MAC Aware Quantization for Distributed Gradient Descent

Wei-Ting Chang Ravi Tandon
Department of Electrical and Computer Engineering
University of Arizona, Tucson, Arizona, USA
Email: {*wchang, tandonr*}@email.arizona.edu

***Abstract*—In this work, we study the problem of federated learning (FL), where distributed users aim to jointly train a machine learning model with the help of a parameter server (PS). In each iteration of FL, users compute local gradients, followed by transmission of the quantized gradients for subsequent aggregation and model updates at PS. One of the challenges of FL is that of communication overhead due to FL's iterative nature and large model sizes. One recent direction to alleviate communication bottleneck in FL is to let users communicate simultaneously over a multiple access channel (MAC), possibly making better use of the communication resources.**

In this paper, we consider the problem of FL learning over a MAC. In particular, we focus on the design of digital gradient transmission schemes over a MAC, where gradients at each user are first quantized, and then transmitted over a MAC to be decoded individually at the PS. When designing digital FL schemes over MACs, there are new opportunities to assign different amount of resources (such as rate or bandwidth) to different users based on a) the informativeness of the gradients at each user, and b) the underlying channel conditions. We propose a stochastic gradient quantization scheme, where the quantization parameters are optimized based on the capacity region of the MAC. We show that such *channel aware quantization* for FL outperforms uniform quantization, particularly when users experience different channel conditions, and when have gradients with varying levels of informativeness.

I. Introduction

Federated Learning (FL) refers to a distributed machine learning (ML) framework that allows distributed machines, or users, to collaboratively train an ML model with the help of a parameter server (PS). Typically, users compute gradients for a global model on their local data, and send gradients to the PS for aggregation and model updates in an iterative fashion. FL is appealing and has gained recent attention due to the fact that it allows natural parallelization, and can be more efficient than centralized approaches in terms of storage. However, communication overhead caused by exchanging gradients remains an issue that needs to be addressed.

Previous works alleviate the communication bottleneck by compressing gradients before transmissions. Two commonly used gradient compression approaches are a) quantization, and b) sparsification. Gradient quantization follows the idea of lossy compression by describing gradients using a small number of bits and these low-precision gradients are transmitted back to the PS. One extreme is to send just 1 bit of information per value [1]. Similar idea was used in signSGD [2] and TernGrad [3], which use 1 and 2 bits to describe each value, respectively. In gradient sparsification, some coordinates of the gradient vector are dropped based on certain criteria [4], [5], which for instance, can depend on the variance and informativeness of the gradients. Other quantization/sparsification techniques include [6]–[10]. However, these stand alone compression techniques are not tuned to the underlying communication channel over which the exchange takes place between the users and the PS, and may not utilize the channel resources to the fullest.

Another line of recent works study FL over wireless channels, and more generally multiple access channels (MACs). The superposition nature of wireless channels allows gradients to be aggregated "over-the-air" and allows for much more efficient training. Several recent works include [11]–[23]. The approaches can be broadly categorized into digital or analog schemes depending on how the gradients are transmitted over the channel. In analog schemes, the local gradients are scaled and directly transmitted over the wireless channel, allowing PS to directly receive a noisy version of the aggregated gradient. In digital schemes, gradients from users are decoded individually, but transmission still occurs over a MAC. Although it has been shown that in terms of bandwidth efficiency, analog schemes can be superior than digital schemes [11], [13], we argue that digital schemes have the following advantages: a) backward compatibility - they can be easily implemented on the existing digital systems, b) they are less prone to slow users, c) they are more reliable due to the fact that various error control codes can be used, and d) digital schemes do not require tight synchronization as required by analog transmission.

Main Contributions: Motivated by the above discussion, we consider FL learning over a MAC and focus on the design of digital gradient transmission schemes, where gradients at each user are first quantized, and then transmitted over a MAC to be decoded individually at the PS. When designing digital FL schemes over MACs, we show that there are new opportunities to assign different amount of resources (such as rate or bandwidth) to different users based on a) the informativeness of the gradients at each user, and b) the underlying channel conditions. We propose a stochastic gradient quantization scheme, where the quantization parameters are optimized based on the capacity region of the MAC. We show that such *channel aware quantization* for FL outperforms channel unaware quantization schemes (such as uniform allocation), particularly when users experience different channel conditions, and when have gradients with varying levels of informativeness.

II. System Model

We consider a distributed machine learning system with a parameter server (PS) and M users, where users are connected to the PS through a Gaussian MAC as shown in Fig. 1. Users want to collaboratively train a machine learning model $\mathbf{w}$ with the help of PS by minimizing an empirical loss function,

$$L(w) = \frac{1}{n_{\text{tot}}} \sum_{m=1}^{M} \sum_{d_n^{(m)} \in \mathcal{D}^{(m)}} \ell(\mathbf{w}, \mathbf{d}_n^{(m)}), \tag{1}$$

where $\mathcal{D}^{(m)}$, $|\mathcal{D}^{(m)}| = n_m$, $m = 1, \dots, M$ denotes the local data set at user m, $n_{\text{tot}} = \sum_{m=1}^{M} n_m$ and $\mathbf{d}_n^{(m)}$ is the n-th data point in $\mathcal{D}^{(m)}$, and $\ell(\cdot)$ is the loss function. The minimization is done by using gradient descent (GD) algorithm. Each user computes the local gradient $\mathbf{g}^{(m)}(\mathbf{w}_t) \in \mathbb{R}^d$ on the local data set $\mathcal{D}^{(m)}$, where $\mathbf{w}_t$ is vector of model parameters at iteration t, and

$$\mathbf{g}^{(m)}(\mathbf{w}_t) = \frac{1}{n_m} \sum_{n=1}^{n_m} \nabla \ell(\mathbf{w}_t, \mathbf{d}_n^{(m)}), \ \mathbf{d}_n^{(m)} \in \mathcal{D}^{(m)}, \ \forall m. \tag{2}$$

At each iteration, each user m sends a function of its computed gradient $\mathbf{x}_t^{(m)} = f_t^{(m)}(\mathbf{g}^{(m)}(\mathbf{w}_t))$ back to the PS through s channel uses of the MAC, where $f_t^{(m)}(\cdot)$ is some pre-processing function the PS assigned to user m at iteration t. We assume that capacity-achieving error control codes are used by users. We note that the capacity region of a Gaussian MAC can be described as follows [24],

$$\sum_{m \in \mathcal{M}} r_m \leq C_{\mathcal{M}}, \ \mathcal{M} \subset [M], \ |\mathcal{M}| = 1, \dots, M, \tag{3}$$

where r_m denotes the transmission rate of user m and $C_{\mathcal{M}}$ denotes the sum capacity of the users in subset $\mathcal{M}$. We assume an average transmit power constraint P_m for user m, and in this case, $C_{\mathcal{M}} = 0.5 \log(1 + \sum_{m \in \mathcal{M}} P_m/\sigma^2)$, where σ^2 denotes variance of the channel noise.

At iteration t, the received signal at the PS $\mathbf{y}_t$ is a function of all $\mathbf{x}_t^{(m)}$. The goal of the PS is to recover the average of the local gradients $\mathbf{g}_{\text{avg}}(\mathbf{w}_t) = (1/n_{\text{tot}}) \sum_{m=1}^{M} n_m \mathbf{g}^{(m)}(\mathbf{w}_t)$ from $\mathbf{y}_t$ using some post-processing function $h_t(\cdot)$. However, due to the pre- and post-processing, and the capacity region of the MAC, the PS can only recover the noisy versions of the local gradients $\hat{\mathbf{g}}^{(m)}(\mathbf{w}_t)$, thus, the noisy version of the average gradient $h_t(\mathbf{y}_t) = \hat{\mathbf{g}}_{\text{avg}}(\mathbf{w}_t) = (1/n_{\text{tot}}) \sum_{m=1}^{M} n_m \hat{\mathbf{g}}^{(m)}(\mathbf{w}_t)$. Therefore, the transmission from the users must ensure that the gradients received at the PS are unbiased estimators of $\mathbf{g}^{(m)}(\mathbf{w}_t)$ and have bounded variance, i.e.,

$$\mathbb{E}\left[\hat{\mathbf{g}}^{(m)}(\mathbf{w}_t)\right] = \mathbf{g}^{(m)}(\mathbf{w}_t), \ \text{Var}(\hat{\mathbf{g}}^{(m)}(\mathbf{w}_t)) \leq \epsilon_m, \tag{4}$$

where the variance bound ϵ_m should be as small as possible.

Problem Statement When jointly transmitting over a MAC, it is critical to allocate resources efficiently to ensure that the gradient aggregation can be done in a timely manner, and the training error is low. Let $\{r_1, \dots, r_M\}$ be the set of rates allocated to users for gradient transmission over the MAC. In this work, we want to understand how one should allocate rates as a function of the capacity region of the MAC, and the underlying informativeness of the gradients at different users. Furthermore, we want to characterize the resulting trade-off between the underlying channel conditions of the MAC and the convergence rate of GD algorithms.

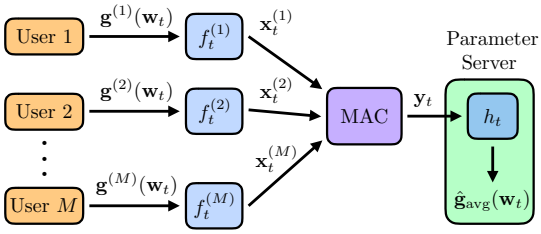

Fig. 1: FL over a MAC. At each iteration, users send their local gradients $\mathbf{g}^{(m)}(\mathbf{w}_t)$ through a MAC. The PS aggregates the gradients, updates the model and sends the updated model back to users for subsequent iteration.

III. Main Results

In this section, we present our proposed stochastic gradient quantization scheme for GD, which is inspired by schemes in [9], [10], [25]. In this scheme, the PS asks users to quantize their local gradients before sending them based on individual quantization budgets. The quantization budgets are found by the PS by solving an optimization problem that aims to minimize the variance of the aggregated gradients, while satisfying the transmission rate constraints imposed by the MAC. The distinction between our scheme and the scheme in [10] is that we allow each user to have its own quantization budget. We first present the proposed scheme for any number of users M, analyze the convergence rate of the scheme, and present a general optimization problem for quantization budget allocation based on the capacity of the MAC. We then show an example with $M = 2$ users and solve for the optimal quantization budget and communication rate for each user.

A. Stochastic Multi-level Gradient Quantization

At each iteration t, each user m computes the local gradient vector $\mathbf{g}^{(m)}(\mathbf{w}_t)$ using its local data set $\mathcal{D}_t^{(m)}$, $m = 1, \dots, M$. For simplicity of notation, we drop the iteration index t in describing the quantization scheme. Each user computes the dynamic range of its local gradient, i.e., $\Delta_m = g_{\max}^{(m)} - g_{\min}^{(m)}$, where $g_{\max}^{(m)}$ and $g_{\min}^{(m)}$ are the maximum and minimum values of the local gradient vector at user m. The user then quantizes its local gradient vector using the stochastic multi-level quantization scheme as we describe next. For every integer $r \in [0, k_m)$, we define

$$G^{(m)}(r) \triangleq g_{\min}^{(m)} + \frac{r \Delta_m}{k_m - 1}, \tag{5}$$

where $k_m \geq 2$ is the quantization budget for user m. For each element i in the local gradient vector, if $g_i^{(m)} \in [G^{(m)}(r), G^{(m)}(r+1))$, then $g_i^{(m)}$ is quantized as follows,

$$Q(g_i^{(m)}) = \begin{cases} G^{(m)}(r+1) & \text{w.p. } \frac{g_i^{(m)} - G^{(m)}(r)}{G^{(m)}(r+1) - G^{(m)}(r)} \\ G^{(m)}(r) & \text{otherwise} \end{cases}. \tag{6}$$

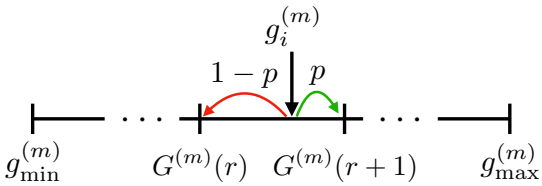

Fig. 2: Stochastic multi-level gradient quantization where the dynamic range of the gradient vector is split into k_m levels. Subsequently, each element of the vector $\mathbf{g}_i^{(m)}$ is quantized to $G^{(m)}(r)$ with probability p as shown in (6), or to $G^{(m)}(r+1)$ with probability $1-p$.

This operation is shown in Fig. 2. Once the entire gradient vector is quantized, user m sends its quantized gradient vector $\mathbf{Q}(\mathbf{g}^{(m)}) = [Q(g_1^{(m)}), \ldots, Q(g_d^{(m)})]$ to the PS over the Gaussian MAC. We assume that before each iteration, each user describes the scalars $g_{\max}^{(m)}$ and $g_{\min}^{(m)}$ (which describe the dynamic range $\Delta_m = g_{\max}^{(m)} - g_{\min}^{(m)}$ of the local gradient) at full resolution to the PS. In addition, as each element in the gradient vector is quantized to be one of the k_m levels, hence, a total of $d\log_2 k_m$ bits are required to describe the quantized gradient vector. Comparing to sending the gradient vector, sending scalars $g_{\max}^{(m)}$ and $g_{\min}^{(m)}$ costs significantly less. Hence, we omit the cost of sending $g_{\max}^{(m)}$ and $g_{\min}^{(m)}$ in this paper. The PS recovers all the quantized gradient vectors by performing optimal decoding over the MAC. Thus, for reliable decoding, the transmission rates of the users, i.e., $r_m = d\log_2 k_m$ must be within the MAC capacity region.

The PS then aggregates the quantized gradients as

$$\hat{\mathbf{g}}_t = \frac{1}{n_{\text{tot}}} \sum_{m=1}^{M} n_m \mathbf{Q}(\mathbf{g}_t^{(m)}), \tag{7}$$

and updates the model using, $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \hat{\mathbf{g}}_t$, where η_t is the learning rate. The updated model is then transmitted back to the users for subsequent iterations.

Suppose that in the tth iteration, the dynamic range of the gradient vector of user m is $\Delta_{t,m}$, and the number of quantization levels used is $k_{t,m}$. Then, it can be readily checked that $Q(g_{t,i}^{(m)})$ is an unbiased estimator of $g_{t,i}^{(m)}$, i.e., $E\left[Q(g_{t,i}^{(m)})\right] = g_{t,i}^{(m)}$. The variance can be computed as, $\text{Var}(Q(g_{t,i}^{(m)})) \leq \Delta_{t,m}^2/4(k_{t,m}-1)^2$. Therefore, the variance of the quantized gradient vector at user m in iteration t can be bounded as

$$\text{Var}(\mathbf{Q}(\mathbf{g}_t^{(m)})) = \sum_{i=1}^{d} \text{Var}(Q(g_{t,i}^{(m)})) \leq \frac{d\Delta_{t,m}^2}{4(k_{t,m}-1)^2}. \tag{8}$$

We next present our first result which shows how the convergence of the above algorithm depends on the parameters of multi-level stochastic quantization at the users.

Theorem 1. *If the loss function $L(\cdot)$ is λ-strongly convex and μ-smooth, with L_g-Lipschitz gradients, then by using a time varying learning rate of $\eta_t = 1/(\lambda t)$, we have the following convergence result:*

$$E\left[L(\mathbf{w}_T)\right] - L(\mathbf{w}^*) \leq \frac{2\mu}{\lambda^2 T^2} \sum_{t=1}^{T} \left(\frac{1}{n_{tot}^2} \sum_{m=1}^{M} \frac{n_m^2 d\Delta_{t,m}^2}{4(k_{t,m}-1)^2} + L_g^2 \right) \tag{9}$$

The proof of this Theorem is presented in Appendix I.

From Theorem 1, we observe that the convergence rate depends directly on the following factors: a) the dynamic range of the gradients ($\{\Delta_{t,m}\}$) computed by the users, and b) the quantization levels assigned to the users in each iteration. The traditional approach is to assign equal quantization levels to all users, i.e., $k_{t,m} = k$, for all m, t. However, the above expression shows that in order to maximize the rate of convergence, users whose gradients have a higher dynamic range must be assigned a higher quantization budget. On the other hand, if the users are communicating to the PS in a communication constrained setting, such as a MAC, then the quantization budget $k_{t,m}$, which is directly related to the transmission rate cannot exceed the constraints imposed by the capacity region of the MAC.

B. MAC Aware Gradient Quantization

Motivated by the above discussion, we propose MAC aware gradient quantization which works as follows. In each iteration t, a) users compute their local gradients $\mathbf{g}_t^{(m)}$, and describe $g_{t,\min}^{(m)}, g_{t,\max}^{(m)}$ to the PS. b) using these scalars, PS computes the dynamic range(s) $\left(\{\Delta_{t,m} = g_{t,\max}^{(m)} - g_{t,\min}^{(m)}\}\right)$ of the gradients for all the users and performs the optimization described in Theorem 2. Subsequently, the PS assigns individual quantization budgets (transmission rates) to each user; c) users subsequently quantize their gradients and transmit over the MAC. In the following Theorem, we present the optimization problem using which we can determine the optimal $k_{t,m}^*$'s that maximize the convergence rate.

Theorem 2. *At each iteration t, the optimal $k_{t,m}^*$'s that give the best convergence rate can be found by solving the following optimization problem,*

$$\min_{\{k_{t,m}\}_{m=1}^M} \sum_{m=1}^{M} \frac{n_m^2 d\Delta_{t,m}^2}{4(k_{t,m}-1)^2} \tag{10}$$
$$\text{s.t.} \sum_{m\in\mathcal{M}} r_{t,m} \leq sC_{\mathcal{M}},\ \mathcal{M} \subset [M],\ |\mathcal{M}| = 1, \ldots, M,$$
$$k_{t,m} \in \mathbb{Z}^+,\ \forall m$$

where s denotes the number of channel uses, $r_{t,m} = d\log_2 k_{t,m}$ denotes the transmission rate of user m and $C_{\mathcal{M}} = 0.5\log(1 + \sum_{m\in\mathcal{M}} P_m/\sigma^2)$ denotes the sum capacity of the users in subset $\mathcal{M}$, where σ^2 denotes variance of the channel noise.

The above optimization problem falls into the category of constrained integer programming since $k_{t,m}$'s take non-negative integer values. In general, integer programming is considered to be NP-hard problem [26]. However, one could obtain sub-optimal solutions by relaxing the constraint on $k_{t,m}$'s. For instance, by allowing $k_{t,m}$'s to be real numbers greater or equal to 2 (so that each user gets at least 1 bit), it is easy to verify that the above problem becomes a convex optimization problem. One could then either use convex solvers or solve the convex problem analytically by checking KKT conditions, and round the results. We next show an example for $M = 2$ users,

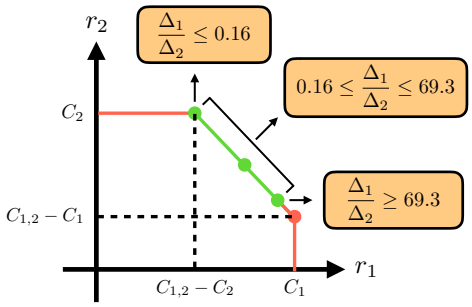

Fig. 3: The capacity region of the Gaussian MAC when $P_1 = 80$, $P_2 = 20$. Green area denote points that achieve maximum sum rate.

	$\frac{\Delta_1}{\Delta_2} \leq 0.16$	$\frac{\Delta_1}{\Delta_2} = 1$	$\frac{\Delta_1}{\Delta_2} \geq 69.3$
Quantization Levels for User 1	4	10	50
Quantization Levels for User 2	21	10	2

TABLE I: Per-user quantization budget based on ratio of dynamic range of the gradients, i.e., Δ_1/Δ_2 and the capacity region of MAC.

and solve the convex relaxation analytically to gain insights on how the dynamic ranges of the gradients, and the capacity region of MAC impact the resulting quantization budgets.

C. Solution for the Relaxed Optimization Problem with $M = 2$

For $M = 2$ users, the relaxed optimization problem $(\mathcal{P})$ is given as follows:

$$\mathcal{P}: \min_{(k_1,k_2)} \frac{n_1^2 d\Delta_1^2}{4(k_1-1)^2} + \frac{n_2^2 d\Delta_2^2}{4(k_2-1)^2} \quad (11)$$
$$\text{s.t. } d\log_2 k_1 \leq sC_1, \quad d\log_2 k_2 \leq sC_2$$
$$d(\log_2 k_1 + \log_2 k_2) \leq sC_{1,2}$$

The three constraints on rates can be rearranged as follows:

$$k_1 \leq 2^{\widetilde{C}_1}, \quad k_2 \leq 2^{\widetilde{C}_2}, \quad k_1 k_2 \leq 2^{\widetilde{C}_{1,2}}, \quad (12)$$

where $\widetilde{C}_m = sC_m/d$, $m = 1, 2$ and $\widetilde{C}_{1,2} = sC_{1,2}/d$. As mentioned earlier, the objective function being minimized is a convex function when k_1 and k_2 are both greater or equal to 2. The $M = 2$-user case can be solved analytically by first forming the following Lagrangian function,

$$J = \frac{n_1^2 d\Delta_1^2}{4(k_1-1)^2} + \frac{n_2^2 d\Delta_2^2}{4(k_2-1)^2} + \lambda_1(k_1 - 2^{\widetilde{C}_1})$$
$$+ \lambda_2(k_2 - 2^{\widetilde{C}_2}) + \lambda_3(k_1 k_2 - 2^{\widetilde{C}_{1,2}}). \quad (13)$$

We note that to fully utilize the channel, the sum-rate constraint in $\mathcal{P}$ should be satisfied with equality, i.e., $d(\log_2 k_1 + \log_2 k_2) = sC_{1,2}$ or equivalently, $k_1 k_2 = 2^{\widetilde{C}_{1,2}}$. By taking the partial derivatives of J with respect to k_1 and k_2 and checking the KKT conditions, we obtain,

$$\lambda_1 = \lambda_2 = 0, \ \lambda_3 = \frac{n_1^2 d\Delta_1^2}{2k_2(k_1-1)^3} = \frac{n_2^2 d\Delta_2^2}{2k_1(k_2-1)^3}. \quad (14)$$

Using this condition and the sum-rate constraint, i.e., $k_1 k_2 = 2^{\widetilde{C}_{1,2}}$, we can solve for the optimal quantization budgets.

Theorem 3. *For a 2-user Gaussian MAC, the optimal quantization budgets k_1^* and k_2^* for $\mathcal{P}$ can be found by solving*

$$\frac{\Delta_1}{\Delta_2} = \frac{n_2}{n_1}\left(\frac{2^{\widetilde{C}_{1,2}} k_1^*(k_1^*-1)^3}{(2^{\widetilde{C}_{1,2}} - k_1^*)^3}\right)^{1/2}, \quad (15)$$

and subsequently $k_2^* = 2^{\widetilde{C}_{1,2}}/k_1^*$, where Δ_1 and Δ_2 are dynamic ranges of gradients at users 1 and 2.

To show the impact of dynamic ranges and the capacity region of the MAC, we let $n_1 = n_2$, and solve k_1^* and k_2^* numerically with the following parameters: we let $d = 7850$, $s = 2d$, $P_1 = 80, P_2 = 20$, so that the individual and sum capacities for this setting are $C_1 = 3.1699$, $C_2 = 2.1962$ and $C_{1,2} = 3.3291$. These lead to $k_1 \leq 80.9$, $k_2 \leq 21$ and $k_1 k_2 \leq 100.9$. We fix $\Delta_2 = 50$ and vary Δ_1 from 1 to 3500 to understand the impact of the ratio of dynamic range Δ_1/Δ_2 on the quantization budgets. It can be seen in Fig. 3 and Table I that by using proposed MAC aware scheme, the PS allocates more rate towards the user whose gradients are more informative (higher dynamic range). For instance, when $\Delta_1/\Delta_2 = 1$, gradients from both users are equally informative, and both users are assigned equal quantization budgets $k_1 = k_2 = 10$. On one extreme, when $\Delta_1/\Delta_2 \leq 0.16$, gradients from user 2 are considered more useful than user 1, the optimal allocation is $k_1 = 4$, $k_2 = 21$. On the other extreme, if $\Delta_1/\Delta_2 \geq 69.28$, gradients from user 1 are more informative, hence we see that $k_1 = 50$, and $k_2 = 2$.

IV. Experiments

To show the performance of our proposed scheme, we consider MNIST image classification task using single layer neural networks trained on 60000 training and 10000 testing samples with $M = 2$ users, and a cross-entropy loss function. The dimensionality of the classifier model is $d = 7850$. We assume that user 1's data set $\mathcal{D}_1$ consists of images belonging to digits '0' and '1', whereas the data set of user 2 consists of all the 10 digits. The channel noise variance is set as $\sigma^2 = 1$, and the total transmit power per iteration is set as $\bar{P} = 100$. We use the MAC for $s = 2d$ channel uses for each iteration.

In Fig. 4, we let $P_1 = 0.95\bar{P}$ and $P_2 = 0.05\bar{P}$, and compare the proposed MAC aware gradient quantization scheme with the following schemes: a) uniform rate allocation subject to MAC capacity constraints, b) a recently proposed digital scheme in [11], c) SignSGD, which uses 1 bit quantization per dimension for each user [2], and d) TernGrad [3], which uses three levels $\{-1, 0, +1\}$ to quantize each dimension of the gradient. We also plot the non-quantized full resolution scheme as a baseline. In the digital scheme proposed in [11], all but the highest q_t and lowest q_t gradient values are set to zero. The remaining gradient values are then split into two groups depending on their signs. The mean of elements in each group is computed, denoted by α_{avg}^+ and α_{avg}^-. If $\alpha_{\text{avg}}^+ > |\alpha_{\text{avg}}^-|$ ($\alpha_{\text{avg}}^+ < |\alpha_{\text{avg}}^-|$), all remaining positive (negative) values will be set to α_{avg}^+ (α_{avg}^-). Each user then transmits the location of q_t non-zero values and a scalar (using c bits) to describe the average value at each iteration. Therefore, the communication cost is $\log_2 \binom{d}{q_t} +$

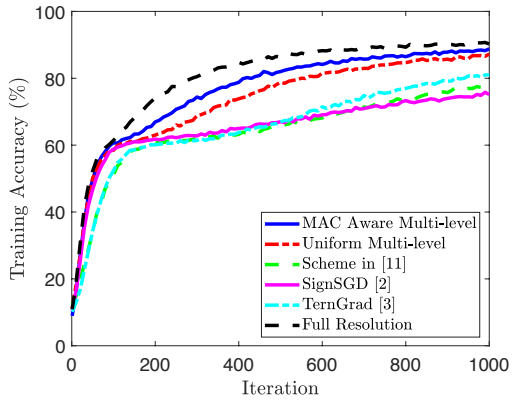

Fig. 4: Training accuracy comparison between MAC aware gradient quantization, uniform rate allocation, digital scheme proposed in [11], SignSGD [2], TernGrad [3], and full resolution when the total transmit power per iteration is $\bar{P} = 100$ and $s = 2d$.

MAC Aware	Uniform	[11]	SignSGD	TernGrad	Full Resolution
79.8%	76.7%	57.9%	52.3%	64.5%	84%

TABLE II: Comparison of test accuracy after $T = 1000$ iterations.

c. This scheme [11] is fundamentally different than the one proposed in this paper, and, moreover, the quantization budget q_t is the same for all users. As shown in Fig. 4, the proposed MAC aware multi-level scheme outperforms the uniform multilevel scheme, the scheme in [11], SignSGD and TernGrad. This is due to the fact that $\log\binom{d}{q_t}$ grows exponentially as q_t increases. In addition, the rates are limited by the user with the worst channel. Therefore, as it reaches the capacity of the user with the worst channel, q_t is still small compared to d. Other schemes such as SignSGD and TernGrad suffer from underutilization of channel resources, as they use a fixed quantization budget (1 bit, and 2 bits respectively per gradient dimension). We also show the testing accuracy of each scheme at the end of 1000 iterations (see Table II). They are consistent with Fig. 4 where our proposed scheme is the closest to full resolution.

For Fig. 5, we set $s = 1.5d$, $P_1 = 0.8\bar{P}$ and $P_2 = 0.2\bar{P}$, and vary $\bar{P}$ to see the impact of increasing power, and thus, a larger capacity region. It can be seen in Fig. 5 that the performance improves monotonically with the increase in total power. The testing accuracy at the end of $T = 1000$ iterations is shown in Table III as a function of the total power.

V. Conclusions

In this paper, we considered the problem of MAC aware gradient quantization for federated learning. We showed that when designing digital FL schemes over MACs, there are new opportunities to assign different amount of resources (such as quantization rates) to different users based on a) the informativeness of the gradients at each user, captured by their dynamic range, and b) the underlying channel conditions. We studied and analyzed a *channel aware quantization* scheme and showed that it outperforms uniform quantization and other existing digital schemes. An interesting future direction is to explore if other quantization schemes (for instance, the scheme

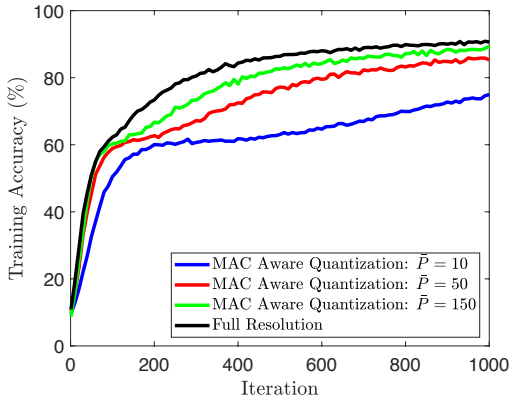

Fig. 5: Training accuracy comparison for MAC aware gradient quantization with total power per iteration $\bar{P} = 10, 50, 150$, and $P_1 = 0.8\bar{P}$ and $P_2 = 0.2\bar{P}$.

MAC Aware	$\bar{P} = 10$	$\bar{P} = 50$	$\bar{P} = 150$	Full Resolution
	51%	74.7%	80.2%	84%

TABLE III: Test accuracy for proposed scheme as a function of total power.

in [11], or gradient sparsification schemes in [4], [5]) can be optimized (with limited interaction with the PS) as a function of the underlying communication channel such as MAC.

Appendix I: Proof of Theorem 1

Standard convergence results in [27] have shown that for a loss function $L(\cdot)$ that is λ-strongly convex and μ-smooth w.r.t. $\mathbf{w}^*$, using SGD with stochastic unbiased gradients, bounded second order moments, i.e., $E[\|\hat{\mathbf{g}}_t\|_2^2] \leq G^2$, with a learning rate of $\eta_t = 1/\lambda t$ can achieve a convergence result:

$$E\left[L(\mathbf{w}_T)\right] - L(\mathbf{w}^*) \leq \frac{2\mu G^2}{\lambda^2 T}. \tag{16}$$

There are two distinctions between our bound and (16). First, the randomness in our scheme comes from quantizing the gradients instead of randomly selecting data points. Second, as users can have different quantization budgets per iteration, the resulting variance is iteration dependent, i.e.,, $E[\|\hat{\mathbf{g}}_t\|_2^2] \leq G_t^2$. By slightly modifying the proof in [27], it is possible to prove the following convergence result (proof omitted due to space):

$$E\left[L(\mathbf{w}_T)\right] - L(\mathbf{w}^*) \leq \frac{2\mu}{\lambda^2 T}\left(\sum_{t=1}^{T} G_t^2/T\right). \tag{17}$$

Theorem 1 now follows directly by plugging in the values of G_t^2, which can be computed as:

$$\begin{aligned} E[\|\hat{\mathbf{g}}_t\|_2^2] &= \text{Var}(\hat{\mathbf{g}}_t) + \|\mathbf{g}_t\|_2^2 \\ &= \frac{1}{n_{\text{tot}}^2}\sum_{m=1}^{M} n_m^2 \text{Var}(Q(\mathbf{g}_t^{(m)})) + \|\mathbf{g}_t\|_2^2 \\ &\overset{(a)}{\leq} \frac{1}{n_{\text{tot}}^2}\sum_{m=1}^{M} \frac{n_m^2 d \Delta_{t,m}^2}{4(k_{t,m}-1)^2} + L_g^2 \triangleq G_t^2, \end{aligned} \tag{18}$$

where (a) follows from (8) and Lipschitz assumption, i.e., $\|\mathbf{g}_t\|_2^2 \leq L_g^2$.

REFERENCES

[1] F. Seide, H. Fu, J. Droppo, G. Li, and D. Yu, "1-Bit Stochastic Gradient Descent and Application to Data-Parallel Distributed Training of Speech DNNs," in *Interspeech 2014*, September 2014.

[2] J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar, "signSGD: Compressed optimisation for non-convex problems," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 80, 10–15 Jul 2018, pp. 560–569.

[3] W. Wen, C. Xu, F. Yan, C. Wu, Y. Wang, Y. Chen, and H. Li, "TernGrad: Ternary Gradients to Reduce Communication in Distributed Deep Learning," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 1509–1519.

[4] A. F. Aji and K. Heafield, "Sparse communication for distributed gradient descent," *CoRR*, vol. abs/1704.05021, 2017. [Online]. Available: http://arxiv.org/abs/1704.05021

[5] J. Wangni, J. Wang, J. Liu, and T. Zhang, "Gradient sparsification for communication-efficient distributed optimization," in *Advances in Neural Information Processing Systems 31*, 2018, pp. 1299–1309.

[6] N. Dryden, S. A. Jacobs, T. Moon, and B. Van Essen, "Communication quantization for data-parallel training of deep neural networks," in *Proceedings of the Workshop on Machine Learning in High Performance Computing Environments*, ser. MLHPC '16. Piscataway, NJ, USA: IEEE Press, 2016, pp. 1–8. [Online]. Available: https://doi.org/10.1109/MLHPC.2016.4

[7] Y. Lin, S. Han, H. Mao, Y. Wang, and W. J. Dally, "Deep gradient compression: Reducing the communication bandwidth for distributed training," *CoRR*, vol. abs/1712.01887, 2017. [Online]. Available: http://arxiv.org/abs/1712.01887

[8] F. Sattler, S. Wiedemann, K. Müller, and W. Samek, "Sparse binary compression: Towards distributed deep learning with minimal communication," *CoRR*, vol. abs/1805.08768, 2018. [Online]. Available: http://arxiv.org/abs/1805.08768

[9] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "QSGD: Communication-Efficient SGD via Gradient Quantization and Encoding," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 1709–1720.

[10] A. T. Suresh, F. X. Yu, S. Kumar, and H. B. McMahan, "Distributed mean estimation with limited communication," in *Proceedings of the 34th International Conference on Machine Learning*, 2017, p. 3329–3337.

[11] M. M. Amiri and D. Gündüz, "Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air," *CoRR*, vol. abs/1901.00844, 2019. [Online]. Available: http://arxiv.org/abs/1901.00844

[12] M. M. Amiri and D. Gündüz, "Over-the-air machine learning at the wireless edge," in *2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, July 2019, pp. 1–5.

[13] M. M. Amiri and D. Gündüz, "Federated learning over wireless fading channels," *CoRR*, vol. abs/1907.09769, 2019. [Online]. Available: http://arxiv.org/abs/1907.09769

[14] M. M. Amiri, T. M. Duman, and D. Gündüz, "Collaborative machine learning at the wireless edge with blind transmitters," *CoRR*, vol. abs/1907.03909, 2019. [Online]. Available: http://arxiv.org/abs/1907.03909

[15] M. S. H. Abad, E. Ozfatura, D. Gündüz, and O. Ercetin, "Hierarchical federated learning across heterogeneous cellular networks," *CoRR*, vol. abs/1909.02362, 2019. [Online]. Available: http://arxiv.org/abs/1909.02362

[16] 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," *CoRR*, vol. abs/1909.07972, 2019. [Online]. Available: http://arxiv.org/abs/1909.07972

[17] K. Yang, T. Jiang, Y. Shi, and Z. Ding, "Federated learning via over-the-air computation," *CoRR*, vol. abs/1812.11750, 2018. [Online]. Available: http://arxiv.org/abs/1812.11750

[18] Q. Zeng, Y. Du, K. K. Leung, and K. Huang, "Energy-efficient radio resource allocation for federated edge learning," *CoRR*, vol. abs/1907.06040, 2019. [Online]. Available: http://arxiv.org/abs/1907.06040

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

[20] Y. Sun, S. Zhou, and D. Gündüz, "Energy-aware analog aggregation for federated learning with redundant data," *CoRR*, vol. abs/1911.00188, 2019. [Online]. Available: http://arxiv.org/abs/1911.00188

[21] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, "When edge meets learning: Adaptive control for resource-constrained distributed machine learning," *CoRR*, vol. abs/1804.05271, 2018. [Online]. Available: http://arxiv.org/abs/1804.05271

[22] T. Sery and K. Cohen, "On analog gradient descent learning over multiple access fading channels," *CoRR*, vol. abs/1908.07463, 2019. [Online]. Available: http://arxiv.org/abs/1908.07463

[23] T. Sery and K. Cohen, "A sequential gradient-based multiple access for distributed learning over fading channels," in *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Sep. 2019, pp. 303–307.

[24] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley-Interscience, 2006.

[25] N. Agarwal, A. T. Suresh, F. Yu, S. Kumar, and H. B. Mcmahan, "cpSGD: Communication-efficient and differentially-private distributed SGD," *CoRR*, vol. abs/1805.10559, 2018. [Online]. Available: http://arxiv.org/abs/1805.10559

[26] A. Schrijver, *Theory of linear and integer programming*. John Wiley & Sons, 1998.

[27] A. Rakhlin, O. Shamir, and K. Sridharan, "Making gradient descent optimal for strongly convex stochastic optimization," *CoRR*, vol. abs/1109.5647, 2012. [Online]. Available: http://arxiv.org/abs/1109.5647