ_t, and trains it using its local data set over H SGD steps, namely,

$$\boldsymbol{\theta}_{t+1}^n = \boldsymbol{\theta}_t^n - \eta_t \nabla f_{i_t^n}(\boldsymbol{\theta}_t^n), \tag{7}$$

where $f_{i_t^n}(\theta) \triangleq l(s_{i_t^n}^n; \theta)$ is the loss evaluated at a single data sample, drawn uniformly from $\{\boldsymbol{s}_i^n\}_{i=1}^{D_n}$, and η_t is the SGD step size. The update rule (7) is repeated H steps to yield $\boldsymbol{\theta}_{t+H}^n$.
3) Each user conveys its trained local model $\boldsymbol{\theta}_{t+H}^n$ (or alternatively, the updates in its trained model $\boldsymbol{\theta}_{t+H}^n - \boldsymbol{\theta}_t^n$) to the central server, which averages them into a global model via[1] $\theta_{t+H} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{\theta}_{t+H}^n$, and sends the new model to the users for another round.

The uplink transmission in this algorithm is typically executed over an error-free channel with limited throughput, where channel noise and fading are assumed to be eliminated [12], [32], [36]. The local SGD algorithm is known to result in a model θ_t whose objective function $F(\theta_t)$ converges to $F^\star \triangleq F(\boldsymbol{\theta}^\star)$ as the number of rounds grows for various families of loss measures under homogeneous data [32]. When the data is heterogeneous, convergence is affected by an additional term encapsulating the *degree of heterogeneity*, defined as $\Gamma \triangleq F^\star - \frac{1}{N}\sum_{n=1}^{N} f_n^\star$, where $f_n^\star \triangleq \min_\theta f_n(\theta)$ [12]. In particular, for convex objectives, convergence of the global model to (2) can be still guaranteed, though at slower rates compared to homogeneous setups [12]. To the best of our knowledge, the convergence of local SGD with heterogeneous data (e.g. non-i.i.d data distribution between users) carried out over noisy fading wireless channels has not been studied to date.

D. Problem Formulation

We consider FL carried out over shared wireless channels using local SGD optimization, as detailed in Subsection II-C. Each round of local SGD consists of two communication phases: downlink transmission of the global model θ_t from the server to the users, and uplink transmissions of the updated local models $\{\boldsymbol{\theta}_{t+H}^n\}$ from each user to the server. An illustration of a single round of local SGD carried out over a wireless MAC of the form (3) is depicted in Fig. 2. This involves the repetitive communication of a large amount of parameters over

[1]While we focus here on conventional averaging of the local models, our framework can be naturally extended to weighted averages.

Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on July 29,2021 at 09:55:20 UTC from IEEE Xplore. Restrictions apply.

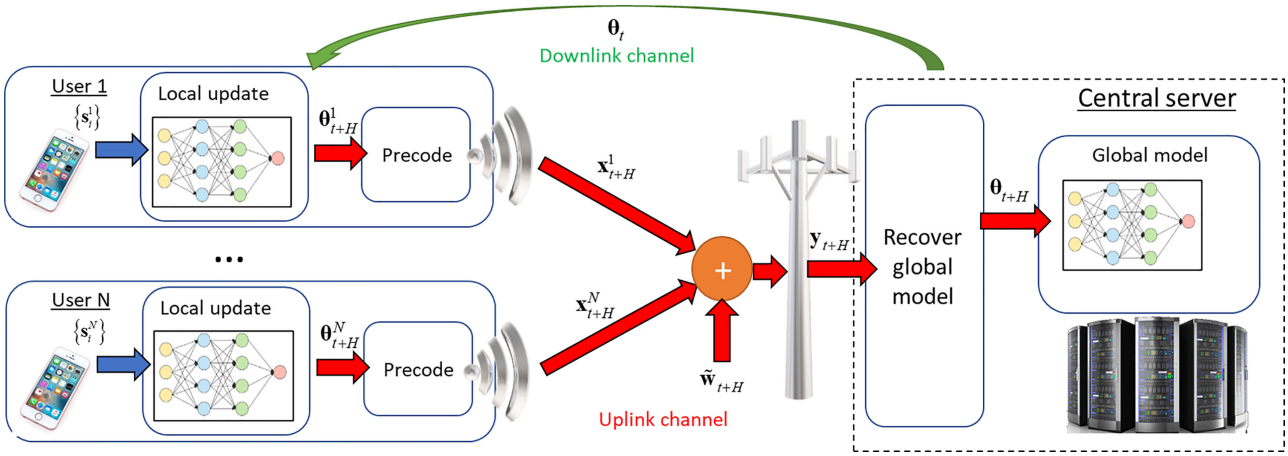

Fig. 2. An illustration of FL over wireless MAC.

wireless channels. This increased communication overhead is considered one of the main challenges of FL [5], [9]. The conventional strategy in the FL literature is to treat the uplink channel as an error-free bit-constrained pipeline, and thus the focus is on deriving methods for compressing and sparsifying the conveyed model updates, such that convergence of θ_t to $\boldsymbol{\theta}^\star$ is preserved [13], [14], [16]. However, the model of error-free channels, which are only constrained in terms of throughput, requires the bandwidth of the wireless channel to be divided between the users and have each user utilize coding schemes with a rate small enough to guarantee accurate recovery. This severely limits the volume of data which can be conveyed as compared to utilizing the full bandwidth.

The task of the server on every communication round in FL is not to recover each model update individually, but to aggregate them into a global model θ_t. This motivates having each of the users exploit the complete spectral and temporal resources by avoiding conventional orthogonality-based strategies and utilizing the wireless MAC (3) on uplink transmissions. The inherent aggregation carried out by the MAC can in fact facilitate FL at high communication rate via OTA computations [23], as was also proposed in the context of distributed learning in [18]–[20]. However, the fact that the channel outputs are corrupted by additive noise is known to degrade the ability of SGD-based algorithms to converge to the desired $\theta^\star$ for convex objectives [30], adding to the inherent degradation due to statistical heterogeneity. For non-convex objectives, noise can contribute to the overall convergence as it reduces the probability of getting trapped in local minima [34], [37]. However, for the learning algorithm to benefit from such additive noise, the level of noise should be limited. It is preferable to have a gradual decay of the noise over time to allow convergence when in the proximity of the desired optimum point, which is not the case when communicating over noisy MACs.

Our objective is to design a communication strategy for FL over wireless channels based on the optimization problem (2). This involves determining a mapping, referred to as precoding, from θ_t^n into $\boldsymbol{x}_t^n$ at each user, as well as a transformation of $\boldsymbol{y}_t$ into θ_t on the server side. To tackle the challenges described above, the algorithm is required to: 1) Mitigate the limited convergence of noisy SGD for convex objectives by properly precoding the model updates into the channel inputs $\{\boldsymbol{x}_t^n\}$; 2) benefit from the presence of noise when trained using non-convex objectives; and 3) allow achieving FL performance which approaches that of FL over noise-free orthogonal channels for convex objectives, while utilizing the complete spectral and temporal resources of the wireless channel. This is achieved by introducing time-varying precoding mapping $\theta_t^n \mapsto \boldsymbol{x}_t^n$ at the users' side, and scaling laws are introduced at the server side for accurate transformation of the received signal to a global model. These rules gradually mitigate the effect of noise on the resulting global model, as detailed in the following section.

III. The Convergent Over-the-Air Federated Learning (COTAF) Algorithm

We now propose the COTAF algorithm. We first describe the COTAF transmission and aggregation protocol in Subsection III-A. Then, we analyze its convergence in Subsection III-B, proving its ability to converge to the loss-minimizing network weights under strongly convex objectives. In Subsection III-C we extend COTAF to fading channels, and discuss its pros and cons in Subsection III-D.

A. Precoding and Reception Algorithm

In COTAF, all users transmit their corresponding signals $\{\boldsymbol{x}_t^n\}$ over a shared channel to the server. The transmitted signals are aggregated over the wireless MAC and are received at the server as a sum, together with additive noise. As in [18], [20], [24], we utilize analog signalling, namely, each vector $\boldsymbol{x}_t^n$ consists of continuous-amplitude quantities, rather than a set of discrete symbols or bits, as common in digital communications. On each communication round, the server recovers the global model directly from the channel output $\boldsymbol{y}_t$, and feedbacks the updated model to the users as in conventional local SGD.

COTAF implements local SGD while communicating over an uplink wireless MAC. As a result, its performance is not measured in its ability to recover the model parameters at the server side, as in conventional OTA computation, but rather in terms

Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on July 29,2021 at 09:55:20 UTC from IEEE Xplore. Restrictions apply.

of the accuracy of the learning algorithm, for a given number of iteration, i.e., the expected objective achieved after T iterations. Let $\mathcal{H}$ be the set of time instances in which transmissions occur, i.e., the integer multiples of H. In order to convey the local trained model after H local SGD steps, i.e., at time instance $t \in \mathcal{H}$, the nth user precodes its model update $\boldsymbol{\theta}_t^n - \boldsymbol{\theta}_{t-H}^n$ into the MAC channel input $\boldsymbol{x}_t^n$ via

$$\boldsymbol{x}_t^n = \sqrt{\alpha_t}\left(\boldsymbol{\theta}_t^n - \boldsymbol{\theta}_{t-H}^n\right), \tag{8}$$

where α_t is a precoding factor set to gradually amplify the model updates as t progresses, while satisfying the power constraint (4). The precoder α_t is given by

$$\alpha_t \triangleq \frac{P}{\max_n \mathbb{E}\left[||\theta_t^n - \theta_{t-H}^n||^2\right]}. \tag{9}$$

The precoding parameter α_t depends on the distribution of the updated model, which depends on the distribution of the data. It can thus be computed by performing offline simulations with smaller data sets and distributing the numerically computed coefficients among the users, as we do in our numerical study in Section IV. Alternatively, when the loss function has bounded gradients, this term can be replaced with a coefficient that is determined by the bound on the norm of the gradients, as we discuss in Subsection III-D.

The channel output (3) is thus given by

$$\boldsymbol{y}_t = \sum_{n=1}^{N} \sqrt{\alpha_t}\left(\boldsymbol{\theta}_t^n - \boldsymbol{\theta}_{t-H}^n\right) + \tilde{\boldsymbol{w}}_t. \tag{10}$$

In order to recover the aggregated global model θ_t from $\boldsymbol{y}_t$, the server sets

$$\theta_t = \frac{\boldsymbol{y}_t}{N\sqrt{\alpha_t}} + \theta_{t-H}, \tag{11}$$

for $t \in \mathcal{H}$, where $\boldsymbol{\theta}_0$ is the initial parameter estimate. The global update rule (11) can be equivalently written as

$$\boldsymbol{\theta}_t = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{\theta}_t^n + \boldsymbol{w}_t, \tag{12}$$

where $\boldsymbol{w}_t \triangleq \frac{\tilde{\boldsymbol{w}}_t}{N\sqrt{\alpha_t}}$ is the equivalent additive noise term distributed via $\boldsymbol{w}_t \sim \mathcal{N}(0, \frac{\sigma_w^2}{N^2\alpha_t}\boldsymbol{I}_d)$. The resulting OTA FL algorithm with R communication rounds is summarized below in Algorithm 1. Here, the local model available at the nth user at time t can be written as:

$$\theta_{t+1}^n = \begin{cases} \theta_t^n - \eta_t \nabla f_{i_t^n}(\theta_t^n), & t{+}1 \notin \mathcal{H}, \\ \frac{1}{N}\sum\limits_{n=1}^{N}\left(\theta_t^n - \eta_t \nabla f_{i_t^n}(\theta_t^n)\right) + \boldsymbol{w}_t, & t{+}1 \in \mathcal{H}. \end{cases} \tag{13}$$

Algorithm 1: COTAF Algorithm.

Init: Fix an initial $\boldsymbol{\theta}_0^n = \boldsymbol{\theta}_0$ for each user $n \in \mathcal{N}$.
for $t = 1, 2, \ldots, RH$ **do**
  Each user $n \in \mathcal{N}$ locally trains $\boldsymbol{\theta}_t^n$ via (7);
  if $t \in \mathcal{H}$ **then**
    Each user $n \in \mathcal{N}$ transmits $\boldsymbol{x}_t^n$ precoded via (8) over the MAC (3);
    The server recovers $\boldsymbol{\theta}_t$ from $\boldsymbol{y}_t$ via (11);
    The server broadcasts $\boldsymbol{\theta}_t$ to the users;
    Each user $n \in \mathcal{N}$ sets $\boldsymbol{\theta}_t^n = \boldsymbol{\theta}_t$;
  end
end
Output: Global model $\boldsymbol{\theta}_{RH}$

B. Performance Analysis

In this section, we analyze the performance of COTAF. Our analysis is carried out under the following assumptions:

- AS1 The objective function $F(\cdot)$ is L-smooth, namely, for all $\boldsymbol{v}_1, \boldsymbol{v}_2$ it holds that $F(\boldsymbol{v}_1) - F(\boldsymbol{v}_2) \leq (\boldsymbol{v}_1 - \boldsymbol{v}_2)^T \nabla F(\boldsymbol{v}_2) + \frac{1}{2}L\|\boldsymbol{v}_1 - \boldsymbol{v}_2\|^2$.
- AS2 The objective function $F(\cdot)$ is μ-strongly convex, namely, for all $\boldsymbol{v}_1, \boldsymbol{v}_2$ it holds that $F(\boldsymbol{v}_1) - F(\boldsymbol{v}_2) \geq (\boldsymbol{v}_1 - \boldsymbol{v}_2)^T \nabla F(\boldsymbol{v}_2) + \frac{1}{2}\mu\|\boldsymbol{v}_1 - \boldsymbol{v}_2\|^2$.
- AS3 The stochastic gradients $\nabla f_{i_t^n}(\boldsymbol{\theta})$ satisfy $\mathbb{E}[\|\nabla f_{i_t^n}(\boldsymbol{\theta})\|^2] \leq G^2$ and $\mathbb{E}[\|\nabla f_{i_t^n}(\boldsymbol{\theta}) - \nabla f_n(\boldsymbol{\theta})\|^2] \leq M_n^2$ for some fixed $G^2 > 0$ and $M_n^2 > 0$, for each $\boldsymbol{\theta} \in \Theta$ and $n \in \mathcal{N}$.

Assumptions *AS1–AS3* imply that the model is trained in a federated manner to optimize a smooth strongly-convex objective with bounded gradients. In particular, *AS1–AS2* hold for objective functions such as those encountered in ℓ_2-norm regularized linear regression and logistic regression [12]. These assumptions are commonly used when studying the convergence of FL schemes, see, e.g., [12], [22], [32]. As a result, analyzing COTAF under *AS1–AS3* facilitates the comparison of its convergence profile to local SGD carried out over noise-free interference-free links, as in [12]. These assumptions are required to maintain an analytically tractable convergence analysis. However, COTAF can be applied for arbitrary learning tasks for which *AS1–AS3* do not necessarily hold, as numerically demonstrated in Section IV.

After $T = RH$ iterations of updating the global model via COTAF, the server utilizes its learned global model for inference. This can be achieved by setting the global model weights according to the instantaneous parameters vector available at this time instance, i.e., θ_T. An alternative approach is to utilize the fact that the server also has access to previous aggregated models, i.e., $\{\theta_r\}$ for each $r \in \mathcal{H}$ such that $r \leq T$. In this case, the server can infer using a model whose parameters are obtained as a weighted average of its previous learned model parameters, denoted by $\hat{\theta}_T$, which can be optimized to reduce the model variance [38] and thus improve the convergence rate.

The error (or the excess risk) of gradient descent type algorithms is commonly defined as the loss in the objective value at iteration t with respect to the optimal value:

$$\mathbb{E}[F(\theta_t)] - F(\theta^\star). \tag{14}$$

We next establish a finite-sample bound on the error, given by the expected loss (14) at iteration T, for both the weighted average model $\hat{\theta}_T$ and instantaneous weights θ_T. We begin with the bound relevant for the average model, stated in the following theorem:

Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on July 29,2021 at 09:55:20 UTC from IEEE Xplore. Restrictions apply.

Theorem 1: Let $\{\boldsymbol{\theta}_t^n\}_{n=1}^N$ be the model parameters generated by COTAF according to (7), and (12) over R rounds, i.e., $t \in \{0, 1, \dots T-1\}$ with $T = RH$. Then, when *AS1–AS3* hold and the step sizes are set to $\eta_t = \frac{4}{\mu(a+t)}$ with shift parameter $a > \max\{16\frac{L}{\mu}, H\}$, and the precoder is set as in (9), it holds that

$$\mathbb{E}[F(\hat{\theta}_T)] - F^\star \leq \frac{4(T+R)}{3\mu S_R}(2a+H+R-1)B + \frac{16\,dTHG^2\sigma_w^2}{3\mu P N^2 S_R}(2a+T+H) + \frac{\mu a^3}{6S_R}||\theta_0-\theta^\star||^2, \quad (15)$$

where $\hat{\theta}_T = \frac{1}{S_R}\sum_{r=1}^R \beta_r \theta_{rH}$, for $\beta_t = (a+t)^2$, $S_R = \sum_{r=1}^R \beta_{rH} \geq \frac{1}{3H}T^3$, and $B = 8H^2G^2 + \frac{1}{N^2}\sum_{n=1}^N M_n^2 + 6L\Gamma$.

Proof: The proof is given in Appendix A. □

The weighted average in $\hat{\theta}_T$ is taken over the models known to the server, i.e., $\{\theta_r\}$ with $r \in \mathcal{H}$. For comparison, in previous convergence studies of local SGD and its variants [32], [36], the weighted average is computed over every past model, including those available only to users and not to the server. In such cases, the resulting bound does not necessarily correspond to an actual model used for inference, since the weighted average is not attainable. Comparing Theorem 1 to the corresponding result in [32], which considered i.i.d data and noise-free channels, we observe that COTAF achieves the same convergence rate, with an additional term which depends on the noise-to-signal ratio σ_w^2/P, and decays as $1/T$ (see corollary 1). When $\sigma_w^2/P = 0$, Theorem 1 specializes into [32, Thm 2.2].

In the next theorem, we establish a finite sample bound on the error for the instantaneous weights θ_T rather than the weighted average $\hat{\theta}_T$:

Theorem 2: Let $\{\boldsymbol{\theta}_t^n\}_{n=1}^N$ be the model parameters generated by COTAF according to (7) and (12) over R rounds, i.e., $t \in \{0, 1, \dots T-1\}$ with $T = RH$. Then, when *AS1–AS3* hold and the step sizes are set to $\eta_t = \frac{2}{\mu(\gamma+t)}$, for $\gamma \geq \max(\frac{8L\rho}{\mu}, H)$, it holds that:

$$\mathbb{E}[F(\theta_T)] - F(\theta^*) \leq \frac{2L\max\left(4\,C, \mu^2\gamma\delta_0\right)}{\mu^2(T+\gamma)}. \quad (16)$$

where $C = B + \frac{4dH^2G^2\sigma_w^2}{PN^2}$, and $\delta_0 = \|\theta_0 - \theta^\star\|^2$ is the initial guess accuracy.

Proof: The proof is given in Appendix B. □

The proofs for both Theorems 1–2 follow the same first steps. Yet in the derivation of Theorem 2 an additional relaxation was applied, implying that the bound in (16) is less tight than (15). For the noise-free case, i.e., $\sigma_w^2/P = 0$, Theorem 2 coincides with [12, Thm. 1].

Theorems 1 and 2 characterize of the effect of three sources of error on the rate of convergence: The accuracy of the initial guess $\|\theta_0 - \theta^\star\|^2$; the effect of statistical heterogeneity encapsulated in Γ, which is linear in B and C; and the noise-to-signal ratio σ_w^2/P induced by the wireless channel. In particular, in (16) all of these quantities, which potentially degrade the accuracy of the learned global model, contribute to the error bound in a manner proportional to $1/(T+\gamma)$, i.e., which decays as the number of rounds grows. The same observation also holds for (15), in which the aforementioned terms contribute in a manner that decays at an order proportional to $1/T$. The fact that the error due to the noise, encapsulated in σ_w^2/P, decays with the number of iterations, indicates the ability of COTAF to mitigate the harmful effect of the MAC noise, as discussed next.

Comparing (16) to the corresponding bound for local SGD with heterogeneous data and without communication constraints in [12, Thm. 1], i.e., over orthogonal channels as in (5) without noise, we observe that the bound takes a similar form as that in [12, Eq. (5)]. The main difference is in the additional term that depends on the noise-to-signal ratio σ_w^2/P in the constant C, which does not appear in the noiseless case in [12]. Consequently, the fact that COTAF communicates over a noisy channel induces an additional term that can be written as σ_w^2/P times some factor which, as the number of FL rounds R grows, is dominated by $\frac{H^2}{N^2(T+\gamma)}$. This implies that the time-varying precoding and aggregation strategy implemented by COTAF results in a gradual decay of the noise effect, and allows its contribution to be further mitigated by increasing the number of users N. Furthermore, Theorems 1–2 yield the same asymptotic convergence rate to that observed for noiseless local SGD in [12], as stated in the following corollary:

Corollary 1: COTAF achieves an asymptotic convergence rate of $\mathcal{O}(\frac{1}{T})$.

Proof: The corollary follows directly from (15) and (16) by letting T grow arbitrarily large while keeping the number of SGD iterations per round H fixed. □

Corollary 1 implies that COTAF allows OTA FL to achieve the same asymptotic convergence rate as local SGD with a strongly convex objective and without communication constraints [12], [32]. This advantage of COTAF adds to its ability to exploit the temporal and spectral resources of the wireless channel, allowing communication at higher throughput compared to conventional designs based on orthogonal communications, as discussed in Subsection III-D.

C. Extension to Fading Channels

We next show how COTAF can be extended to fading MACs of the form (6), while preserving its proven convergence. As detailed in Subsection II-B, we focus on scenarios in which the participating entities have CSI.

In fading MACs, the signal transmitted by each user undergoes a fading coefficient denoted $h_t^n e^{j\phi_t^n}$ (6). Following the scheme proposed in [19] for conveying sparse model updates, each user can utilize its CSI to cancel the fading effect by amplifying the signal by its inverse channel coefficient. However, weak channels might cause an arbitrarily high amplification, possibly violating the transmission power constraint (4). Therefore, a threshold h_{min} is set, and users observing fading coefficients of a lesser magnitude than h_{min} do not transmit in that communication round. As channels typically attenuate their signals, it holds that $h_{min} < 1$. Under this extension of COTAF, (8) becomes

$$\boldsymbol{x}_t^n = \begin{cases} \frac{\sqrt{\alpha_t}h_{min}}{h_t^n}e^{-j\phi_t^n}\left(\boldsymbol{\theta}_t^n - \boldsymbol{\theta}_{t-H}^n\right), & h_t^n > h_{min}, \\ 0, & h_t^n \leq h_{min}. \end{cases} \quad (17)$$

Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on July 29,2021 at 09:55:20 UTC from IEEE Xplore. Restrictions apply.

Here, $e^{-j\phi_t^n}$ is a phase correction term as in [20]. Note that the energy constraint (4) is preserved as $\mathbb{E}[\|\boldsymbol{x}_t^n\|^2] \leq P$.

To formulate the server aggregation, we let $\mathcal{K}_t \subset \mathcal{N}$ be the set of user indices whose corresponding channel at time t satisfies $h_t^n > h_{min}$. As the server has CSI, it knows $\mathcal{K}_t$, and can thus recover the aggregated model θ_t in a similar manner as in (11)–(12) via $\theta_t = \frac{\boldsymbol{y}_t}{|\mathcal{K}_t|\sqrt{\alpha_t}h_{min}} + \theta_{t-H}$, i.e.,

$$\theta_t = \frac{1}{|\mathcal{K}_t|}\sum_{n\in\mathcal{K}_t}\theta_t^n + \frac{N}{|\mathcal{K}_t|h_{min}}\boldsymbol{w}_t. \tag{18}$$

Comparing (18) to the corresponding equivalent formulation in (12), we note that the proposed extension of COTAF results in two main differences from the fading-free scenario: 1) the presence of fading is translated into an increase in the noise power, encapsulated in the constant $\frac{N}{|\mathcal{K}_t|h_{min}} > 1$; and 2) less models are aggregated in each round as $|\mathcal{K}_t| \leq N$. The set of participating users $\mathcal{K}_t$ depends on the distribution of the fading coefficients. Thus, in order to analytically characterize how the convergence is affected by fading compared to the scenario analyzed in Subsection III-B, we introduce the following assumption:

AS4 At each communication round, the participating users set $\mathcal{K}_t$ contains $K \leq N$ users and is uniformly distributed over all the subsets of $\mathcal{N}$ of cardinality K.

Note that Assumption *AS4* can be imposed by a simple distributed mechanism using an opportunistic carrier sensing [39]. Specifically, each user maps its h_t^n to a backoff time b_t^n based on a predetermined common function $f(h)$, which is a decreasing function with h (truncated at h_{min}). Then, each user with $h_t^n \geq h_{min}$ listens to the channel and transmits a low-power beacon when its backoff time expires, which can be sensed by other users. If K transmissions have been identified, the corresponding K users transmit their data signal to the server. Otherwise, the users wait (which occurs with a small probability as N increases, and h_{min} decreases) to the next time step. This mechanism guarantees $|\mathcal{K}_t| = K$ at each update. We point out that Assumption *AS4* is needed for theoretical analysis only.

Next, we characterize the convergence of the instantaneous global model, as stated in the following theorem:

Theorem 3: Let $\{\boldsymbol{\theta}_t^n\}_{n=1}^N$ be the model parameters generated by the extension of COTAF to fading channels over R rounds, i.e., $t \in \{0, 1, \ldots T-1\}$ with $T = RH$. Then, when *AS1–AS4* hold and the step sizes are set to $\eta_t = \frac{2}{\mu(\gamma+t)}$, for $\gamma \geq \max(\frac{8L\rho}{\mu}, H)$, it holds that:

$$\mathbb{E}[F(\theta_T)] - F(\theta^*) \leq \frac{2L\max\left(4(\tilde{C}+D), \mu^2\gamma\delta_0\right)}{\mu^2(T+\gamma)}. \tag{19}$$

where $\tilde{C} = B + \frac{4dH^2G^2\sigma_w^2}{PK^2h_{min}^2}$ and $D = \frac{4(N-K)}{K(N-1)}H^2G^2$.

Proof: The proof is given in Appendix C. □

Comparing Theorem 3 to the corresponding convergence bound for fading-free channels in Theorem 3 reveals that the extension of COTAF allows the trained model to maintain its asymptotic convergence rate of $O(\frac{1}{T})$ also in the presence of fading channel conditions. However, the aforementioned differences in the equivalent global model due to fading are translated here into additive terms increasing the bound on the distance between the expected instantaneous objective $\mathbb{E}[F(\theta_T)]$ and its desired optimal value. In particular, the fact that not all users participate in each round induces the additional positive term D in (19), which equals zero when $K = N$ and grows as K decreases. Furthermore, the increased equivalent noise results in the additive term $\tilde{C}$ being larger than the corresponding symbol C in (16) due to the increased equivalent noise-to-signal ratio which stems from the scaling by h_{min} at the precoder and the corresponding aggregation at the server side. Despite the degradation due to the presence of fading, COTAF is still capable of guaranteeing convergence and approach the performance of fading and noise-free local SGD when training in light of a smooth convex objective in a federated manner, as also numerically observed in our simulation study in Section IV.

D. Discussion

COTAF is designed to allow FL systems operating over shared wireless channels to exploit the full spectral and temporal resources of the media. This is achieved by accounting for the task of aggregating the local models into a global one as a form of OTA computation [23]. Unlike conventional orthogonality-based transmissions, such as FDM and TDM, in OTA FL the available band and/or transmission time of each user does not decrease with the number of users N, allowing the simultaneous participation of a large number of users without limiting the throughput of each user. Compared to previous strategies for OTA FL, COTAF allows the implementation of local SGD, which is arguably the most widely used FL scheme, over wireless MACs with proven convergence. This is achieved without having to restrict the model updates to be sparse with an identical sparsity pattern shared among all users [18], [19], or requiring the users to repeatedly compute the gradients over the full data set as in [20].

A major challenge in implementing SGD as an OTA computation stems from the presence of the additive channel noise, whose contribution does not decay over time [30]. Under strongly convex objectives, noisy distributed learning can be typically shown to asymptotically converge to some distance from the minimal achievable loss, unlike noise-free local SGD which is known to converge to desired $F^\star$ at a rate of $\mathcal{O}(\frac{1}{T})$ [32]. COTAF involves additional precoding and scaling steps which result in an effective decay of the noise contribution, thus allowing to achieve convergence results similar to noise-free local SGD with strongly convex objectives while operating over shared noisy wireless channels. The fact that COTAF mitigates the effect of noise in a gradual manner allows benefiting from the advantages of such noise profiles under non-convex objectives, where a controllable noise level was shown to facilitate convergence by reducing the probability of the learning procedure being trapped in local minima [34], [37]. This behavior is numerically demonstrated in Section IV.

Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on July 29,2021 at 09:55:20 UTC from IEEE Xplore. Restrictions apply.

COTAF consists of an addition of simple precoding and scaling stages to local SGD. This precoding stage is necessary for assuring a steady convergence rate, while keeping power consumption under control. Implementing the time-varying precoding in (9) implies that every user has to know $\max_n \mathbb{E}[||\theta_t^n - \theta_{t-H}^n||^2]$, for each communication round $t \in \mathcal{H}$. When operating with a decaying step size, as is commonly required in FL, and when *AS3* holds, this term is upper bounded by $H^2\eta_{t-H}^2 G^2$ (see Lemma A.2 in Appendix A), and the upper bound can be used instead in (9), while maintaining the convergence guarantees of Theorems 1–2. Alternatively, since α_t should be proportional to the inverse of the maximal difference of consecutively transmitted models, one can numerically estimate these values by performing offline simulation over a smaller, global data set. Such data can be either obtained from server-side data, or from some level of initial sharing of non-private data [31]. Once these values are numerically computed, the server can distribute them to the users over the downlink channel. Finally, one can also have the users convey their instantaneous model updates norm $||\theta_t^n - \theta_{t-H}^n||^2$ to the server before each communication round, as proposed in [40] for user selection. Doing so allows the server to distribute $\max_n ||\theta_t^n - \theta_{t-H}^n||^2$ to the users, to be used for setting α_t, while inducing only a minor communication overhead since the exchanged quantities are positive scalars.

COTAF involves analog transmissions over MAC, which allows the superposition carried out by the MAC to aggregate the parameters as required in FL. As a result, COTAF is subject to the challenges associated with such signalling, e.g., the need for accurate synchronization among all users. Finally, OTA FL schemes such as COTAF require the participating users to share the same wireless channel, i.e., reside in the same geographical area, while FL systems can be trained using data aggregated from various locations. We conjecture that COTAF can be combined in multi-stage FL, such as clustered FL [8]. We leave this for future study.

IV. Numerical Evaluations

In this section, we provide numerical examples to illustrate the performance of COTAF in two different settings. We begin with a scenario of learning a linear predictor of the release year of a song from audio features in Subsection IV-A. In this setup, the objective is strongly convex, and the model assumptions under which COTAF is analyzed hold. In the second setting detailed in Subsection IV-B, we consider a more involved setup, in which the loss surface with respect to the learned weights is not convex. Specifically, we train a CNN for classification on the CIFAR-10 dataset.

A. Linear Predictor Using the Million Song Dataset

We start by examining COTAF for learning how to predict the release year of a song from audio features in an FL manner. We use the Million Song Dataset [33], that contains songs which are mostly western, commercial tracks ranging from 1922 to 2011. Each song is associated with a release year and 90 audio attributes. Consequently, each data sample $\boldsymbol{s}$ takes the form $\boldsymbol{s} = \{\boldsymbol{s}_s, s_y\}$, where $\boldsymbol{s}_s$ is the audio attributes vector and s_y is the year. The system task is to train a linear estimator θ with $d = 90$ entries in an FL manner using data available at N users, where each user has access to $D_n = 9200$ samples. The predictor is trained using the regularized linear least-squares loss, given by:

$$f(\boldsymbol{\theta}, \{\boldsymbol{s}_s, s_y\}) = \frac{1}{2}(\boldsymbol{s}_s^T\boldsymbol{\theta} - s_y)^2 + \frac{\lambda}{2}||\boldsymbol{\theta}||^2, \qquad (20)$$

where we used $\lambda = 0.5$. We note that the loss measure (20) is strongly convex and has a Lipschitz gradient, and thus satisfies the conditions of Theorem 1. In every FL round, each user performs H SGD steps (7) where the step size is set via Theorem 1. In particular, the parameters L and μ are numerically evaluated before transmitting the model update to the server over the MAC. The precoding coefficient α_t is computed via (9) using numerical averaging, i.e., we carried out an offline simulation of local SGD without noise and with 20% of the data samples, and computed the averaged norm of the resulting model updates.

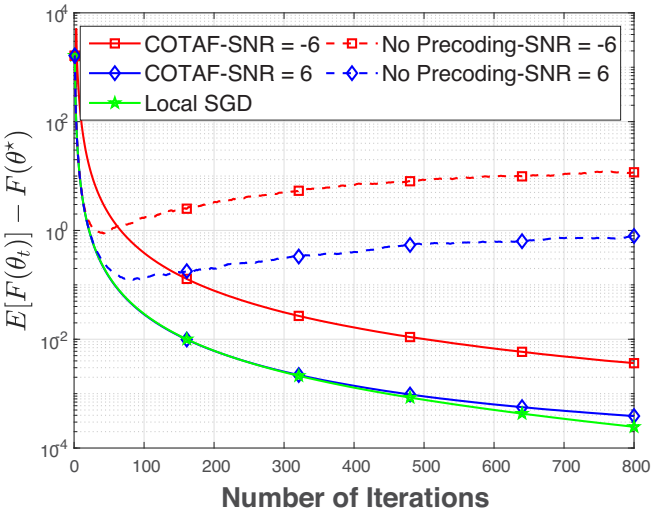

Fig. 3. Linear predictor, Million Song dataset, $H = 40$, $N = 50$.

We numerically evaluate the gap from the achieved expected objective and the loss-minimizing one, i.e., $\mathbb{E}[F(\theta_t)] - F^\star$. Using this performance measure, we compare COTAF to the following FL methods: (i) Local SGD, in which every user conveys its model updates over a noiseless individual channel; (ii) Non-precoded OTA FL, where every user transmits its model updates over the MAC without time-varying precoding (9) and with a constant amplification as in [20], i.e., $\boldsymbol{x}_t^n = P(\theta_t^n - \theta_{t-H}^n)$. The stochastic expectation is evaluated by averaging over 50 Monte Carlo trials, where in each trial the initial θ_0 is randomized from zero-mean Gaussian distribution with covariance $5\boldsymbol{I}_d$.

We simulate MACs with signal-to-noise ratios (SNRs) of $P/\sigma_w^2 = -6\,\text{dB}$ and $P/\sigma_w^2 = 6\,\text{dB}$. In Fig. 3, we present the performance evaluation when the number of users is set to $N = 50$ and the number of SGD steps is $H = 40$. It can be seen in Fig. 3 that COTAF achieves performance within a minor gap of $5.8 \cdot 10^{-}4$ and $3.2 \cdot 10^{-3}$ in the objective function for SNRs 6 and -6, respectively, from that of local SGD carried out over ideal orthogonal noiseless channels. For comparison, the performance of non-precoded OTA FL is within a much larger

Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on July 29,2021 at 09:55:20 UTC from IEEE Xplore. Restrictions apply.

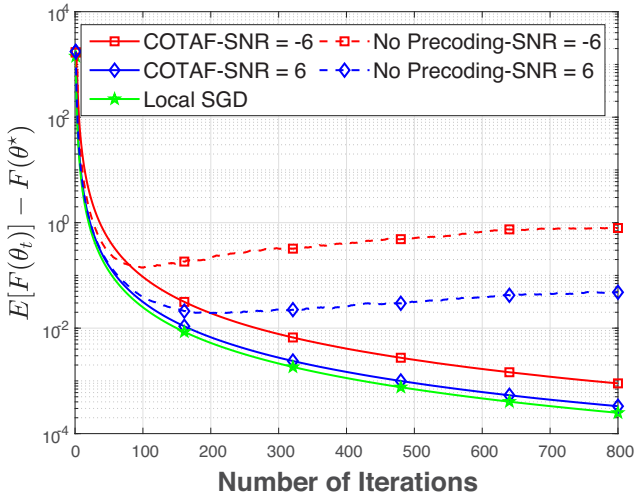

Fig. 4. Linear predictor, Million Song dataset, $H = 40$, $N = 200$.

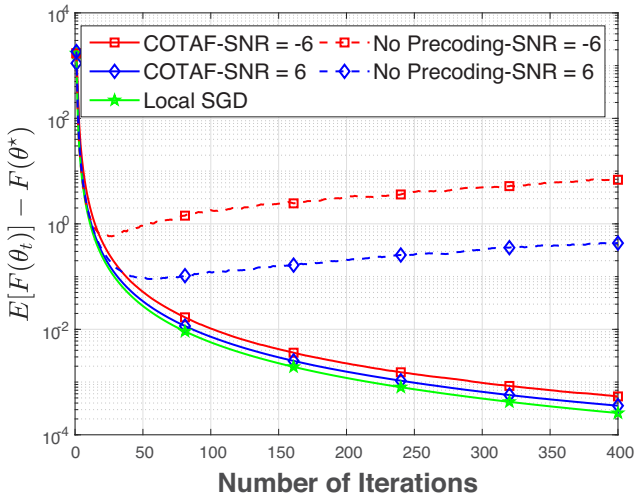

Fig. 5. Linear predictor, Million Song dataset, $H = 80$, $N = 50$.

gap of 0.2 for SNR of 6 dB, and 11.5 for SNR of -6 dB. This improved performance of COTAF is achieved without requiring the users to divide the spectral and temporal channel resources among each other, thus to communicate at higher throughput uplink communications as compared to the local SGD. This is due to the precoding scheme of COTAF, which allows gradually mitigating the effect of channel noise, while OTA FL without such time-varying precoding results in a dominant error floor due to presence of non-vanishing noise.

Next, we repeat the simulation study of Fig. 3 while increasing the number of users to be $N = 200$ in Fig. 4, and with setting the number of SGD steps to $H = 80$ in Fig. 5. The number of gradient computations $T = RH$ and the overall number of training samples ND_n is kept constant throughout the simulations. Figs. 4–5 thus demonstrate the dependence of COTAF performance on two key system parameters: The number of users, N, and the number of SGD steps, H, between communication rounds.

Observing Fig. 4 and comparing it to Fig. 3, we note that increasing the number of users improves the performance of both OTA FL schemes, despite the fact that each user holds less training samples. In particular, COTAF effectively coincides with the performance of noise-free local SGD here, while the non-precoded OTA FL achieves improved performance compared to the setting with $N = 50$, yet it is still notably outperformed by COTAF. The gain in increasing the number of users follows from the fact that averaging over a larger number of users at the server side mitigates the contribution of the channel noise, as theoretically established for COTAF in Subsection III-B. When using orthogonal transmissions, as implicitly assumed in conventional local SGD, increasing the number of users implies that the channel resources must be shared among more users, hence the throughput of each users decreases. However, in OTA FL the throughput is invariant of the number of users. Comparing Fig. 5 to Fig. 3 reveals that increasing the SGD steps H can improve the performance of OTA FL as the channel noise is induced less frequently. Nonetheless, the gains here are far less dominant than those achievable by allowing more users to participate in the FL procedure, as observed in Fig. 4. The results depicted in Figs. 3–5 demonstrate the benefits of COTAF, as an OTA FL scheme which accounts for both the convergence properties of local SGD as well as the unique characteristics of wireless communication channels.

Next, we simulate the effect of fading channels on COTAF. In particular, we apply the extension of COTAF to fading channels detailed in Subsection III-C and compare the results with ECESA-DSGD and GBMA, which are OTA FL schemes for fading MACs proposed in [19] and [20], respectively. In particular, both ECESA-DSGD and GBMA are designed for a single gradient transmission, i.e., $H = 1$, where ECESA-DSGD incorporates an error accumulation technique to retain the accuracy of local stochastic gradients, while GBMA transmits the non-stochastic gradients computed over the complete local dataset. To guarantee fair comparison with COTAF, which is based on local SGD optimization, we simulate a variant of these methods, which transmits the local models at each transmission round of H local iterations.

The MAC input-output relationship is given by (6), and block fading channel coefficients $\{h_t^n\}$ are sampled from a Rayleigh distribution in an i.i.d. fashion, while the remaining parameters are the same as those used in the scenario simulated in Fig. 3. The threshold h_{min} in (17) is set such that on average 40 out of the $N = 50$ users participate in each communication round. For fairness, the same conditions are applied in the no precoding setting, i.e., the users utilize their CSI to cancel the effect of the channel as in [19]. This comparison illustrates the significance of the dedicated precoding introduced by COTAF.

The results, depicted in Fig. 6, demonstrate that COTAF maintains its ability to approach the performance of noise-free local SGD, observed in Figs. 3–5 for additive noise MACs. We also note that the error accumulation incorporated by ECESA-DSGD allows it to achieve an accuracy which is only slightly improved compared to OTA FL with no precoding. The gain of GBMA, which computes full gradients at the users with increased computational complexity, over non-precoded local SGD, is more dominant compared to that of ECESA-DSGD. Nonetheless, both previously proposed OTA FL schemes are

Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on July 29,2021 at 09:55:20 UTC from IEEE Xplore. Restrictions apply.

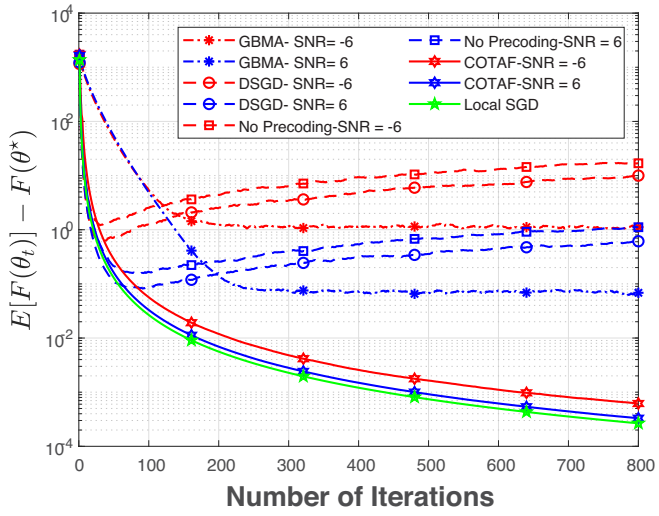

Fig. 6. Linear predictor, Million Song dataset, $H = 40$, $N = 50$, Rayleigh fading channels.

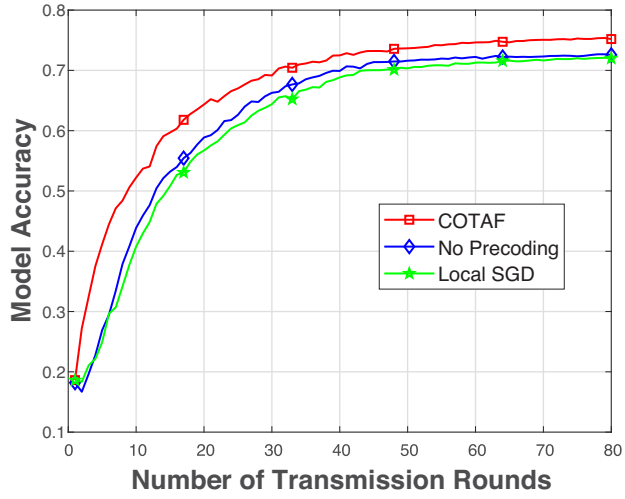

Fig. 7. CNN, CIFAR-10 dataset, i.i.d. data.

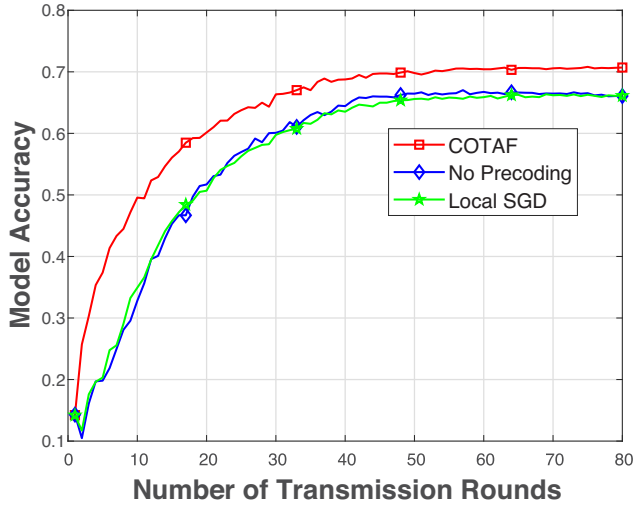

Fig. 8. CNN, CIFAR-10 dataset, heterogeneous data.

notably outperformed by COTAF, illustrating the importance of the time-varying precoder in handling the effect of channel noise.

B. CNN Classifier Using the CIFAR-10 Dataset

Next, we consider an image classification problem, based on the CIFAR-10 dataset, which contains train and test images from ten different categories. The classifier model is the DNN architecture detailed in [41], which consists of three conventional layers and two fully-connected layers. When trained in a centralized setting, this architecture achieves an accuracy of roughly 70% [41]. Here, we train this network to minimize the empirical cross-entropy loss in an FL manner, where the data set is distributed among $N = 10$ users. Each user holds 5000 images, and carries out its local training with a minibatch size of 60 images, while aggregation is done every $H = 84$ iterations over a MAC with SNR of -4 dB. We consider two divisions of the training data among the users: *i.i.d. data*, where we split the data between the users in an i.i.d fashion, i.e. each user holds the same amount of figures from each class; and *heterogeneous data*, where approximately 20% of the training data of each user is associated with a single label, which differs among the different users, the remaining 80% of the training data is uniformly distributed between all labels. This division causes heterogeneity between the users, as each user holds more images from a unique class. Under both divisions, an image can only be assigned to a single user, i.e., the data is private and not shared between users and the server. The model accuracy versus the transmission round achieved for the considered FL schemes is depicted in Figs. 7–8 for the i.i.d. case and the heterogeneous case, respectively.

Observing Figs. 7 and 8, we note that the global model trained using COTAF converges to an accuracy of approximately 73% and 69% respectively. This implies that training under an i.i.d distribution achieves a model whose accuracy is larger by 4% compared to being trained using non-i.i.d local data sets. This is achieved while allowing each user to fully utilize its available temporal and spectral channel resources, thus communicating at higher throughput as compared to orthogonal transmissions. Furthermore, we point out the following advantages of COTAF when applied to CIFAR-10:

1) COTAF Achieves the Desired Sublinear Convergence Rate: While the objective in training the CNN to minimize the cross-entropy loss is not a convex function of the weights, we observe the same rate of convergence for COTAF as that of noise-free local SGD. This result suggests a generalization of the theoretical analysis for the convex case and indicates that even in cases in which assumptions *AS1–AS3* do not hold, COTAF is still able to converge in a sub-linear rate. We deduce that COTAF can be applied in settings less restrictive than the analyzed case introduced in Subsection III-B and still achieve good results, as numerically illustrated in the current study.

2) COTAF Benefits From the Presence of Noise: The simulation results indicate that the additive noise caused by the channel improves the convergence rate and generalization of

Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on July 29,2021 at 09:55:20 UTC from IEEE Xplore. Restrictions apply.

the CNN model. The fact that COTAF gradually mitigates the effective noise allows it to benefit from its presence in non-convex settings, while notably outperforming direct OTA FL with no time-varying precoding operating in the same channel. Here, COTAF not only demonstrates an improved learning curve compared to local OTA FL without precoding and local SGD, but also results in an improved model once convergence is achieved. Specifically, we observe that after 30 transmission rounds, COTAF allows to train a model with accuracy of 70% and 65% for the i.i.d. and heterogeneous data divisions, respectively. The corresponding accuracy of OTA FL without such precoding are 66% and 60%, and local SGD achieves 65% and 60% after 30 transmission rounds. The resulting gains of COTAF are thus consistently around 5% in accuracy over both competing FL approaches. These gains follow from the fact that the presence of noise whose level gradually decreases when training DNNs is known to have positive effects such as reducing overfitting and avoiding local minima [34], [37], [42]. Furthermore, for the heterogeneous data case in Fig. 8, the gap between COTAF and local SGD is increased as compared to the i.i.d case in Fig. 7. This indicates that the noise has a smoothing effect as well. It allows better generalizations in the non-i.i.d setting, which are exploited by COTAF in a manner that contributes to its accuracy more effectively as compared to OTA FL with no precoding.

V. Conclusions

In this work we proposed the COTAF algorithm for implementing FL over wireless MACs. COTAF maintains the convergence properties of local SGD with heterogeneous data across users, with convex objectives carried out over ideal channels, without requiring the users to divide the channel resources. This is achieved by introducing a time-varying precoding and scaling scheme which facilitates the aggregation and gradually mitigates the noise effect. We prove that for convex objectives, models trained using COTAF with heterogeneous data converge to the loss minimizing model with the same asymptotic convergence rate of local SGD over orthogonal channels, thus theoretically guaranteeing the effectiveness of COTAF. Our numerical study demonstrates the ability of COTAF to learn accurate models over wireless channels using non-synthetic datasets. Furthermore, the simulation results empirically demonstrate the effectiveness of COTAF by showing that it converges in non-convex settings in a sub-linear rate, and outperforms not only OTA FL without precoding, but also the local SGD algorithm in carried out in an orthogonal fashion without errors induced by the channel.

Appendix

A. Proof of Theorem 1

In the following, we detail the proof of Theorem 1, introduced in Subsection III-B. The intermediate derivations detailed below are used in proving Theorem 2 in Appendix B as well. The outline of the proof is as follows: First, we define a virtual sequence $\{\bar{\theta}_t\}$ that represents the averaged parameters over all users at every iteration in (A.1), i.e. as if the local SGD framework is replaced with mini-batch SGD. While $\bar{\theta}_t$ can not be explicitly computed at each time instance by any of the users or the server, it facilitates utilizing bounds established for mini-batch SGD, as was done in [12], [32]. Next, we provide in Lemma A.1 a single step recursive bound for the error $E[||\bar{\theta}_t - \theta^\star||^2]$. The bound consists of four terms, in Lemmas A.2 and A.3 we upper bound these quantities. Finally, we obtain a non-recursive bound from the recursive expression in Lemma A.4, with which we prove Theorem 1.

Recursive error formulation: Following the steps used in the corresponding convergence analysis of FL without communication constraints [12], [32], we first define the virtual sequence $\{\bar{\theta}_t\}_{t\geq 0}$. Broadly speaking, $\{\bar{\theta}_t\}$ represents the weights obtained when the weights trained by the users are aggregated and averaged over the true channel on each H SGD steps, and over a virtual noiseless channel on the remaining SGD iterations. This virtual sequence is given by

$$\bar{\theta}_t \triangleq \frac{1}{N}\sum_{n=1}^{N}\theta_t^n + \frac{1}{\sqrt{\alpha_t}N}\tilde{\boldsymbol{w}}_t \mathbb{1}_{t\in\mathcal{H}}, \tag{A.1}$$

where $\mathbb{1}_{(\cdot)}$ is the indicator function. Rearranging (A.1) to fit our transmission scheme yields:

$$\bar{\theta}_t = \bar{\theta}_{t-H} + \frac{1}{\sqrt{\alpha_t}N}\sum_{n=1}^{N}\sqrt{\alpha_t}\left(\theta_t^n - \bar{\theta}_{t-H}\right) + \boldsymbol{w}_t\mathbb{1}_{t\in\mathcal{H}}, \tag{A.2}$$

with $\bar{\theta}_t \triangleq \theta_0$ for $t \leq 0$. The scaled noise $\boldsymbol{w}_t = \frac{1}{\sqrt{\alpha_t}N}\tilde{\boldsymbol{w}}_t$, and the sequence $\{\theta_t^n\}_{t\geq 0}$ are defined in Subsection III-A.

Notice that $\{\bar{\theta}_t\}$ is not computed explicitly, and that $\bar{\theta}_t = \theta_t^n$ for each $n \in \mathcal{N}$ whenever $t \in \mathcal{H}$. We also define

$$\boldsymbol{g}_t \triangleq \frac{1}{N}\sum_{n=1}^{N}\nabla f_{i_t^n}(\theta_t^n), \quad \bar{\boldsymbol{g}}_t \triangleq \frac{1}{N}\sum_{n=1}^{N}\nabla F(\theta_t^n). \tag{A.3}$$

Since the indices i_t^n used in each SGD iteration are uniformly distributed, it follows that $\mathbb{E}[\boldsymbol{g}_t] = \bar{\boldsymbol{g}}_t$. By writing $\bar{\boldsymbol{w}}_t \triangleq \boldsymbol{w}_{t+1}\mathbb{1}_{t+1\in\mathcal{H}} - \boldsymbol{w}_t\mathbb{1}_{t\in\mathcal{H}}$, we have that

$$\bar{\theta}_{t+1} = \bar{\theta}_t - \eta_t\boldsymbol{g}_t + \bar{\boldsymbol{w}}_t. \tag{A.4}$$

The equivalent noise vector $\bar{\boldsymbol{w}}_t$ is zero-mean and satisfies

$$\mathbb{E}[\|\bar{\boldsymbol{w}}_t\|^2] \leq \frac{d\sigma_w^2}{N^2\min(\alpha_t, \alpha_{t+1})}\mathbb{I}_t, \tag{A.5}$$

where $\mathbb{I}_t \triangleq \mathbb{1}_{(t\in\mathcal{H})\cup(t+1\in\mathcal{H})}$. Theorem 1 is obtained from definitions (A.1) and (A.3) via the following lemma:

Lemma A.1: Let $\{\theta_t^n\}$ and $\{\bar{\theta}_t\}$ be as defined in (13) and (A.1), respectively. Then, when *AS1–AS2* are satisfied and the SGD step size satisfies $\eta_t \leq \frac{1}{4L}$, it holds that

$$\begin{aligned}\mathbb{E}\left[||\bar{\theta}_{t+1} - \theta^\star||^2\right] &\leq (1-\mu\eta_t)\mathbb{E}\left[||\bar{\theta}_t - \theta^\star||^2\right] \\ &+ \eta_t^2\mathbb{E}\left[\left\|\boldsymbol{g}_t - \bar{\boldsymbol{g}}_t + \frac{\bar{\boldsymbol{w}}_t}{\eta_t}\right\|^2\right] - \frac{3}{2}\eta_t\mathbb{E}\left[F(\bar{\theta}_t) - F^\star\right] \\ &+ \frac{2}{N}\sum_{n=1}^{N}\mathbb{E}\left[||\bar{\theta}_t - \theta_t^n||^2\right] + 6L\eta_t^2\Gamma. \end{aligned} \tag{A.6}$$

Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on July 29,2021 at 09:55:20 UTC from IEEE Xplore. Restrictions apply.

Proof: Using the update rule we have:

$$||\bar{\theta}_{t+1}-\theta^\star||^2 = \|\bar{\theta}_t-\eta_t\bar{\boldsymbol{g}}_t-\theta^\star\|^2+\eta_t^2\left\|\bar{\boldsymbol{g}}_t-\boldsymbol{g}_t+\frac{\bar{\boldsymbol{w}}_t}{\eta_t}\right\|^2 + 2\eta_t\left\langle\bar{\theta}_t-\theta^\star-\eta_t\bar{\boldsymbol{g}}_t,\bar{\boldsymbol{g}}_t-\boldsymbol{g}_t+\frac{\bar{\boldsymbol{w}}_t}{\eta_t}\right\rangle. \quad (A.7)$$

Observe that $E[\langle\bar{\theta}_t-\theta^\star-\eta_t\bar{\boldsymbol{g}}_t,\bar{\boldsymbol{g}}_t-\boldsymbol{g}_t+\frac{\bar{\boldsymbol{w}}_t}{\eta_t}\rangle]=0$. Following the proof steps in [12, Lemma 1], we obtain $||\bar{\theta}_t-\eta_t\bar{\boldsymbol{g}}_t-\theta^\star||^2\le(1-\mu\eta_t)||\bar{\theta}_t-\theta^\star||^2+\frac{1}{N}\sum_{n=1}^N||\bar{\theta}_t-\theta_n^t||^2+A$, with $A\triangleq\frac{4L\eta_t^2}{N}\sum_{n=1}^N(f_n(\theta_t^n)-f_n^\star)-\frac{2\eta_t}{N}\sum_{n=1}^N f_n(\theta_t^n)-f_n(\theta^\star)$. The Value of A satisfies:

$$A=(4L\eta_t^2-2\eta_t)\frac{1}{N}\sum_{n=1}^N(f_n(\theta_t^n)-F^\star)+4L\eta_t^2\Gamma, \quad (A.8)$$

where we used the definitions of $F^\star$ and Γ. Notice that $4L\eta_t^2-2\eta_t\le\eta_t-2\eta_t\le-\eta_t$.

Next, we use the following inequality obtained in [12] $\frac{1}{N}\sum_{n=1}^N(f_n(\theta_t^n)-F^\star)\ge(F(\bar{\theta}_t)-F^\star)-\frac{1}{N}\sum_{n=1}^N[\eta_t L(f_n(\bar{\theta}_t)-f_n^\star)+\frac{1}{2\eta_t}||\theta_t^n-\bar{\theta}_t||^2]$. Substituting this into (A.8) and using the fact that $\eta_t L-1\le-\frac{3}{4}$, (2) $2\eta_t-4L\eta_t^2\le2\eta_t$ yields

$$A\le6\,L\eta_t^2\Gamma-\frac{3\eta_t}{2}(F(\bar{\theta}_t)-F^\star)+\frac{1}{N}\sum_{n=1}^N||\theta_t^n-\bar{\theta}_t||^2. \quad (A.9)$$

Consequently, we have that:

$$||\bar{\theta}_t-\theta^\star-\eta_t\bar{\boldsymbol{g}}_t||^2\le(1-\mu\eta_t)||\bar{\theta}_t-\theta^\star||^2 +\frac{2}{N}\sum_{n=1}^N||\bar{\theta}_t-\theta_t^n||^2+6\,L\eta_t^2\Gamma-\frac{3\eta_t}{2}(F(\bar{\theta}_t)-F^\star). \quad (A.10)$$

Finally, by taking the expected value of both sides of (A.7) and using (A.10) we complete the proof. ■

Upper bounds on the additive terms: Next, we prove the theorem by bounding the summands constituting the right hand side of (A.6). First, we bound $\mathbb{E}[\|\boldsymbol{g}_t-\bar{\boldsymbol{g}}_t+\frac{\bar{\boldsymbol{w}}_t}{\eta_t}\|^2]$, as stated in the following lemma:

Lemma A.2: When the step size sequence $\{\eta_t\}$ consists of decreasing positive numbers satisfying $\eta_t\le2\eta_{t+H}$ for all $t\ge0$ and *AS3* holds, then

$$\mathbb{E}\left[\left\|\boldsymbol{g}_t-\bar{\boldsymbol{g}}_t+\frac{\bar{\boldsymbol{w}}_t}{\eta_t}\right\|^2\right]\le\frac{1}{N^2}\sum_{n=1}^N M_n^2+\frac{4\,dH^2\,G^2\sigma_w^2}{PN^2}\mathbb{I}_t.$$

Proof: The lemma follows since the noise term $\bar{\boldsymbol{w}}_t$ is zero-mean and independent of the stochastic gradients, hence $\mathbb{E}[\|\bar{\boldsymbol{g}}_t-\boldsymbol{g}_t+\frac{\bar{\boldsymbol{w}}_t}{\eta_t}\|^2]\le\frac{1}{N^2}\sum_{n=1}^N M_n^2+\mathbb{E}[\|\frac{\bar{\boldsymbol{w}}_t}{\eta_t}\|^2]$, by *AS3*. From (A.5) we obtain

$$\mathbb{E}\left[\left\|\frac{\bar{\boldsymbol{w}}_t}{\eta_t}\right\|^2\right]\le\frac{d\sigma_w^2\mathbb{I}_t}{\eta_t^2\,N^2\min(\alpha_t,\alpha_{t+1})}. \quad (A.11)$$

Next, we bound $\frac{1}{\alpha_t}=\frac{1}{P}\max_n\mathbb{E}[||\theta_t^n-\theta_{t-H}^n||^2]$ via:

$$\frac{1}{\alpha_t}\overset{(a)}{\le}\frac{1}{P}\max_n\left(H\eta_{t-H}\sum_{t'=t-H}^{t-1}\mathbb{E}\left[||\nabla f_{i_h^k}(\bar{\theta}_{t'}^n)||^2\right]\right) \overset{(b)}{\le}\frac{1}{P}H^2\eta_{t-H}^2\,G^2\overset{(c)}{\le}\frac{1}{P}4H^2\eta_t^2\,G^2, \quad (A.12)$$

where (a) follows from (7), using the inequality $\|\sum_{t'=t-H}^t\boldsymbol{r}_t\|^2\le H\sum_{t'=t-H}^t\|\boldsymbol{r}_t\|^2$, which holds for any multivariate sequence $\{\boldsymbol{r}_t\}$, while noting that the step size is monotonically non-increasing; (b) holds by *AS3*; and (c) holds as $\eta_t\le2\eta_{t+H}$ for all $t\ge0$. Finally, notice that

$$\frac{1}{\min(\alpha_t,\alpha_{t+1})}\le\frac{1}{P}4H^2\eta_t^2\,G^2, \quad (A.13)$$

as $\{\eta_t\}$ is monotonically decreasing. Substituting (A.13) into (A.11) completes the proof. ■

In the next lemma, we bound $\frac{1}{N}\sum_{n=1}^N\mathbb{E}[||\bar{\theta}_t-\theta_t^n||^2]$:

Lemma A.3: When the step size sequence $\{\eta_t\}$ consists of decreasing positive numbers satisfying $\eta_t\le2\eta_{t+H}$ for all $t\ge0$ and *AS3* holds, then $\mathbb{E}[||\bar{\theta}_t-\theta_t^n||^2]\le4\eta_t^2\,G^2H^2$.

Proof: The lemma follows directly from [32, Lem. 3.3]. ■

Obtaining a non-recursive convergence bound: Combining Lemmas A.1–A.3 yields a recursive relationship which allows us to characterize the convergence of COTAF. To complete the proof, we next establish the convergence bound from the recursive equations, based on the following lemma:

Lemma A.4: Let $\{\delta_t\}_{t\ge0}$ and $\{e_t\}_{t\ge0}$ be two positive sequences satisfying

$$\delta_{t+1}\le(1-\mu\eta_t)\delta_t-\eta_t e_t A+\eta_t^2 B+\eta_t^2 D_t, \quad (A.14)$$

for $\eta_t=\frac{4}{\mu(a+t)}$ with constants $A>0,B,C\ge0,\mu>0,a>1$ and $D_t=D\mathbb{1}_{t\in\mathcal{H}}$. Then, for any positive integer R it holds that $\frac{A}{S_R}\sum_{r=1}^R\beta_{rH}e_{rH}\le\frac{\mu a^3}{4S_R}\delta_0+\frac{2R(H+1)}{\mu S_R}B(2a+H+R-1)+\frac{2R}{\mu S_R}D(2a+HR+H)$, for $\beta_t=(a+t)^2$, $T=RH$ and $S_R=\sum_{r=1}^R\beta_{rH}\ge\frac{1}{3H}T^3=\frac{1}{3}T^2R$.

Proof: To prove the lemma, we first note that by [36, Eqn.(45)], it holds that $(1-\mu\eta_t)\frac{\beta_t}{\eta_t}\le\frac{\mu(a+t-1)^3}{4}=\frac{\beta_{t-1}}{\eta_{t-1}}$. Therefore, multiplying (A.14) by $\frac{\beta_t}{\eta_t}$ yields

$$\delta_{t+1}\frac{\beta_t}{\eta_t}\le(1-\mu\eta_t)\frac{\beta_t}{\eta_t}\delta_t-\beta_t e_t A+\beta_t\eta_t B+\beta_t\eta_t D_t. \quad (A.15)$$

Next, we extract the relations between two sequential rounds, i.e. δ_t,δ_{t+H}. Note that in each round a single D_t is activated, i.e., only a single entry in the set $\{D_\tau\}_{\tau=t}^{t+H}$ is non-zero. Therefore, by repeating the recursion (A.15) over H time instances, we get

$$\delta_{t+H}\frac{\beta_{t+H-1}}{\eta_{t+H-1}}\le(1-\mu\eta_t)\frac{\beta_t}{\eta_t}\delta_t+\frac{2B}{\mu}(H+1)(H+2t+2a) -\beta_t e_t A+\beta_{t_0}\eta_{t_0}D, \quad (A.16)$$

where t_0 is the only time instance in the interval $[t,t+H)$ such that $t_0\in\mathcal{H}$. In the last inequality we used the fact that $A\beta_t e_t>0$. Recursively applying (A.16) R times yields: $\delta_R\frac{\beta_R}{\eta_R}\le(1-\mu\eta_0)\frac{\beta_0}{\eta_0}\delta_0-\sum_{r=1}^R\beta_{rH}e_{rH}A+\sum_{r=1}^R\beta_{rH}\eta_{rH}D+\sum_{r=1}^R\frac{2B}{\mu}(H+1)(H+2rH+2a)$. As $\delta_t\frac{\beta_t}{\eta_t}>0$ for each t, this implies that $A\sum_{r=1}^R\beta_{rH}e_{rH}\le\frac{\beta_0}{\eta_0}\delta_0+\frac{2B(H+1)}{\mu}\sum_{r=1}^R(H+2rH+2a)+\sum_{r=1}^R\beta_{rH}\eta_{rH}D$.

Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on July 29,2021 at 09:55:20 UTC from IEEE Xplore. Restrictions apply.

Next, recalling that $\beta_t\eta_t = \frac{4(a+t)}{\mu}$, we obtain:

$$\sum_{r=1}^{R}\beta_{rH}\eta_{rH}D = \frac{4}{\mu}D\left(aR + \frac{HR}{2} + \frac{HR^2}{2}\right). \quad \text{(A.17)}$$

For the current setting of β_t and η_t it holds that $\frac{\beta_0}{\eta_0} = \frac{\mu a^3}{4}$. Further, $\sum_{r=1}^{R}(H + 2rH + 2a) = R(2a + T + 2H)$. Substituting this and (A.17) into the above inequality proves the lemma. ■

We complete the proof of the theorem by combining Lemmas A.1–A.4 as follows: By defining $\delta_t \triangleq \mathbb{E}[||\bar{\theta}_t - \theta^\star||^2]$, it follows from Lemma A.1 combined with the bounds stated in Lemmas A.2–A.3 that:

$$\delta_{t+1} \le (1-\mu\eta_t)\delta_t + \eta_t^2\left(\frac{1}{N^2}\sum_{n=1}^{N}M_n^2 + \frac{4\,dH^2\,G^2\sigma_w^2}{PN^2}\mathbb{I}_t\right) - \frac{3}{2}\eta_t\mathbb{E}[F(\bar{\theta}_t) - F^\star] + 8\eta_t^2H^2\,G^2 + 6L\eta_t^2\Gamma. \quad \text{(A.18)}$$

In the non-trivial case where $H > 1$, at most one element of $\{t_0+1, t_0\}$ can be in $\mathcal{H}$ for any t_0. Therefore, without loss of generality, we reduce the set over which the indicator function in (A.18) is defined to be $\{t \in \mathcal{H}\}$. By defining

$$A \triangleq \frac{3}{2};\quad B \triangleq 8H^2\,G^2 + \frac{1}{N^2}\sum_{n=1}^{N}M_n^2 + 6L\Gamma;$$

$$e_t \triangleq \mathbb{E}\left[F(\bar{\theta}_t) - F^\star\right];\quad D_t \triangleq \frac{4\,dH^2\,G^2\sigma_w^2}{PN^2}\mathbb{1}_{t\in\mathcal{H}},$$

and plugging these notations into Lemma A.4, we obtain $\frac{1}{S_R}\sum_{r=1}^{R}\beta_{rH}\mathbb{E}[F(\bar{\theta}_{rH}) - F^\star] \le \frac{\mu a^3}{6S_R}||\theta_0 - \theta^\star||^2 + \frac{4(T+R)}{3\mu S_R}(2a+H+R-1)B + \frac{16\,dTHG^2\sigma_w^2}{3\mu PN^2S_R}(2a+T+H)$. Finally, by the convexity of the objective function, it holds that $\mathbb{E}[F(\hat{\theta}_T) - F^\star] \le \frac{1}{S_R}\sum_{r=1}^{R}\beta_{rH}\mathbb{E}[F(\bar{\theta}_{rH}) - F^\star]$, thus proving Theorem 1. □

B. Proof of Theorem 2

The proof of Theorem 2 utilizes Lemmas A.1–A.3, stated in Appendix A, while formulating an alternative non-recursive bound compared to that used in Appendix A. To obtain the convergence bound in (16), we first recall the definition $\delta_t \triangleq \mathbb{E}\{\|\bar{\theta}_{t+1} - \theta^*\|^2\}$. When $t \in \mathcal{H}$, the term δ_t represents the ℓ_2 norm of the error in the weights of the global model. We can upper bound (A.18) and formulate the following recursive relationship on the weights error

$$\delta_{t+1} \le (1-\eta_t\mu)\delta_t + \eta_t^2\,C, \quad \text{(B.1)}$$

where $C = B + \frac{4\,dH^2\,G^2\sigma_w^2}{PN^2}$. The inequality is obtained from (A.18) since $-\eta_t e_t\mathbb{E}[F(\bar{\theta}_t) - F^*] \le 0$ and as $D\mathbb{1}_{t\in\mathcal{H}} \le D$, for $D \ge 0$. The convergence bound is achieved by properly setting the step-size and the FL systems parameters in (B.1) to bound δ_t, and combining the resulting bound with the strong convexity of the objective. In particular, we set the step size η_t to take the form $\eta_t = \frac{\rho}{t+\gamma}$ for some $\rho > \frac{1}{\mu}$ and $\gamma \ge \max(4L\rho, H)$, for which $\eta_t \le \frac{1}{4L}$ and $\eta_t \le 2\eta_{t+H}$, implying that Lemmas A.2–A.3 hold.

Under such settings, we show that there exists a finite ν such that $\delta_t \le \frac{\nu}{t+\gamma}$ for all integer $l \ge 0$. We prove this by induction, noting that setting $\nu \ge \gamma\delta_0$ guarantees that it holds for $t = 0$. We next show that if $\delta_t \le \frac{\nu}{t+\gamma}$, then $\delta_{t+1} \le \frac{\nu}{t+1+\gamma}$. It follows from (B.1) that $\delta_{t+1} \le \frac{1}{t+\gamma}((1-\frac{\rho}{t+\gamma}\mu)\nu + \frac{\rho^2}{t+\gamma}C)$. Consequently, $\delta_{t+1} \le \frac{\nu}{t+1+\gamma}$ holds when

$$\left(1 - \frac{\rho}{t+\gamma}\mu\right)\nu + \frac{\rho^2}{t+\gamma}C \le \frac{t+\gamma}{t+1+\gamma}\nu. \quad \text{(B.2)}$$

By setting $\nu \ge \frac{\rho^2\,C}{\rho\mu-1}$, the left hand side of (B.2) satisfies $(1-\frac{\rho}{t+\gamma}\mu)\nu + \frac{\rho^2}{t+\gamma}C \le \frac{t-1+\gamma}{t+\gamma}\nu$ since $\nu \ge \frac{\rho^2\,C}{\rho\mu-1}$. As this bound is not larger than that of (B.2), it follows that (B.2) holds for the current setting, proving that $\delta_{t+1} \le \frac{\nu}{t+1+\gamma}$. Finally, the smoothness of the objective implies that

$$\mathbb{E}\{F(\theta_t)\} - F(\theta^*) \le \frac{L}{2}\delta_t \le \frac{L\nu}{2(t+\gamma)}, \quad \text{(B.3)}$$

which, in light of the above setting, holds for $\nu = \max(\frac{\rho^2\,C}{\rho\mu-1}, \gamma\delta_0)$, $\gamma \ge \max(H, 4\rho L)$, and $\rho > 0$. In particular, setting $\rho = \frac{2}{\mu}$ results in $\gamma = \max(H, 8L/\mu)$, $\nu = \max(\frac{4\,C}{\mu^2}, \gamma\delta_0)$ and $\mathbb{E}[F(\theta_t)] - F(\theta^*) \le \frac{2L\max(4\,C, \mu^2\gamma\delta_0)}{\mu^2(t+\gamma)}$, thus concluding the proof of Theorem 2. □

C. Proof of Theorem 3

First, as done in Appendix A, we the virtual sequence $\{\bar{\theta}_t\}$, which here is given by

$$\bar{\theta}_{t+1} = \begin{cases} \frac{1}{N}\sum_{n=1}^{N}\theta_t^n, & t+1 \notin \mathcal{H}, \\ \frac{1}{K}\sum\limits_{n\in\mathcal{K}_t}\theta_t^n + \frac{N}{Kh_{min}}\boldsymbol{w}_t, & t+1 \in \mathcal{H}. \end{cases} \quad \text{(C.1)}$$

Let $\bar{\boldsymbol{v}}_t \triangleq \frac{1}{N}\sum_{n=1}^{N}\theta_t^n + \frac{N}{Kh_{min}}\boldsymbol{w}_t\mathbb{1}_{t\in\mathcal{H}}$ be the virtual sequence of the averaged model over all users. Therefore $\bar{\boldsymbol{v}}_t = \bar{\theta}_t$ when $t \notin \mathcal{H}$. Under this notation, Theorem 2 characterizes the convergence of $\mathbb{E}[F(\bar{\boldsymbol{v}}_t)] - F(\theta^\star)$. We use the following lemmas, proved in [12, Appendix B.4].

Lemma C.1: Under assumption *AS4* $\bar{\theta}_t$ is an unbiased estimation of $\bar{\boldsymbol{v}}_t$, i.e. $\mathbb{E}_{\mathcal{K}_t}[\bar{\theta}_t] = \bar{\boldsymbol{v}}_t$.

Lemma C.2: The expected difference between $\bar{\theta}_t$ and $\bar{\boldsymbol{v}}_t$ is bounded by: $\mathbb{E}_{\mathcal{K}_t}[\|\bar{\boldsymbol{v}}_t - \bar{\theta}_t\|^2] \le \frac{4(N-K)}{(N-1)K}\eta_t^2H^2\,G^2$.

We next use these lemmas to prove the theorem, as $\|\bar{\theta}_{t+1} - \theta^\star\|^2 = A_1 + A_2 + A_3$ with $A_1 \triangleq \|\bar{\theta}_{t+1} - \bar{\boldsymbol{v}}_{t+1}\|^2$, $A_2 \triangleq \|\bar{\boldsymbol{v}}_{t+1} - \theta^\star\|^2$, and $A_3 \triangleq 2\langle\bar{\theta}_{t+1} - \bar{\boldsymbol{v}}_{t+1}, \bar{\boldsymbol{v}}_{t+1} - \theta^\star\rangle$. The term $E_{\mathcal{K}_t}[A_3] = 0$ since $\bar{\theta}_t$ is unbiased by Lemma C.1. Further, using Lemma C.2, Theorem 2, and the equivalent global model in (18) to bound A_1 and A_2 respectively:

$$\mathbb{E}[\|\bar{\theta}_{t+1} - \theta^\star\|^2] \le (1-\eta_t\mu)\mathbb{E}[\|\bar{\theta}_t - \theta^\star\|^2] + \eta_t^2(\tilde{C} + D) \quad \text{(C.2)}$$

where $D = \frac{4(N-K)}{K(N-1)}H^2\,G^2$. Notice the difference between equations (B.1) and (C.2) is in the additional constant D, and the scaling of the noise-to-signal ratio in $\tilde{C}$ compared to C in Theorem 2. The same arguments used in proving Theorem 2 can now be applied to (C.2) to prove the theorem. □

Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on July 29,2021 at 09:55:20 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] T. Sery, N. Shlezinger, K. Cohen, and Y. C. Eldar, "COTAF: Convergent over-the-air federated learning," in *Proc. IEEE Global Commun. Conf.*, 2020, pp. 1–6.

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–453, 2015.

[3] J. Chen and X. Ran, "Deep learning with edge computing: A review," *Proc. IEEE*, vol. 107, no. 8, pp. 1655–1674, 2019.

[4] H. B. McMahan, E. Moore, D. Ramage, and S. Hampson, "Communication-efficient learning of deep networks from decentral-ized data," 2016, *arXiv:1602.05629*.

[5] P. Kairouz *et al.*, "Advances and open problems in federated learning," 2019, *arXiv:1912.04977*.

[6] T. Gafni, N. Shlezinger, K. Cohen, Y. C. Eldar, and H. V. Poor, "Federated learning: A signal processing perspective," 2021, *arXiv:2103.17150*.

[7] V. Smith, C.-K. Chiang, M. Sanjabi, and A. S. Talwalkar, "Federated multi-task learning," in *Proc. NeurIPS*, 2017, pp. 4424–4434.

[8] N. Shlezinger, S. Rini, and Y. C. Eldar, "The communication-aware clustered federated learning problem," in *Proc. IEEE Int. Symp. Inf. Theory*, 2020, pp. 2610–2615.

[9] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 50–60, May 2020.

[10] speedtest.net, "Speedtest united states market report," 2019. [Online]. Available: http://www.speedtest.net/reports/united-states/

[11] 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, Jan. 2021.

[12] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-iid data," 2019, *arXiv:1907.02189*.

[13] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "QSGD: Communication-efficient SGD via gradient quantization and encoding," in *Proc. NeurIPS*, 2017, pp. 1709–1720.

[14] N. Shlezinger, M. Chen, Y. C. Eldar, H. V. Poor, and S. Cui, "UVeQFed: Universal vector quantization for federated learning," *IEEE Trans. Signal Process.*, vol. 69, pp. 500–514, 2021.

[15] A. F. Aji and K. Heafield, "Sparse communication for distributed gradient descent," 2017, *arXiv:1704.05021*.

[16] D. Alistarh, T. Hoefler, M. Johansson, N. Konstantinov, S. Khirirat, and C. Renggli, "The convergence of sparsified gradient methods," in *Proc. NeurIPS*, 2018, pp. 5973–5983.

[17] A. Goldsmith, *Wireless Communications*. Cambridge Univ. Press, 2005.

[18] 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.

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

[20] T. Sery and K. Cohen, "On analog gradient descent learning over multiple access fading channels," *IEEE Trans. Signal Process.*, 2020, vol. 68, pp. 2897–2911, 2020.

[21] 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, Mar. 2020.

[22] H. Guo, A. Liu, and V. K. Lau, "Analog gradient aggregation for federated learning over wireless networks: Customized design and convergence analysis," *IEEE Internet Things J.*, vol. 8, no. 1, pp. 197–210, Jan. 2021.

[23] O. Abari, H. Rahul, and D. Katabi, "Over-the-air function computation in sensor networks," 2016, *arXiv:1612.02307*.

[24] K. Liu and A. Sayeed, "Type-based decentralized detection in wireless sensor networks," *IEEE Trans. Signal Process.*, vol. 55, no. 5, pp. 1899–1910, May 2007.

[25] K. Cohen and A. Leshem, "Performance analysis of likelihood-based multiple access for detection over fading channels," *IEEE Trans. Inf. Theory*, vol. 59, no. 4, pp. 2471–2481, Apr. 2013.

[26] M. Seif, R. Tandon, and M. Li, "Wireless federated learning with local differential privacy," in *Proc. IEEE Int. Symp. Inf Theory (ISIT)*, 2020.

[27] Z. Yang, M. Chen, W. Saad, C. S. Hong, and M. Shikh-Bahaei, "Energy efficient federated learning over wireless communication networks," 2019, *arXiv:1911.02417*.

[28] N. H. Tran, W. Bao, A. Zomaya, M. N. H. Nguyen, and C. S. Hong, "Federated learning over wireless networks: Optimization model design and analysis," in *Proc. IEEE Conf. Comput. Commun.*, 2019, pp. 1387–1395.

[29] W. Ni, Y. Liu, Z. Yang, H. Tian, and X. Shen, "Federated learning in multi-RIS aided systems," 2020, *arXiv:2010.13333*.

[30] N. Cesa-Bianchi, S. Shalev-Shwartz, and O. Shamir, "Online learning of noisy data," *IEEE Trans. Inf. Theory*, vol. 57, no. 12, pp. 7907–7931, Dec. 2011.

[31] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-iid data," 2018, *arXiv:1806.00582*.

[32] S. U. Stich, "Local SGD converges fast and communicates little," 2018, *arXiv:1805.09767*.

[33] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere, "The million song dataset," in *Proc. ISMIR*, 2011.

[34] G. An, "The effects of adding noise during backpropagation training on a generalization performance," *Neural Comput.*, vol. 8, no. 3, pp. 643–674, 1996.

[35] W.-T. Chang and R. Tandon, "Communication efficient federated learning over multiple access channels," 2020, *arXiv:2001.08737*.

[36] S. U. Stich, J.-B. Cordonnier, and M. Jaggi, "Sparsified SGD with memory," in *Proc. NeurIPS*, 2018, pp. 4447–4458.

[37] A. Neelakantan *et al.*, "Adding gradient noise improves learning for very deep networks," 2015, *arXiv:1511.06807*.

[38] H. Chen, S. Lundberg, and S.-I. Lee, "Checkpoint ensembles: Ensemble methods from a single training process," 2017, *arXiv:1710.03282*.

[39] K. Cohen and A. Leshem, "A time-varying opportunistic approach to lifetime maximization of wireless sensor networks," *IEEE Trans. Signal Process.*, vol. 58, no. 10, pp. 5307–5319, Oct. 2010.

[40] M. Chen, N. Shlezinger, H. V. Poor, Y. C. Eldar, and S. Cui, "Communication efficient federated learning," *Proc. Nat. Acad. Sci.*, vol. 118, no. 17, pp. 1–8, 2021.

[41] *MathWorks Deep Learning Toolbox Team*, "Deep learning tutorial series," *MATLAB Central File Exchange*, 2020. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/62990-deep-learning-tutorial-series

[42] L. Holmstrom and P. Koistinen, "Using additive noise in back-propagation training," *IEEE Trans. Neural Netw.*, vol. 3, no. 1, pp. 24–38, Jan. 1992.

Tomer Sery received the B.Sc. and M.Sc. degrees in electrical and computer engineering from Ben-Gurion University, Beersheba, Israel, in 2019, and 2020 respectively. He is a parallel processing architect in GSI technology, Tel Aviv, Israel.

Nir Shlezinger (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in electrical and computer engineering from Ben-Gurion University, Beer-Sheva, Israel, in 2011, 2013, and 2017, respectively. From 2017 to 2019, he was a Postdoctoral Researcher with Technion, and from 2019 to 2020, he was a Postdoctoral Researcher with the Weizmann Institute of Science, Rehovot, Israel, where he was awarded the FGS prize for outstanding research achievements. He is currently an Assistant Professor with the School of Electrical and Computer Engineering, Ben-Gurion University, Israel. His research interests include communications, information theory, signal processing, and machine learning.

Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on July 29,2021 at 09:55:20 UTC from IEEE Xplore. Restrictions apply.

Kobi Cohen (Senior Member, IEEE) received the B.Sc. and Ph.D. degrees in electrical engineering from Bar-Ilan University, Ramat Gan, Israel, in 2007 and 2013, respectively. In October 2015, he joined the Department of Electrical and Computer Engineering, Ben-Gurion University of the Negev (BGU), Beer-Sheva, Israel, where he is currently a Senior Lecturer (tenured Assistant Professor). He is also a Member of the Cyber Security Research Center, and the Data Science Research Center, BGU. Before joining BGU, he was with the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Champaign, IL, USA, from August 2014 to July 2015, and the Department of Electrical and Computer Engineering, University of California, Davis, CA, USA, as a Postdoctoral Research Associate, from November 2012 to July 2014. His main research interests include decision theory, stochastic optimization, and statistical inference and learning, with applications and analysis in communication networks, cyber systems, and large-scale systems. He was the recipient of various awards, including the Best Paper Award in the International Symposium on Modeling and Optimization in Mobile, Ad hoc and Wireless Networks (WiOpt) 2015, the Feder Family Award (second prize), granted by the Advanced Communication Center, Tel Aviv University, Tel Aviv, Israel, in 2011, and President Fellowship from 2008 to 2012 and top honor list's prizes from Bar-Ilan University, in 2006, 2010, and 2011, respectively.

Yonina C. Eldar (Fellow, IEEE) received the B.Sc. degree in physics in 1995 and the B.Sc. degree in electrical engineering in 1996 from Tel-Aviv University, Tel-Aviv, Israel, and the Ph.D. degree in electrical engineering and computer science in 2002 from the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA.

She is currently a Professor with the Department of Mathematics and Computer Science, Weizmann Institute of Science, Rehovot, Israel. She was previously a Professor with the Department of Electrical Engineering, Technion, where she held the Edwards Chair of engineering. She is also a Visiting Professor with MIT, a Visiting Scientist with Broad Institute, and an Adjunct Professor with Duke University, Durham, NC, USA, and was a Visiting Professor with Stanford University, Stanford, CA, USA. She is the author of the book *Sampling Theory: Beyond Bandlimited Systems* and also coauthor of four other books published by Cambridge University Press. Her research interests include statistical signal processing, sampling theory and compressed sensing, learning and optimization methods, and their applications to biology, medical imaging and optics.

He was the recipient of many awards for excellence in research and teaching, including the IEEE Signal Processing Society Technical Achievement Award (2013), the IEEE/AESS Fred Nathanson Memorial Radar Award (2014), the IEEE Kiyo Tomiyasu Award (2016), the Michael Bruno Memorial Award from the Rothschild Foundation, the Weizmann Prize for Exact Sciences, the Wolf Foundation Krill Prize for Excellence in Scientific Research, the Henry Taub Prize for Excellence in Research (twice), the Hershel Rich Innovation Award (three times), the Award for Women with Distinguished Contributions, the Andre and Bella Meyer Lectureship, the Career Development Chair at the Technion, the Muriel David Jacknow Award for Excellence in Teaching, and the Technion's Award for Excellence in Teaching (two times), various Best Paper awards and best demo awards together with her research students and colleagues, including the SIAM outstanding Paper Prize, the UFFC Outstanding Paper Award, the Signal Processing Society Best Paper Award and the IET Circuits, Devices and Systems Premium Award, was selected as one of the 50 most influential women in Israel and in Asia, and is a highly cited Researcher.

She was a Member of the Young Israel Academy of Science and Humanities and the Israel Committee for Higher Education. She is the Editor-in-Chief of Foundations and Trends in Signal Processing, a Member of the IEEE Sensor Array and Multichannel Technical Committee and is on various other IEEE committees. In the past, she was a Signal Processing Society Distinguished Lecturer, Member of the IEEE Signal Processing Theory and Methods and Bio Imaging Signal Processing technical committees, and was an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the *EURASIP Journal of Signal Processing*, the *SIAM Journal on Matrix Analysis and Applications*, and the *SIAM Journal on Imaging Sciences*. She was the Co-Chair and Technical Co-Chair of various international conferences and workshops. She was a Horev Fellow of the Leaders in Science and Technology program at the Technion and an Alon Fellow.

Authorized licensed use limited to: Weizmann Institute of Science. Downloaded on July 29,2021 at 09:55:20 UTC from IEEE Xplore. Restrictions apply.