COTAF: Convergent Over-the-Air Federated Learning

Tomer Sery, Nir Shlezinger, Kobi Cohen, and Yonina C. Eldar

GLOBECOM 2020 - 2020 IEEE Global Communications Conference | 978-1-7281-8298-8/20/$31.00 ©2020 IEEE | DOI: 10.1109/GLOBECOM42002.2020.9322580

***Abstract*—Federated learning (FL) is a framework for distributed learning of centralized models. In FL, a set of edge devices train a model using their local data, while repeatedly exchanging their trained model with a central server, allowing to tune a global model without having the users share their possibly private data. A major challenge in FL is to reduce the bandwidth and energy consumption due to the repeated transmissions of large volumes of data by a large number of users over the wireless channel. Recently, over-the-air (OTA) FL has been suggested to achieve this goal. In this setting, all users transmit their data signal simultaneously over a Multiple Access Channel (MAC), and the computation is done over the wireless channel. In this paper, we develop a novel convergent OTA FL (COTAF) algorithm, which induces precoding and scaling upon transmissions to gradually mitigate the effect of the noisy channel, thus facilitating FL convergence. We analyze the convergence of COTAF to the loss minimizing model theoretically, showing its ability to achieve a convergence rate similar to that achievable over error-free channels. Our simulations demonstrate the improved convergence of COTAF for training using non-synthetic datasets.**

***Index terms*— Federated learning, wireless communications, multiple access channels (MAC).**

I. Introduction

Recent years have witnessed an unprecedented success of machine learning methods in a broad range of applications [1]. These systems utilize highly parameterized models, such as deep neural networks, trained using a massive amount of labeled data samples. Since data samples are available at remote users, e.g., smartphones and other edge devices, the common strategy is to gather these samples at a computationally powerful server, which uses them to train its model [2].

Often, data samples, such as images and text messages, contain private information, and thus the user may not be willing to share them with the server. To allow centralized training without data sharing, federated learning (FL) was proposed in [3] as a method combining distributed training with central aggregation, and has become the focus of growing research attention over the last few years [4]. FL exploits the increased computational capabilities of modern edge devices to train the parametric model on the users side, while having the server periodically collect these local models into a global one and broadcast it back to the users.

The need to repeatedly convey a massive amount of model parameters between the server and a large number of users induces a significant load on the wireless channel over which the users communicate with the server [5]. This is particularly relevant in uplink communications, which are commonly more limited as compared to its downlink counterpart [6]. A commonly adopted 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 [7], [8], or by compressing the conveyed model parameters via quantization [9]–[11] or sparsification [12], [13]. All these methods model 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 [14], 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 each user utilize channel codes to overcome the noise. This, however, results in each user being assigned a dedicated bandwidth which decreases with the number of participating users, increasing in the energy consumption requried to meet a desirable communication rate, and decreasing in the overall throughput and the training speed.

An alternative approach is to allow the users to simultaneously utilize the complete temporal and spectral resources of the uplink multiple access channel (MAC) in a non-orthogonal manner. In this method, referred to as over-the-air (OTA) FL [15]–[17], the users transmit their model updates via analog signalling, i.e., without converting to discrete coded symbols which should be decoded at the user side, and exploit the inherent aggregation carried out by the shared channel as a form of OTA computation [18]. Early studies in the sensor network literature considered model-dependent inference over mulitple access channels (MACs) [19]–[27]. Although the theoretical performance analysis has been established rigorously under a wide class of problem settings, all these studies assume that the observation distributions are known to the nodes or to the network edge. Therefore, developing efficient inference algorithms over MAC in the online learning context, where the observation distributions are unknown, becomes extremely important to expand their applicability to real-world problems, and have attracted a growing interest recently. In particular, the works [15], [16] studied OTA FL in which the model updates are sparse with an identical sparsity pattern, while [17] proposed an algorithm for computing gradient descent using OTA computations. The main advantage in this approach is that it allows the users to transmit at increased throughput, being allowed to exploit the complete available bandwidth regardless of the number of participating users. However, a major drawback of such uncoded analog signalling is that the noise induced by the channel is not handled by channel coding and thus affects the training procedure, as the convergence of learning algorithms such as stochastic

This project has received funding from the Benoziyo Endowment Fund for the Advancement of Science, the Estate of Olga Klein – Astrachan, the European Union's Horizon 2020 research and innovation program under grant No. 646804-ERC-COG-BNYQ, the Israel Science Foundation under grants No. 0100101 and 2640/20 and the U.S.-Israel Binational Science Foundation (BSF) under grant 2017723. T. Sery and K. Cohen are with the School of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel (e-mail:seryt@post.bgu.ac.il; yakovsec@bgu.ac.il). N. Shlezinger and Y. C. Eldar are with the Math and CS Faculty, Weizmann Institute of Science, Rehovot, Israel (e-mail:nirshlezinger1@gmail.com; yonina.eldar@weizmann.ac.il).

978-1-7281-8298-8/20/$31.00 ©2020 IEEE

gradient descent (SGD) is known to be sensitive to noisy observations [28]. This motivates the design and analysis of an FL scheme for wireless channels which exploits the high throughput of OTA computations while maintaining the convergence properties of conventional FL methods. Other recent studies considered FL digital gradient transmissions [29], and privacy over MACs [30].

Next, we summarize our main contributions: We propose a convergent OTA FL (COTAF) algorithm, which facilitates high throughput FL over wireless channels, while preserving the convergence properties of the common local SGD method for distributed learning. COTAF overcomes the need to divide the channel resources among the users by allowing the users to simultaneously share the uplink channel while aggregating the global model via OTA computations. To that aim, we introduce a time-varying precoding and scaling scheme which allows COTAF to gradually mitigate the effect of channel noise, thus achieving convergence properties similar to that of local SGD carried out over error-free independent channels.

We theoretically analyze the convergence of models trained by COTAF to the minimal achievable loss function, when the objective is a strongly convex and smooth function with bounded gradients, as commonly utilized in FL convergence studies [8], [31]–[33]. We rigorously prove that when using COTAF, the usage of analog transmissions over shared noisy channels does not affect the asymptotic convergence rate of local SGD compared to communicating over error-free separate channels, while allowing the users to communicate at high throughput by avoiding the need to divide the channel resources. Our numerical results, which consider distributed training with the Million Song Dataset [34], demonstrate that COTAF achieves accuracy within a minor gap from that of unconstrained local SGD, while notably outperforming OTA FL strategies without time-varying precoding designed to facilitate convergence.

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.

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

We start by describing the local SGD algorithm, which is the common training method used in FL in Subsection II-A. Then, we present the uplink channel model over which the users communicate with the server in Subsection II-B.

A. Local SGD

Consider a central server which trains a model consisting of d parameters, represented by the vector $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$, using data available at N users, indexed by the set $\mathcal{N} = \{1, 2, ..., N\}$. Each user of index $n \in \mathcal{N}$ has access to a data set of D_n samples, denoted by $\{\boldsymbol{s}_i^n\}_{i=1}^{D_n}$. All $\boldsymbol{s}_i^n$ are assumed to be i.i.d samples from some distribution $\mathcal{X}$, i.e. $\boldsymbol{s}_i^n \overset{\text{i.i.d}}{\sim} \mathcal{X}$. Let $f(\boldsymbol{\theta}, \boldsymbol{s})$ be the loss function of model $\boldsymbol{\theta}$ evaluated at sample $\boldsymbol{s}$. The global loss is given by

$$F(\boldsymbol{\theta}) \triangleq \frac{1}{N}\sum_{n=1}^{N}\frac{1}{D_n}\sum_{i=1}^{D_n} f_{i^n}(\boldsymbol{\theta}), \tag{1}$$

where $f_{i^n}(\boldsymbol{\theta}) \triangleq f(\boldsymbol{\theta}, \boldsymbol{s}_i^n)$. Consequently, the objective of the training process is to approach the following minimizer:

$$\boldsymbol{\theta}^{\star} = \underset{\boldsymbol{\theta}\in\Theta}{\arg\min}\; F(\boldsymbol{\theta}). \tag{2}$$

Local SGD aims at recovering (2) without having the users share their local data by carrying out multiple training rounds, each consisting of the following three phases:
a) The server shares its current model at time instance t, denoted $\boldsymbol{\theta}_t$, with the users.
b) Each user sets its local model $\boldsymbol{\theta}_t^n$ to $\boldsymbol{\theta}_t$, and trains it using its local data set using H SGD steps, namely, via

$$\boldsymbol{\theta}_{t+1}^n = \boldsymbol{\theta}_t^n - \eta_t \nabla f_{i_t^n}(\boldsymbol{\theta}_t^n), \tag{3}$$

where $f_{i_t^n}(\cdot)$ 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.
c) 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] $\boldsymbol{\theta}_{t+H} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\theta}_{t+H}^n$, and sends the new model to the users for another round. This iterative algorithm is known to converge to (2) as the number of rounds grows, for various families of loss measures [8], [31], [33].

Each round of FL consists of two communication phases: downlink transmission of the global model $\boldsymbol{\theta}_t$ 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. This involves the repetitive communication of a large amount of parameters over wireless channels. This increased communication overhead is considered one of the main challenges of FL [4], [5], particularly the uplink communication, which is known to be more rate limited in cellular systems [6]. 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 $\boldsymbol{\theta}_t$ to $\boldsymbol{\theta}^{\star}$ is preserved [9]–[11], [13]. 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 limits severely the volume of data which can be conveyed as compared to utilizing the full bandwidth. To overcome this issue, we next describe the communications model over a MAC used in the suggested COTAF algorithm.

B. Communication Channel Model

FL is commonly carried out over wireless channels. We consider here FL setups in which the N users communicate with the server using the same wireless network, either

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

directly or via some wireless access point, focusing on the uplink transmissions. 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 $\boldsymbol{x}_t^n$ can be written as

$$\boldsymbol{y}_t = \sum_{n=1}^{N} \boldsymbol{x}_t^n + \tilde{\boldsymbol{w}}_t, \tag{4}$$

where $\tilde{\boldsymbol{w}}_t \sim \mathcal{N}(0, \sigma_w^2 \boldsymbol{I}_d)$ is $d \times 1$ vector of additive noise. The channel input is subject to an average power constraint

$$\mathbb{E}\left[\|\boldsymbol{x}_t^n\|^2\right] \leq P, \tag{5}$$

where $P > 0$ represents the available transmission power. An illustration of a single round of local SGD carried out over such a wireless MAC is depicted in Fig. 1. The channel in (4) represents an additive noise MAC model, 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 seconds.

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 K distinct bands, or via time division multiplexing (TDM), in which the temporal block is divided into K 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

$$\boldsymbol{y}_t^n = \boldsymbol{x}_t^n + \tilde{\boldsymbol{w}}_t^n, \quad n \in \mathcal{N}. \tag{6}$$

While orthogonalization of the channels in (6) facilitates recovery of each $\boldsymbol{x}_t^n$ individually, the fact that each user has access only to $1/K$ of the channel resources implies that its throughput, i.e., the volume of data which can be conveyed reliably, is reduced accordingly [14, Ch. 4].

Our goal is to design a communication strategy for FL over wireless channels of the form (4), namely, to determine a mapping from $\boldsymbol{\theta}_t^n$ into $\boldsymbol{x}_t^n$ at each user as well as a transformation of $\boldsymbol{y}_t$ into $\boldsymbol{\theta}_t$ on the server side. The fact that in FL, the task of the server on every communication round is not to recover each model update individually, but to aggregate them into a global model $\boldsymbol{\theta}_t$, motivates having each of the users exploiting the complete spectral and temporal resources by avoiding conventional orthogonality-based strategies and utilizing the wireless MAC (4) on uplink transmissions. The inherent aggregation carried out by the MAC can in fact facilitate FL at high communication rate via OTA computations [18], as was also proposed in the context of distributed learning in [15]–[17]. However, the fact that the channel outputs are corrupted by additive noise is known to degrade the ability of SGD-based FL algorithms to converge to the desired $\boldsymbol{\theta}^\star$ [28]. This motivates the COTAF protocol as detailed in the following section. COTAF tackles the limited convergence of noisy SGD by properly precoding the model updates into the channel inputs $\{\boldsymbol{x}_t^n\}$. This allows to achieve FL performance which approaches that of FL over noise-free orthogonal channels, while utilizing the complete spectral and temporal resources of the wireless channel.

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

The COTAF algorithm is designed to facilitate OTA FL at high rates by letting each user exploit the full spectral and temporal resources of the wireless MAC. We start by describing the COTAF transmission and aggregation protocol in Subsection III-A. Then, we provide theoretical convergence analysis in Subsection III-B, rigorously proving its ability to converge to the loss-minimizing network weights. Finally, we discuss the pros and cons of COTAF in Subsection III-C.

A. Precoding and Reception Algorithm

In COTAF, all users transmit in their channel output $\{\boldsymbol{x}_t^n\}$ a function of their locally trained models to the server simultaneously. As in [15], [17], [20], [21], 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 $\boldsymbol{y}_t$ as detailed in the sequel, and feedbacks the updated model to the users as in conventional FL.

In particular, COTAF implements the local SGD algorithm detailed in Subsection II-A while communicating over an uplink wireless MAC as illustrated in Fig. 1. In order to convey the local trained model after H local SGD steps, i.e., at a communication round occurring at time instance t, 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

$$\boldsymbol{x}_t^n = \sqrt{\alpha_t}\left(\boldsymbol{\theta}_t^n - \boldsymbol{\theta}_{t-H}^n\right), \tag{7}$$

where α_t is a precoding factor set to satisfy the power constraint (5), and is given by:

$$\alpha_t \triangleq \frac{P}{\max_n \mathbb{E}\left[||\boldsymbol{\theta}_t^n - \boldsymbol{\theta}_{t-H}^n||^2\right]}. \tag{8}$$

We discuss the computation of α_t in Subsection III-C.

The channel output (4) is thus given by

$$\boldsymbol{y}_t = \sum_{n=1}^{N} \sqrt{\alpha_t}\left(\boldsymbol{\theta}_t^n - \boldsymbol{\theta}_{t-H}^n\right) + \tilde{\boldsymbol{w}}_t. \tag{9}$$

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

$$\theta_t = \frac{\boldsymbol{y}_t}{N\sqrt{\alpha_t}} + \theta_{t-H}, \tag{10}$$

for $t = 0, 1, 2, \cdots$, where $\boldsymbol{\theta}_0$ is the initial parameter estimate. The global update rule (10) can be equivalently written as

$$\boldsymbol{\theta}_t = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{\theta}_t^n + \boldsymbol{w}_t, \tag{11}$$

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 as Algorithm 1. By letting $\mathcal{H}$ be the set of time instances in which transmissions occur, i.e., the integer

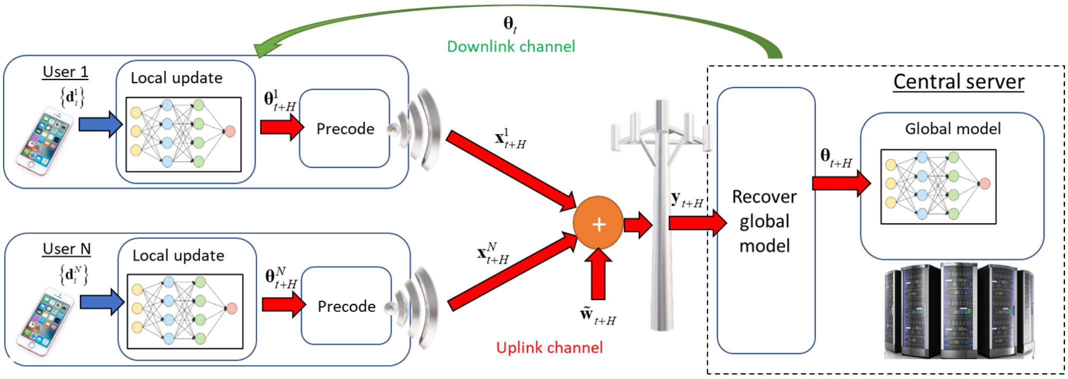

Fig. 1. FL over wireless MAC illustration.

Algorithm 1: COTAF algorithm

```
Init: Fix an initial $\theta_0^n = \theta_0$ for each user $n \in \mathcal{N}$.
for $t = 1, 2, \dots, RH$ do
    Each user $n \in \mathcal{N}$ locally trains $\theta_t^n$ via (3);
    if $t \in \mathcal{H}$ then
        Each user $n \in \mathcal{N}$ transmits $x_t^n$ precoded via
          (7) over the MAC (4);
        The server recovers $\theta_t$ from $y_t$ via (10);
        The server broadcasts $\theta_t$ to the users;
        Each user $n \in \mathcal{N}$ sets $\theta_t^n = \theta_t$;
    end
end
Output: Global model $\theta_{RH}$
```

multiples of H, the local model available at the nth user at time t can be written as:

$$\theta_{t+1}^n = \begin{cases} \theta_t^n - \eta_t \nabla f_{i_t^n}(\theta_t^n), & t+1 \notin \mathcal{H}, \\ \frac{1}{N} \sum\limits_{n=1}^{N} \left(\theta_t^n - \eta_t \nabla f_{i_t^n}(\theta_t^n)\right) + \boldsymbol{w}_t, & t+1 \in \mathcal{H}. \end{cases} \quad (12)$$

B. Convergence Analysis

In this section we theoretically characterize the convergence of COTAF to the optimal model parameters $\theta^\star$, i.e., the vector θ which minimizes the global loss function. Our analysis is carried out under the following assumptions:

- *A1* The objective function $F(\cdot)$ is L-smooth, namely, for all $\boldsymbol{v}_1, \boldsymbol{v}_2$ it holds that $F(\boldsymbol{v}_1) - F(\boldsymbol{v}_2) \leq (\boldsymbol{v}_1 - \boldsymbol{v}_2)^T \nabla F(\boldsymbol{v}_2) + \frac{1}{2} L \|\boldsymbol{v}_1 - \boldsymbol{v}_2\|^2$.
- *A2* The objective function $F(\cdot)$ is μ-strongly convex, namely, for all $\boldsymbol{v}_1, \boldsymbol{v}_2$ it holds that $F(\boldsymbol{v}_1) - F(\boldsymbol{v}_2) \geq (\boldsymbol{v}_1 - \boldsymbol{v}_2)^T \nabla F(\boldsymbol{v}_2) + \frac{1}{2} \mu \|\boldsymbol{v}_1 - \boldsymbol{v}_2\|^2$.
- *A3* The stochastic gradients $\nabla f_{i_t^n}(\theta)$ satisfy $\mathbb{E}[\|\nabla f_{i_t^n}(\theta)\|^2] \leq G^2$ and $\mathbb{E}[\|\nabla f_{i_t^n}(\theta) - \nabla F(\theta)\|^2] \leq M^2$ for some fixed $G^2 > 0$ and $M^2 > 0$, for any $\theta \in \Theta$.

Assumptions *A1*-*A3* are commonly used when studying the convergence of FL schemes, see, e.g., [8], [31]. In particular, *A1*-*A2* hold for a broad range of objective functions used in FL systems, including ℓ_2-norm regularized linear regression and logistic regression [8], while *A3* represents having bounded second-order statistical moments of the stochastic gradients [31]. It is also emphasized that these assumptions are required to maintain an analytically tractable convergence analysis, and that COTAF can be applied for arbitrary learning tasks for which *A1*-*A3* do not necessarily hold.

The error of SGD-type algorithms is commonly defined as the expected loss in the objective value at iteration $T = RH$ between a weighted average of learned model parameters, denoted $\hat{\theta}_T$, with respect to $F^\star$. Such bounds are commonly used in the FL literature to characterize convergence [31]–[33], and can be commonly used to obtain convergence bounds for the instantaneous weights, i.e., θ_T instead of $\hat{\theta}_T$, see, e.g., [8] as well as the discussion regarding FL convergence results in [4, Sec. 3.2]. We establish a finite-sample bound on the error in the following theorem:

Theorem 1. *Let $\{\theta_t^n\}_{n=1}^N$ be the model parameters generated by COTAF according to (3) and (11) over R rounds, i.e., $t \in \{0, 1, \dots T-1\}$ with $T = RH$. Then, when A1-A3 hold and the step sizes are set to $\eta_t = \frac{4}{\mu(a+t)}$ with shift parameter $a > \max\{16\frac{L}{\mu}, H\}$, it holds that:*

$$\mathbb{E}[F(\hat{\theta}_T)] - F^\star \leq \frac{\mu a^3}{2S_T}\|\theta_0 - \theta^\star\|^2 + \frac{4T(T+2a)}{\mu N S_T} M^2 + \frac{256T}{\mu^2 S_T} G^2 H^2 L + \frac{16dH^2G^2\sigma_w^2}{\mu P N^2 S_T} \frac{T(R+1)+2aR}{2}, \quad (13)$$

where $\hat{\theta}_T = \frac{1}{NS_T}\sum_{n=1}^N \sum_{t=0}^T \beta_t \theta_t^n$, for $\beta_t = (a+t)^2$, and $S_T = \sum_{t=0}^T \beta_t \geq \frac{1}{3}T^3$.

Comparing (13) to the corresponding bound for local SGD without communication constraints [31, Thm. 2.2], i.e., over the orthogonal channels as in (6) without noise, we observe that the first three summands of (13) are identical to [31, Eq. (5)]. Consequently, the fact that COTAF communicates over a noisy channel induces an additional term which can be written as the noise-to-signal ratio σ_w^2/P times some factor which, as the number of FL rounds R grows, is dominated by $HT^2/S_T \leq \frac{1}{R}$. This implies that the time varying precoding and aggregation strategy implemented by COTAF results in a gradual decay of the noise effect, which in turn yields the same asymptotic convergence rate to that observed in [31], as stated in the following corollary:

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

Corollary 1 implies that COTAF results in an FL system which achieves the same asymptotic convergence rate as local SGD without communication constraints [31]. 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 we discuss in the following section.

C. Discussion

COTAF is designed to allow FL systems operating over shared wireless channels to exploit the full spectral and temporal resources of the media, by accounting for the task of aggregating the local models into a global one as a form of OTA computation [18]. 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.

A major challenge in implementing SGD as an OTA computation stems from the presence of an additive noise, whose contribution does not decay over time [28]. 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 $\mathcal{O}(\frac{1}{T})$ [31]. 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 while operating over shared noisy wireless channels. Compared to previous strategies for OTA FL, COTAF allows the implementation of local SGD, which is the arguably the most widely used FL scheme, over wireless MACs with proven convergence without having to restrict the model updates to be sparse with an identical sparsity pattern shared among all users [15], [16], or requiring the users to repeatedly compute the gradients over the full data set [17].

Our derivation and analysis of COTAF is carried out for a relatively simplified model for FL and wireless communication, which may not accurately reflect their characteristics in practice. For once, the wireless channel is modeled as an additive noise MAC, while such channels often induce fading in addition to noise. Furthermore, the training data is assumed to be drawn from a single distribution in an i.i.d. manner, while FL systems are typically trained with statistically heterogeneous data sets [4], [35]. We leave the extension and analysis of COTAF for heterogeneous FL systems over fading wireless channels for future research.

While COTAF consists of an addition of simple precoding and scaling stages to local SGD, its implementation involves several challenges: First, by (8), each user has to know $\max_n \mathbb{E}\left[||\boldsymbol{\theta}_t^n - \boldsymbol{\theta}_{t-H}^n||^2\right]$, for each communication round $t \in \mathcal{H}$. When operating with a decaying step size, as is commonly required in FL, and *A3* holds, it can be shown that this term is upper bound by $H^2\eta_{t-H}^2 G^2$, and the upper bound can be used instead in (8), while maintaining the convergence guarantees of Theorem 1. Alternatively, one can numerically tune this value offline, and have the server distribute it to the users over the downlink channel. Additionally, we note that 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, such as 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 a multitude of different locations. We thus conjecture that COTAF can be combined in multi-stage FL, such as clustered FL [35]. We leave this for future study.

IV. Numerical Evaluations

In this section we evaluate COTAF for distributed learning with non-synthetic data in a numerical study. We simulate the distributed learning for prediction of a release year of a song from audio features using a federated learning setting. We use the dataset available by UCI Machine Learning Repository [36], extracted from the Million Song Dataset collaborative project between The Echo Nest and LabROSA [34]. The Million Song Dataset contains songs which are mostly western, commercial tracks ranging from 1922 to 2011. Each song is associated with a released 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. Here, the system task is to train a linear estimator $\boldsymbol{\theta}$ with $d = 90$ entries in a federated manner using data available at $N = 50$ users, where each user has access to $D_n = 9200$ samples. The predictor is trained using the regularized linear least squares loss, given by:

$$f(\boldsymbol{\theta}, \{\boldsymbol{s}_s, s_y\}) = \frac{1}{2}(\boldsymbol{s}_s^T\boldsymbol{\theta} - s_y)^2 + \frac{\lambda}{2}||\boldsymbol{\theta}||^2, \qquad (14)$$

where we used $\lambda = 0.5$. We note that the loss measure (14) is strongly convex and has a Lipschitz gradient, and thus satisfies the conditions of Theorem 1. In every FL round, each user performs $H = 40$ SGD steps (3) where the step size is via Theorem 1, while L and μ are numerically evaluated, before transmitting the model update to the server over a MAC with signal to noise ratio P/σ_w^2 of 14 dB. The precoding coefficient α_t was computed via (8) using numerical averaging.

In Fig. 2 we depict the numerically evaluated gap from the achieved expected objective and the loss-minimizing one, i.e., $\mathbb{E}\big[F(\boldsymbol{\theta}_t)\big] - F^\star$, for COTAF compared 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 (8) and with a constant amplification as in [17]. In particular, we set transmitting $\boldsymbol{x}_t^n = P(\boldsymbol{\theta}_t^n - \boldsymbol{\theta}_{t-H}^n)$. The stochastic expectation is evaluated by averaging over 50 Monte Carlo trials, where in each trial the initial $\boldsymbol{\theta}_0$ is randomized from zero-mean Gaussian distribution with covariance $5\boldsymbol{I}_d$.

Observing Fig. 2, we note that COTAF achieves performance within a minor gap from that of local SGD carried out over ideal orthogonal noiseless channels. This performance of COTAF is achieved without requiring the users to divide the spectral and temporal channel resources among each other, thus allowing high throughput uplink communications. This is due to the precoding scheme of COTAF, which allows to gradually mitigate the effect of the channel noise, while OTA FL without such time-varying precoding results in a dominant error floor due to presence of non-vanishing noise. These results demonstrate the benefits

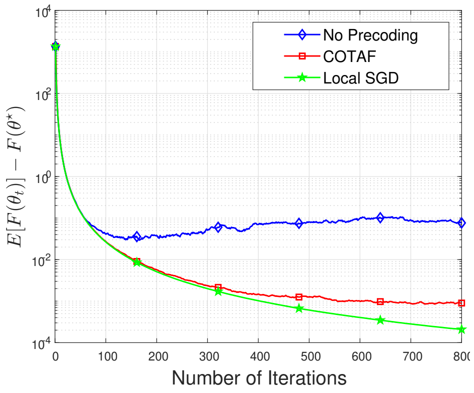

Fig. 2. Simulation results for prediction of a release year of a song. The empirical errors are presented as a function of the number of iterations.

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.

V. Conclusions

In this work we proposed COTAF, which implements FL over wireless MACs without requiring the users to divide the channel resources, while maintaining the convergence properties of local SGD carried out over ideal channels. This is achieved by introducing a time-varying precoding and scaling scheme which facilitates aggregation and gradually mitigates the noise effect. We rigorously proved that models trained using COTAF converge to the loss minimizing model with the same asymptotic convergence rate of local SGD without communication constraints. Our numerical study demonstrates the ability of COTAF to learn accurate models in a federated manner over wireless channels using non-synthetic datasets.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[2] J. Chen and X. Ran, "Deep learning with edge computing: A review," *Proceedings of the IEEE*, 2019.

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

[4] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings *et al.*, "Advances and open problems in federated learning," *arXiv preprint arXiv:1912.04977*, 2019.

[5] 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, 2020.

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

[7] 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," *arXiv preprint arXiv:1909.07972*, 2019.

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

[9] 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.

[10] A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie, and R. Pedarsani, "Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization," *arXiv preprint arXiv:1909.13014*, 2019.

[11] N. Shlezinger, M. Chen, Y. C. Eldar, H. V. Poor, and S. Cui, "Federated learning with quantization constraints," in *Proc. IEEE ICASSP*, 2020.

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

[13] 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.

[14] A. Goldsmith, *Wireless communications*. Cambridge university press, 2005.

[15] 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.

[16] ——, "Federated learning over wireless fading channels," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3546–3557, 2020.

[17] T. Sery and K. Cohen, "On analog gradient descent learning over multiple access fading channels," *IEEE Trans. Signal Process.*, pp. 1–1, 2020.

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

[19] G. Mergen and L. Tong, "Type based estimation over multiaccess channels," *IEEE Trans. Signal Process.*, vol. 54, no. 2, pp. 613–626, 2006.

[20] G. Mergen, V. Naware, and L. Tong, "Asymptotic detection performance of type-based multiple access over multiaccess fading channels," *IEEE Trans. Signal Process.*, vol. 55, no. 3, pp. 1081 – 1092, Mar. 2007.

[21] 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.

[22] S. Marano, V. Matta, T. Lang, and P. Willett, "A likelihood-based multiple access for estimation in sensor networks," *IEEE Trans. Signal Process.*, vol. 55, no. 11, pp. 5155–5166, Nov. 2007.

[23] A. Anandkumar and L. Tong, "Type-based random access for distributed detection over multiaccess fading channels," *IEEE Trans. Signal Process.*, vol. 55, no. 10, pp. 5032–5043, 2007.

[24] 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, 2013.

[25] I. Nevat, G. W. Peters, and I. B. Collings, "Distributed detection in sensor networks over fading channels with multiple antennas at the fusion centre," *IEEE Trans. Signal Process.*, vol. 62, no. 3, pp. 671–683, 2014.

[26] P. Zhang, I. Nevat, G. W. Peters, and L. Clavier, "Event detection in sensor networks with non-linear amplifiers via mixture series expansion," *IEEE Sensors J.*, vol. 16, no. 18, pp. 6939–6946, 2016.

[27] K. Cohen and A. Leshem, "Spectrum and energy efficient multiple access for detection in wireless sensor networks," *IEEE Trans. Signal Process.*, vol. 66, no. 22, pp. 5988–6001, 2018.

[28] 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, 2011.

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

[30] M. Seif, R. Tandon, and M. Li, "Wireless federated learning with local differential privacy," *arXiv preprint arXiv:2002.05151*, 2020.

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

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

[33] B. Woodworth, K. K. Patel, S. U. Stich, Z. Dai, B. Bullins, H. B. McMahan, O. Shamir, and N. Srebro, "Is local SGD better than minibatch SGD?" *arXiv preprint arXiv:2002.07839*, 2020.

[34] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere, "The million song dataset," in *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.

[35] N. Shlezinger, S. Rini, and Y. C. Eldar, "The communication-aware clustered federated learning problem," in *Proc. IEEE ISIT*, 2020.

[36] M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml