UVeQFed: Universal Vector Quantization for Federated Learning

Nir Shlezinger, *Member, IEEE*, Mingzhe Chen, Yonina C. Eldar, *Fellow, IEEE*, H. Vincent Poor, *Fellow, IEEE*, and Shuguang Cui, *Fellow, IEEE*

***Abstract*—Traditional deep learning models are trained at a centralized server using data samples collected from users. Such data samples often include private information, which the users may not be willing to share. Federated learning (FL) is an emerging approach to train such learning models without requiring the users to share their data. FL consists of an iterative procedure, where in each iteration the users train a copy of the learning model locally. The server then collects the individual updates and aggregates them into a global model. A major challenge that arises in this method is the need of each user to repeatedly transmit its learned model over the throughput limited uplink channel. In this work, we tackle this challenge using tools from quantization theory. In particular, we identify the unique characteristics associated with conveying trained models over rate-constrained channels, and propose a suitable quantization scheme for such settings, referred to as universal vector quantization for FL (UVeQFed). We show that combining universal vector quantization methods with FL yields a decentralized training system in which the compression of the trained models induces only a minimum distortion. We then theoretically analyze the distortion, showing that it vanishes as the number of users grows. We also characterize how models trained with conventional federated averaging combined with UVeQFed converge to the model which minimizes the loss function. Our numerical results demonstrate the gains of UVeQFed over previously proposed methods in terms of both distortion induced in quantization and accuracy of the resulting aggregated model.**

Manuscript received July 18, 2020; revised November 22, 2020; accepted December 12, 2020. Date of publication December 23, 2020; date of current version January 21, 2021. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Byonghyo Shim. The work of Yonina C. Eldar was supported in part by 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 646804-ERC-COG-BNYQ, and in part by the Israel Science Foundation under Grant 0100101. The work of H. Vincent Poor was supported in part by the U.S. National Science Foundation under Grants CCF-0939370 and CCF-1908308. The work of Shuguang Cui was supported in part by the Key Area R&D Program of Guangdong Province under Grant 2018B030338001. This paper was presented in part at the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2020 as the paper [1]. *(Corresponding author: Nir Shlezinger.)*

Nir Shlezinger is with the School of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Be'er-Sheva 8410501, Israel (e-mail: nirshlezinger1@gmail.com).

Mingzhe Chen is with the Electrical Engineering Department, Princeton University, Princeton, NJ 08544 USA, and also with the Chinese University of Hong Kong, Shenzhen 518172, China (e-mail: mingzhec@princeton.edu).

Yonina C. Eldar is with the Faculty of Mathematics and Computer Science, Weizmann Institute of Science, Rehovot 7610001, Israel (e-mail: yonina.eldar@weizmann.ac.il).

H. Vincent Poor is with the Electrical Engineering Department, Princeton University, Princeton, NJ 08544 USA (e-mail: poor@princeton.edu).

Shuguang Cui is with the Shenzhen Research Institute of Big Data and Future Network of Intelligence Institute (FNii), Chinese University of Hong Kong, Shenzhen 518172, China (e-mail: shuguangcui@cuhk.edu.cn).

Digital Object Identifier 10.1109/TSP.2020.3046971

***Index Terms*—Federated learning, quantization.**

I. Introduction

MACHINE learning methods have demonstrated unprecedented performance in a broad range of applications [2]. This is achieved by training a deep network model based on a large number of labeled training samples. Often, these samples are gathered on end devices or users, such as smartphones, while the deep model is maintained by a computationally powerful centralized server [3]. Traditionally, the users send their labeled data to the server, who in turn uses the massive amount of samples to train the model. However, data often contains private information, which the users may prefer not to share, and having each user transmit large volumes of training data to the server may induce a substantial load on the communication link. This gives rise to the need to adapt the network on the end-devices., i.e., train a centralized model in a distributed fashion [4]. Federated learning (FL) proposed in [5], is a method to update such decentralized models. Instead of requiring the users to share their possibly private labeled data, each user trains the network locally, and conveys its trained model updates to the server. The server then iteratively aggregates these updates into a global network [6], [7], commonly using some weighted average, also known as *federated averaging* [5].

One of the major challenges of FL is the transfer of a large number of updated model parameters over the uplink communication channel from the users to the server, whose throughput is typically constrained [5], [7]–[9]. This challenge can be tackled by reducing the number of participating users, via, e.g., scheduling policies [10], [11]. An alternative strategy is to reduce the volume of data each user conveys, via sparsification or scalar quantization [12]–[21]. The work [12] proposed various methods for compressing the updates sent from the users to the server. These methods include random masks, subsampling, and probabilistic quantization. Sparsifying masks for compressing the gradients were proposed in [13]–[15]. Additional forms of probabilistic scalar quantization for FL were considered in [16]–[20]. However, these approaches are suboptimal from a quantization theory perspective, as, e.g., discarding a random subset of the gradients can result in dominant distortion, while scalar quantization is inferior to vector quantization [22, Ch. 23]. This motivates the design and analysis of quantization methods for facilitating updated model transfer in FL, which minimize the error induced by quantization in the aggregated global model.

Here, we design quantizers for distributed training by tackling the uplink compression in FL problem from a quantization theory perspective. We first discuss the requirements which one have to account for, and can possibly exploit, when designing quantization schemes for FL. We specifically identify that such quantization schemes are required to operate without knowing the distribution of the model updates, as such knowledge is unlikely to be available in FL. We also note that the repeated communications between the server and users imply that they can share a source of local randomness, by, e.g., sharing a random seed, which can be utilized by the quantization mechanism.

Based on these properties, we propose a scheme following concepts from universal quantization [23], referred to as unviersal vector quantization for federated learning (UVeQFed). UVeQFed implements *subtractive dithered lattice quantization*, which is based on solid information theoretic arguments. In particular, such schemes are known to approach the most accurate achievable finite-bit representation, dictated by rate-distortion theory, to within a controllable gap [24], as well as achieve more accurate quantized representation compared to scalar quantization methods (probabilistic or deterministic) used in existing FL works, while meeting the aforementioned requirements. Consequently, UVeQFed allows FL to operate reliably under strict bit rate constraints, due to its ability to reduce the distortion induced by the need to quantize the model updates, which results in more accurate learned models with faster convergence compared to previously proposed methods.

We theoretically analyze the ability of the server to accurately recover the updated model when UVeQFed is employed. We show that the error induced by UVeQFed is mitigated by conventional federated averaging, and analyze the convergence of the global model to the one which minimize the loss function. Specifically, our analysis reveals that the resulting quantization error can be bounded by a term which vanishes as the number of users grows, regardless of the statistical model from which the data of each user is generated. This rigorously proves that the quantization distortion can be made arbitrarily small when a sufficient number of users contribute to the overall model. Then, we study the convergence of stochastic gradient descent (SGD)-based federated averaging with UVeQFed in a statistically heterogeneous setup, where the training available at each user obeys a different distribution, as is commonly the case in FL [7], [9]. We prove that for strongly convex and smooth objectives, the expected distance between the resulting FL performance and the optimal one asymptotically decays as one over the number of iterations, which is the same order of convergence reported for FL without communication constraints in heterogeneous setups [25]. Finally, we show that these theoretical gains translate into FL performance gains in a numerical study. We demonstrate that FL with UVeQFed yields more accurate global models and faster convergence compared to previously proposed quantization approaches for such setups when operating under tight bit constraints of two and four bits per sample, considering synthetic data as well as the MNIST and CIFAR-10 data sets.

The rest of this paper is organized as follows: Section II presents the system model and identifies the requirements of FL quantization. Section III details the proposed quantization system, and Section IV theoretically anaylzes its performance. Experimental results are presented in Section V. Section VI concludes the paper. Proofs of the results stated in the paper are detailed in the Appendix.

Throughout the paper, we use boldface lower-case letters for vectors, e.g., $\boldsymbol{x}$; Matrices are denoted with boldface upper-case letters, e.g., $\boldsymbol{M}$, where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix; calligraphic letters, such as $\mathcal{X}$, are used for sets. The ℓ_2 norm and stochastic expectation are written as $\|\cdot\|$ and $\mathbb{E}\{\cdot\}$, respectively. Finally, $\mathcal{R}$ and $\mathcal{Z}$ are the sets of real numbers and integers, respectively.

II. System Model

In this section we detail the considered setup of FL with bit-constrained model updates. To that aim, we first review the conventional FL setup in Section II-A. Then, in Section II-B, we formulate the problem and identify the unique requirements of quantizers utilized in FL systems.

A. Federated Learning

We consider the conventional FL framework proposed in [5]. Here, a centralized server is training a model consisting of m parameters based on labeled samples available at a set of K remote users, in order to minimize some loss function $\ell(\cdot;\cdot)$. Letting $\{\boldsymbol{x}_i^{(k)}, \boldsymbol{y}_i^{(k)}\}_{i=1}^{n_k}$ be the set of n_k labeled training samples available at the kth user, $k \in \{1,\ldots,K\} \triangleq \mathcal{K}$, FL aims at recovering the $m \times 1$ weights vector $\boldsymbol{w}^{\rm o}$ satisfying

$$\boldsymbol{w}^{\rm o} = \arg\min_{\boldsymbol{w}} \left\{ F(\boldsymbol{w}) \triangleq \sum_{k=1}^{K} \alpha_k F_k(\boldsymbol{w}) \right\}. \tag{1}$$

Here, the weighting average coefficients $\{\alpha_k\}$ are non-negative satisfying $\sum \alpha_k = 1$, and the local objective functions are defined as the empirical average over the corresponding training set, i.e.,

$$F_k(\boldsymbol{w}) \equiv F_k\left(\boldsymbol{w}; \{\boldsymbol{x}_i^{(k)}, \boldsymbol{y}_i^{(k)}\}_{i=1}^{n_k}\right) \triangleq \frac{1}{n_k} \sum_{i=1}^{n_k} \ell\left(\boldsymbol{w}; (\boldsymbol{x}_i^{(k)}, \boldsymbol{y}_i^{(k)})\right).$$

Federated averaging [5] aims at recovering $\boldsymbol{w}^{\rm o}$ using iterative subsequent updates. In each update of time instance t, the server shares its current model, represented by the vector $\boldsymbol{w}_t \in \mathcal{R}^m$, with the users. The kth user, $k \in \mathcal{K}$, uses its set of n_k labeled training samples to retrain the model $\boldsymbol{w}_t$ over τ time instances into an updated model $\tilde{\boldsymbol{w}}_{t+\tau}^{(k)} \in \mathcal{R}^m$.

Having updated the model weights, the kth user should convey its model update, denoted as $\boldsymbol{h}_{t+\tau}^{(k)} \triangleq \tilde{\boldsymbol{w}}_{t+\tau}^{(k)} - \boldsymbol{w}_t$, to the server. Since uploading throughput is typically more limited compared to its downloading counterpart [26], the kth user needs to communicate a finite-bit quantized representation of its model update. Quantization consists of encoding the model update into a set of bits, and decoding each bit combination into a recovered model update [27]. The kth model update $\boldsymbol{h}_{t+\tau}^{(k)}$ is therefore encoded into a digital codeword of R_k bits denoted as $u_t^{(k)} \in \{0,\ldots,2^{R_k}-1\} \triangleq \mathcal{U}_k$, using an encoding function

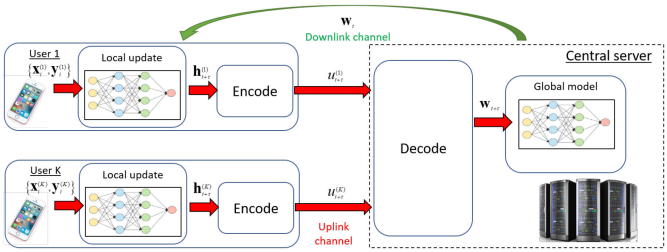


Fig. 1. Federated learning with bit rate constraints.

whose input is $\boldsymbol{h}_{t+\tau}^{(k)}$, i.e.,

$$e_{t+\tau}^{(k)} : \mathcal{R}^m \mapsto \mathcal{U}_k. \tag{2}$$

The uplink channel is modeled as a bit-constrained link, as commonly assumed in the FL literature [12]–[21]. In such channels, each R_k bit codeword is recovered by the server without errors, representing, e.g., coded communications at rates below the channel capacity, where arbitrarily small error rates can be guaranteed by proper channel coding. The server uses the received codewords $\{u_{t+\tau}^{(k)}\}_{k=1}^K$ to reconstruct $\hat{\boldsymbol{h}}_{t+\tau} \in \mathcal{R}^m$, obtained via a joint decoding function

$$d_{t+\tau} : \mathcal{U}_1 \times \ldots \times \mathcal{U}_K \mapsto \mathcal{R}^m. \tag{3}$$

The recovered $\hat{\boldsymbol{h}}_{t+\tau}$ is an estimate of the weighted average $\sum_{k=1}^K \alpha_k \boldsymbol{h}_{t+\tau}^{(k)}$. Finally, the global model $\boldsymbol{w}_{t+\tau}$ is updated via

$$\boldsymbol{w}_{t+\tau} = \boldsymbol{w}_t + \hat{\boldsymbol{h}}_{t+\tau}. \tag{4}$$

An illustration of this FL procedure is depicted in Fig. 1. Clearly, if the number of allowed bits is sufficiently large, the distance $\|\hat{\boldsymbol{h}}_{t+\tau} - \sum_{k=1}^K \alpha_k \boldsymbol{h}_{t+\tau}^{(k)}\|^2$ can be made arbitrarily small, allowing the server to update the global model as the desired weighted average, denoted $\boldsymbol{w}_{t+\tau}^{\text{des}}$, via:

$$\boldsymbol{w}_{t+\tau}^{\text{des}} = \sum_{k=1}^K \alpha_k \tilde{\boldsymbol{w}}_{t+\tau}^{(k)}. \tag{5}$$

In the presence of a limited bit budget, i.e., small values of $\{R_k\}$, distortion is induced which can severely degrade the ability of the server to update its model. To tackle this issue, various methods have been proposed for quantizing the model updates, commonly based on sparsification or probabilistic scalar quantization. These approaches are suboptimal from a quantization theory perspective, namely, the gap between the distortion they achieve in quantizing a signal using a given number of bit and the most accurate achievable finite-bit representation, dictated by rate-distortion theory, can be further reduced by, e.g., using vector quantization [22, Ch. 23]. This motivates the study of efficient and practical quantization methods for FL.

B. Problem Formulation

Our goal is to propose an encoding-decoding system which mitigates the effect of quantization errors on the ability of the server to accurately recover the updated model (5). To faithfully represent the FL setup, we design our quantization strategy in light of the following requirements and assumptions:

- *A*1 All users share the same encoding function, denoted as $e_t^{(k)}(\cdot) = e_t(\cdot)$ for each $k \in \mathcal{K}$. This requirement, which was also considered in [12], significantly simplifies FL implementation.
- *A*2 *a-priori* knowledge or distribution of $\boldsymbol{h}_{t+\tau}^{(k)}$ is assumed.
- *A*3 As in [12], the users and the server share a source of common randomness. This is achieved by, e.g., letting the server share with each user a random seed along with the weights. Once a different seed is conveyed to each user, it can be used to obtain a dedicated source of common randomness shared by server and each of the users for the entire FL procedure.

Requirement *A*2 gives rise to the need for a *universal quantization* approach, namely, a scheme which operates reliably regardless of the distribution of the model updates and without its prior knowledge. In light of the above requirements, we propose UVeQFed in the following section.

III. UVeQFed

We now propose UVeQFed, which conveys the model updates $\{\boldsymbol{h}_{t+\tau}^{(k)}\}$ from the users to the server over the rate-constrained channel using a universal quantization method. Specifically, the scheme encodes each model update using *subtractive dithered lattice quantization* [23], which operates in the same manner for each user, satisfying *A*1. UVeQFed allows the server to recover the updates with small average error regardless of the distribution of $\{\boldsymbol{h}_{t+\tau}^{(k)}\}$, as required in *A*2, by exploiting the source of common randomness assumed in *A*3. In addition to its compliance with the model requirements stated in Section II-B, the proposed approach is particularly suitable for FL, as the distortion is mitigated by federated averaging, as we prove in Section IV. This significantly improves the overall FL capabilities, as numerically demonstrated in Section V. The proposed quantization method is detailed in Section III-A, followed by a discussion in Section III-B.

A. Quantization Scheme

Here, we present the encoding and decoding functions, $e_{t+\tau}(\cdot)$ and $d_{t+\tau}(\cdot)$. Following requirement *A*1, we utilize universal vector quantization, i.e., a quantization scheme which maps each set of continuous-amplitude values into a discrete representation in a manner which is ignorant of the underlying distribution. Common universal quantization methods are based on selection from an ensemble of source codes [28], or alternatively, on subtractive dithering [23], [29]–[33], where the latter is simpler to implement being based on adding dither, i.e., noise, to the discretized quantity, but requires knowledge of the dither as it is subtracted from the discrete quantity when parsing the quantized value. The source of common randomness assumed in *A*3 implies that the server and the users can generate the same realizations of a dither signal. We thus design UVeQFed based on dithered vector quantization, and particularly, on lattice quantization, detailed in the following.

Let L be a fixed positive integer, referred to henceforth as the lattice dimension, and a let $\boldsymbol{G}$ be a non-singular $L \times L$ matrix,

which denotes the lattice generator matrix. For simplicity, we assume that $M \triangleq \frac{m}{L}$ is an integer, where m is the number of model parameters, although the scheme can also be applied when this does not hold by replacing M with $\lceil M \rceil$. Next, we use $\mathcal{L}$ to denote the lattice, which is the set of points in $\mathcal{R}^L$ that can be written as an integer linear combination of the columns of $\boldsymbol{G}$, i.e., the set of all points $\boldsymbol{x} \in \mathcal{R}^L$ which can be written as $\boldsymbol{Gl}$ with $\boldsymbol{l}$ having integer entries:

$$\mathcal{L} \triangleq \{\boldsymbol{x} = \boldsymbol{Gl} : \boldsymbol{l} \in \mathcal{Z}^L\}. \tag{6}$$

A lattice quantizer $Q_{\mathcal{L}}(\cdot)$ maps each $\boldsymbol{x} \in \mathcal{R}^L$ to its nearest lattice point, i.e., $Q_{\mathcal{L}}(\boldsymbol{x}) = \boldsymbol{l}_{\boldsymbol{x}}$ where $\boldsymbol{l}_{\boldsymbol{x}} \in \mathcal{L}$ if $\|\boldsymbol{x} - \boldsymbol{l}_{\boldsymbol{x}}\| \leq \|\boldsymbol{x} - \boldsymbol{l}\|$ for every $\boldsymbol{l} \in \mathcal{L}$. Finally, let $\mathcal{P}_0$ be the basic lattice cell [24], i.e., the set of points $\boldsymbol{x} \in \mathcal{R}^L$ which are closer to $\boldsymbol{0}$ than to any other lattice point:

$$\mathcal{P}_0 \triangleq \{\boldsymbol{x} \in \mathcal{R}^L : \|\boldsymbol{x}\| < \|\boldsymbol{x} - \boldsymbol{p}\|, \forall \boldsymbol{p} \in \mathcal{L}/\{\boldsymbol{0}\}\}. \tag{7}$$

As $\mathcal{P}_0$ represents the set of points which are closer to the origin than to any other lattice point, its shape depends on the lattice $\mathcal{L}$, and in particular on the generator matrix $\boldsymbol{G}$. For instance, when $\boldsymbol{G} = \Delta \cdot \boldsymbol{I}_L$ for some $\Delta > 0$, then $\mathcal{L}$ is the square lattice, for which $\mathcal{P}_0$ is the set of vectors $\boldsymbol{x} \in \mathcal{R}^L$ whose ℓ_∞ norm is not larger than $\frac{\Delta}{2}$. In the two-dimensional case, such a generator matrix results in $\mathcal{P}_0$ being a square centered at the origin. For this setting, $Q_{\mathcal{L}}(\cdot)$ implements entry-wise scalar uniform quantization with spacing Δ [22, Ch. 23]. In general, the basic cell can take different shapes, such as hexagons for two-dimensional hexagonal lattices.

Using the above definitions in lattice quantization, we now present the encoding and decoding procedures of UVeQFed, which are based on subtractive dithered lattice quantization:

Encoder: The proposed encoding function $e_{t+\tau}(\cdot)$ implements dithered lattice quantization in four stages. It first normalizes the model updates and partitions it into sub-vectors of the lattice dimension, where the normalization is used to prevent overloading the finite lattice. Then, each vector is dithered before it is quantized to result in a distortion term which is not deterministically determined by the model updates, and is thus reduced by averaging. The quantized representation is compressed in lossless manner using entropy coding to further reduce its volume without inducing additional distortion. These steps are detailed in the following:

*E*1 **Normalize and partition:** The kth user scales $\boldsymbol{h}_{t+\tau}^{(k)}$ by $\zeta\|\boldsymbol{h}_{t+\tau}^{(k)}\|$ for some $\zeta > 0$, and divides the result into M distinct $L \times 1$ vectors, denoted $\{\bar{\boldsymbol{h}}_i^{(k)}\}_{i=1}^M$. The scalar quantity $\zeta\|\boldsymbol{h}_{t+\tau}^{(k)}\|$ is quantized separately from $\{\bar{\boldsymbol{h}}_i^{(k)}\}_{i=1}^M$ using some fine-resolution quantizer.

*E*2 **Dithering:** The encoder utilizes the source of common randomness, e.g., a shared seed, to generate the set of $L \times 1$ dither vectors $\{\boldsymbol{z}_i^{(k)}\}_{i=1}^M$, which are randomized in an i.i.d. fashion, independently of $\boldsymbol{h}_{t+\tau}^{(k)}$, from a uniform distribution over $\mathcal{P}_0$.

*E*3 **Quantization:** The vectors $\{\bar{\boldsymbol{h}}_i^{(k)}\}_{i=1}^M$ are discretized by adding the dither vectors and applying lattice quantization, i.e., by computing $\{Q_{\mathcal{L}}(\bar{\boldsymbol{h}}_i^{(k)} + \boldsymbol{z}_i^{(k)})\}$.

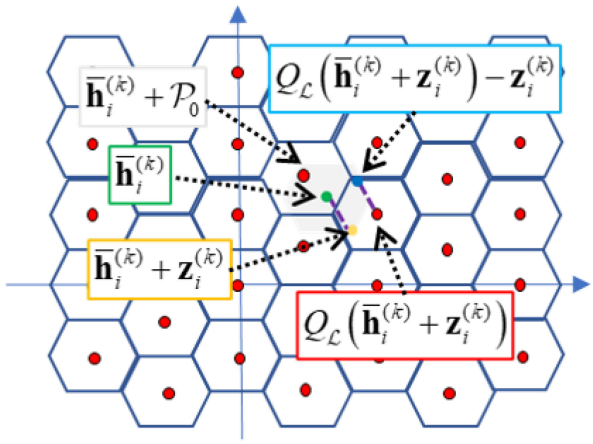


Fig. 2. Subtractive dithered lattice quantization illustration.

*E*4 **Entropy coding:** The discrete values $\{Q_{\mathcal{L}}(\bar{\boldsymbol{h}}_i^{(k)} + \boldsymbol{z}_i^{(k)})\}$ are encoded into a digital codeword $u_{t+\tau}^{(k)}$ in a lossless manner.

In order to utilize entropy coding in step *E*4, the discretized $\{Q_{\mathcal{L}}(\bar{\boldsymbol{h}}_i^{(k)} + \boldsymbol{z}_i^{(k)})\}$ must take values on a *finite set*. This is achieved by the normalization in Step *E*1, which guarantees that $\{\bar{\boldsymbol{h}}_i^{(k)}\}_{i=1}^M$ all reside inside the L-dimensional ball with radius ζ^{-1}, in which the number of lattice points is not larger than $\frac{\pi^{L/2}}{\zeta^L \Gamma(1+L/2)\det(\boldsymbol{G})}$ [34, Ch. 2], where $\Gamma(\cdot)$ is the Gamma function. The overhead in accurately quantizing the single scalar quantity $\zeta\|\boldsymbol{h}^{(k)}\|$ is typically negligible compared to the number of bits required to convey the set of vectors $\{\bar{\boldsymbol{h}}_i^{(k)}\}_{i=1}^M$, hardly affecting the overall quantization rate.

Decoder: The decoding mapping $d_{t+\tau}(\cdot)$ is comprised of four stages. The purpose of the first three steps is to invert the encoding procedure by decoding the lossless entropy code used in *E*4, subtracting the dither added in *E*2, and reforming the full model update vector from the partitioned sub-vectors generated in *E*1. The final stage uses the recovered model update to compute the aggregated global model. These stages are detailed in the following:

*D*1 **Entropy decoding:** The server first decodes each digital codeword $u_{t+\tau}^{(k)}$ into the discrete value $\{Q_{\mathcal{L}}(\bar{\boldsymbol{h}}_i^{(k)} + \boldsymbol{z}_i^{(k)})\}$. Since the encoding is carried out using a lossless source code, the discrete values are recovered without any errors.

*D*2 **Dither subtraction:** Using the source of common randomness, the server generates the dither vectors $\{\boldsymbol{z}_i^{(k)}\}$, which can be carried out rapidly and at low complexity using random number generators as the dither vectors obey a uniform distribution. The server then subtracts the corresponding vector from each lattice point, i.e., compute $\{Q_{\mathcal{L}}(\bar{\boldsymbol{h}}_i^{(k)} + \boldsymbol{z}_i^{(k)}) - \boldsymbol{z}_i^{(k)}\}$. An illustration of the subtractive dithered lattice quantization procedure is illustrated in Fig. 2.

*D*3 **Collecting and scaling:** The values $\{Q_{\mathcal{L}}(\boldsymbol{h}_i^{(k)} + \boldsymbol{z}_i^{(k)}) - \boldsymbol{z}_i^{(k)}\}$ are collected into an $m \times 1$ vector $\hat{\boldsymbol{h}}_{t+\tau}^{(k)}$ using the

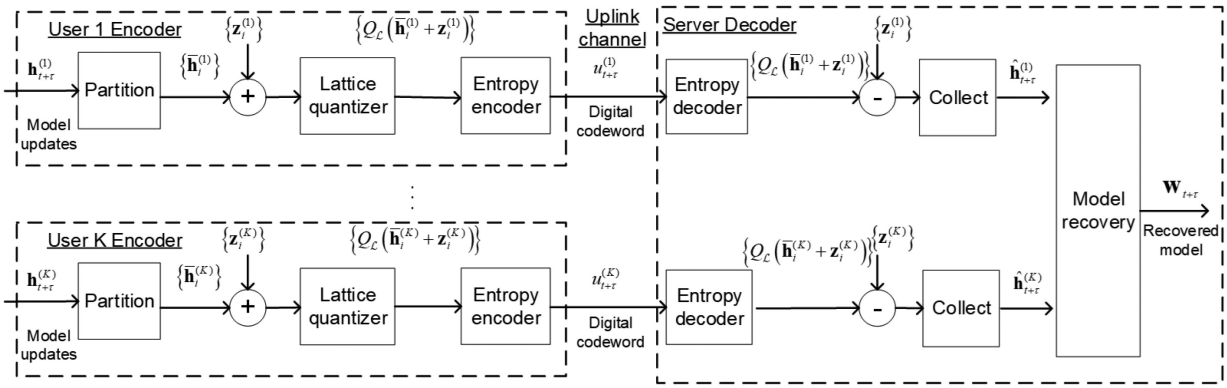


Fig. 3. UVeQFed encoding-decoding block diagram.

inverse operation of the partitioning and normalization in Step *E*1.

*D*4 **Model recovery:** The recovered matrices are combined into an updated model based on (4). Namely,

$$\boldsymbol{w}_{t+\tau} = \boldsymbol{w}_t + \sum_{k=1}^{K} \alpha_k \hat{\boldsymbol{h}}_{t+\tau}^{(k)}. \tag{8}$$

A block diagram of the proposed scheme is depicted in Fig. 3. The usage of subtractive dithered lattice quantization in Steps *E*2–*E*3 and *D*2 allow obtaining a digital representation which is relatively close to the true quantity, as illustrated in Fig. 2, without relying on prior knowledge of its distribution. The joint decoding aspect of the proposed scheme is introduced in the final model recovery Step *D*4. The remaining encoding-decoding procedure, i.e., Steps *E*1–*D*3 is carried out independently for each user.

B. Discussion

UVeQFed has several clear advantages. While it is based on information theoretic arguments, the resulting architecture is rather simple to implement. Both subtractive dithered quantization as well as entropy coding are concrete and established methods which can be realized with relatively low complexity and feasible hardware requirements. In particular, the main novel aspect of UVeQFed, i.e., the usage of subtractive dithered lattice quantization, first requires generating the dither signal via, e.g., the methods discussed in [35] for randomizing uniformly distributed random vectors. Then, the encoder carries out lattice projection of each sub-vector which, for finite and small L as used in the numerical study in Section V, involves a complexity term which only grows linearly with the number of parameters m. This resulting additional complexity is of the same order as previous quantized FL strategies, e.g., QSGD [17] which also uses entropy coding, and is typically dominated by the computational burden involved in training a deep model with m parameters. The source of common randomness needed for generating the dither vectors can be obtained by sharing a common seed between the server and users, as also assumed in [12]. The statistical characterization of the quantization error of such quantizers does not depend on the distribution of the model updates. This analytical tractability allows us to rigorously show that its combination with federated averaging mitigates the quantization error in Section IV. A similar approach was also used in the analysis of probabilistic quantization schemes for average consensus problems [36].

As the updates under the considered system model are quantized for a specific task, i.e., to obtain the global model by averaging, FL with bit constraints can be treated as a task-based quantization scenario [37]–[40]. In UVeQFed, this task is accounted for in the selection of the quantization scheme, using one for which the distortion vanishes by averaging regardless of the values of $\{\boldsymbol{h}^{(k)}\}$. In addition, UVeQFed is derived based on modeling the uplink channel as a bit-limited pipeline. While this model is widely adopted in the FL literature, it may not accurately reflect the true nature of wireless communication channels, being a noisy and shared media. This property of wireless communications is known to affect the design of deep learning systems operating over such channels [41]–[47]. Traditionally, transmitted bit streams such as the codewords produced by UVeQFed are protected using a separate channel code for mitigating the errors induced by the noisy channel. Alternatively, one can also consider extending UVeQFed to directly map the model updates into a channel codeword as a form of task-based joint source channel coding. We leave this extension of UVeQFed for future work.

The encoding Steps *E*1–*E*3 can be viewed as a generalization of probabilistic scalar quantizers, used in, e.g., QSGD [17]. When the lattice dimension is $L = 1$ and $\zeta = 1$, Steps *E*1–*E*3 implement the same encoding as QSGD. However, the decoder is not the same as in QSGD due to the dither subtraction in Step *D*2, which is known to reduce the distortion and yield an error term that does not depend on the model updates [30]. Furthermore, UVeQFed allows using vector quantizers, i.e., setting $L > 1$, which is known to further improve the quantization accuracy [24]. Specifically, the usage of vector quantizers allows UVeQFed to combine dimensionality reduction methods with quantization schemes by jointly mapping sets of samples into discrete representations. The gains of subtracting the dither at the decoder and of using vector quantizers over scalar ones are numerically demonstrated in our experimental study in Section V.

The usage of lossless source coding in Steps *E*4 and *D*1 allows exploiting the typically non-uniform distribution of the quantizer outputs. A similar approach was also used in QSGD [17],

where Elias codes were utilized. Since Steps *E*4 and *D*1 involve multiple encoders and a single decoder, improved compression can be achieved by utilizing distributed source coding methods, e.g., Slepian-Wolf coding [48, Ch. 15.4]. In such cases, the server decodes the received codewords $\{u_{t+\tau}^{(k)}\}$ into $\{Q_{\mathcal{L}}(\bar{\boldsymbol{h}}_i^{(k)}+\boldsymbol{z}_i^{(k)})\}$ in a joint manner, instead of decoding each $Q_{\mathcal{L}}(\bar{\boldsymbol{h}}_i^{(k)}+\boldsymbol{z}_i^{(k)})$ from its corresponding $u_{t+\tau}^{(k)}$ separately. Similarly, the distributed nature of FL can be exploited to optimize the reconstruction fideitly for a given bit budget using Wyner-Ziv coding [49]. However, such distributed coding schemes typically require a-priori knowledge of the joint distribution of $\{\bar{\boldsymbol{h}}_i^{(k)}\}$, and utilize different encoding mappings for each user, thus not meeting requirements *A*1-*A*2.

Finally, we note that the FL performance is affected by the selection of the lattice $\mathcal{L}$ and the coefficient ζ. In general, lattices of higher dimensions typically result in more accurate representations, at the cost of increased complexity. Methods for designing the lattice generator matrix $\boldsymbol{G}$ can be found in [50]. The coefficient ζ should be set to allow the usage of a limited number of lattice points, which is translated into less bits, without concentrating the resulting vectors such that they become indistinguishable after quantization. For example, using $\zeta=1$ results in most quantized values mapped to zero, as also observed in [17]. A reasonable setting is $\zeta=3\frac{1}{\sqrt{M}}$, resulting in $\zeta\|\boldsymbol{h}^{(k)}\|$ approaching 3 times the standard deviation of the quantized vectors when they are zero-mean and i.i.d., and thus assuring that they reside inside the unit L-ball with probability of over 88% by Chebyshev's inequality [51].

IV. Performance Analysis

Next, we analyze the performance of UVeQFed, characterizing its distortion and studying its convergence properties. We consider the conventional local SGD training method, detailed in Section IV-A, and characterize the resulting distortion of UVeQFed and the convergence of the global model in Sections IV-B-IV-C, respectively.

A. Local SGD

Local SGD is arguably the most common training method used for federated averaging [52]. Here, each user updates the weights using τ SGD iterations before sending the updated model to the server for aggregation. Let $\boldsymbol{\epsilon}_t^{(k)}$ denote the error induced in quantizing the model update $\boldsymbol{h}_t^{(k)}$, and let $i_t^{(k)}$ be the sample index chosen uniformly from the local data of the kth user at time t. By defining the gradient computed at a single sample of index i as $\nabla F_k^i(\tilde{\boldsymbol{w}}) \triangleq \nabla F_k(\tilde{\boldsymbol{w}};(\boldsymbol{x}_i^{(k)},\boldsymbol{y}_i^{(k)}))$, the local weights at the kth user, denoted $\tilde{\boldsymbol{w}}_t^{(k)}$, are updated via:

$$\tilde{\boldsymbol{w}}_{t+1}^{(k)}=\begin{cases}\tilde{\boldsymbol{w}}_t^{(k)}-\eta_t\nabla F_k^{i_t^{(k)}}\left(\tilde{\boldsymbol{w}}_t^{(k)}\right), & t+1\notin\mathcal{T}_\tau,\\ \sum_{k'=1}^{K}\alpha_{k'}\left(\tilde{\boldsymbol{w}}_t^{(k')}-\eta_t\nabla F_{k'}^{i_t^{(k')}}\left(\tilde{\boldsymbol{w}}_t^{(k')}\right)\right.\\ \qquad\left.+\boldsymbol{\epsilon}_{t+1}^{(k')}\right), & t+1\in\mathcal{T}_\tau,\end{cases}\tag{9}$$

where η_t is the learning rate at time instance t, and $\mathcal{T}_\tau$ is the set of positive integer multiples of τ.

We focus on the case in which the users compute a single stochastic gradient in each time instance. Hence, the performance in terms of convergence rate can be further improved by using mini-batches [52], i.e., replacing the random index $i_t^{(k)}$ with a set of random indices. The fact that the model updates are quantized when conveyed to the server is encapsulated in the per-user model update quantization error $\boldsymbol{\epsilon}_t^{(k)}$.

B. Quantization Error Bound

The need to represent the model updates $\boldsymbol{h}_{t+\tau}^{(k)}$ using a finite number of bits inherently induces some distortion, i.e., the recovered vector is $\hat{\boldsymbol{h}}_{t+\tau}^{(k)}=\boldsymbol{h}_{t+\tau}^{(k)}+\boldsymbol{\epsilon}_{t+\tau}^{(k)}$. The error in representing $\zeta\|\boldsymbol{h}_{t+\tau}^{(k)}\|$ is assumed to be negligible. For example, the normalized quantization error is of the order of 10^{-7} for 12 b quantization of a scalar value, and decreases exponentially with each additional bit [22, Ch. 23]. Letting $\bar{\sigma}_{\mathcal{L}}^2$ be the normalized second order lattice moment, defined as $\bar{\sigma}_{\mathcal{L}}^2\triangleq\int_{\mathcal{P}_0}\|\boldsymbol{x}\|^2d\boldsymbol{x}/\int_{\mathcal{P}_0}d\boldsymbol{x}$ [53], the moments of the quantization error satisfy the following:

Theorem 1: The quantization error vector $\boldsymbol{\epsilon}_{t+\tau}^{(k)}$ has zero-mean entries and satisfies

$$\mathbb{E}\left\{\left\|\boldsymbol{\epsilon}_{t+\tau}^{(k)}\right\|^2\middle|\boldsymbol{h}_{t+\tau}^{(k)}\right\}=\zeta^2\|\boldsymbol{h}_{t+\tau}^{(k)}\|^2\,M\bar{\sigma}_{\mathcal{L}}^2.\tag{10}$$

Proof: See Appendix A.

Theorem 1 characterizes the distortion in quantizing the model updates using UVeQFed. Unlike the corresponding characterization of previous quantizers used in FL which obtained an upper bound on the quantization error, e.g., [17, Lem. 1], the dependence of the expected error norm on the number of bits is not explicit in (10), but rather encapsulated in the lattice moment $\bar{\sigma}_{\mathcal{L}}^2$. To observe that (10) indeed represents lower distortion compared to previous FL quantization schemes, we note that even when scalar quantizers are used, i.e., $L=1$ for which $\frac{1}{L}\bar{\sigma}_{\mathcal{L}}^2$ is known to be largest [53], the resulting quantization is reduced by a factor of 2 compared to conventional probabilistic scalar quantizers, such as QSGD, due to the subtraction of the dither upon decoding in Step *D*2 [30, Thms. 1-2].

The model updates are recovered in order to update the global model via $\boldsymbol{w}_{t+\tau}=\sum\alpha_k\tilde{\boldsymbol{w}}_{t+\tau}^{(k)}$ at the server. We next show that the statistical characterization of the distortion in Theorem 1 contributes to the accuracy in recovering the desired $\boldsymbol{w}_{t+\tau}^{\mathrm{des}}$ (5) via $\boldsymbol{w}_{t+\tau}$. To that aim, we introduce the following assumption on the stochastic gradients, which is often employed in distributed learning studies [25], [52], [54]:

*AS*1 The expected squared ℓ_2 norm of the random vector $\nabla F_k^i(\boldsymbol{w})$, representing the stochastic gradient evaluated at $\boldsymbol{w}$, is bounded by some $\xi_k^2>0$ for all $\boldsymbol{w}\in\mathcal{R}^m$.

We can now bound the distance between the desired model $\boldsymbol{w}_{t+\tau}^{\mathrm{des}}$ and the recovered one $\boldsymbol{w}_{t+\tau}$, as stated in the following theorem:

Theorem 2: When *AS*1 holds, the mean-squared distance between $\boldsymbol{w}_{t+\tau}$ and $\boldsymbol{w}_{t+\tau}^{\mathrm{des}}$ satisfies

$$\mathbb{E}\left\{\left\|\boldsymbol{w}_{t+\tau}-\boldsymbol{w}_{t+\tau}^{\mathrm{des}}\right\|^2\right\}\leq M\zeta^2\bar{\sigma}_{\mathcal{L}}^2\tau\left(\sum_{t'=t}^{t+\tau-1}\eta_{t'}^2\right)\sum_{k=1}^{K}\alpha_k^2\xi_k^2. \quad (11)$$

Proof: See Appendix B.

Theorem 2 implies that the recovered model can be made arbitrarily close to the desired one by increasing K, namely, the number of users. For example, when $\alpha_k = 1/K$, i.e., conventional averaging, it follows from Theorem 2 that the mean-squared error in the weights decreases as $1/K$. In particular, if $\max_k \alpha_k$ decreases with K, which essentially means that the updated model is not based only on a small part of the participating users, then the distortion vanishes in the aggregation process. Furthermore, when the step size η_t gradually decreases, which is known to contribute to the convergence of FL [25], it follows from Theorem 2 that the distortion decreases accordingly, further mitigating its effect as the FL iterations progress. As shown in Appendix B, the bound in (11) is obtained by exploiting the mutual independence of the subtractive dithered quantization error and the quantized value. Hence, our ability to rigorously upper bound the distance in Theorem 2 is a direct consequence of this universal method.

C. FL Convergence Analysis

We next study the convergence of FL with UVeQFed. Our analysis is carried out under the following assumptions, commonly used in FL convergence studies [25], [52]:

AS2 The local objective functions $\{F_k(\cdot)\}$ are all ρ_s-smooth, namely, for all $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathcal{R}^m$ it holds that

$$F_k(\boldsymbol{v}_1) - F_k(\boldsymbol{v}_2) \leq (\boldsymbol{v}_1-\boldsymbol{v}_2)^T\nabla F_k(\boldsymbol{v}_2) + \frac{1}{2}\rho_s\|\boldsymbol{v}_1-\boldsymbol{v}_2\|^2.$$

AS3 The local objective functions $\{F_k(\cdot)\}$ are all ρ_c-strongly convex, namely, for all $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathcal{R}^m$ it holds that

$$F_k(\boldsymbol{v}_1) - F_k(\boldsymbol{v}_2) \geq (\boldsymbol{v}_1-\boldsymbol{v}_2)^T\nabla F_k(\boldsymbol{v}_2) + \frac{1}{2}\rho_c\|\boldsymbol{v}_1-\boldsymbol{v}_2\|^2.$$

Assumptions *AS*1–*AS*2 are commonly used in FL convergence studies [25], [52], and hold for a broad range of objective functions used in FL systems, including ℓ_2-norm regularized linear regression and logistic regression [25].

We do not restrict the labeled data of each of the users to be generated from an identical distribution, i.e., we consider a statistically heterogeneous scenario, thus faithfully representing FL setups [7], [9]. Such heterogeneity is in line with assumption *A*2, which does not impose any specific distribution structure on the underlying statistics of the training data. Following [25], we define the heterogeneity gap,

$$\psi \triangleq F(\boldsymbol{w}^{\mathrm{o}}) - \sum_{k=1}^{K}\alpha_k \min_{\boldsymbol{w}} F_k(\boldsymbol{w}). \quad (12)$$

The value of ψ quantifies the degree of heterogeneity. If the training data originates from the same distribution, then ψ tends to zero as the training size grows. However, for heterogeneous data, its value is positive. The convergence of UVeQFed with federated averaging is characterized in the following theorem:

Theorem 3: Set $\gamma = \tau\max(1, 4\rho_s/\rho_c)$ and consider a UVeQFed setup satisfying *AS*1–*AS*3. Under this setting, local SGD with step size $\eta_t = \frac{\tau}{\rho_c(t+\gamma)}$ for each $t \in \mathcal{N}$ satisfies

$$\mathbb{E}\{F(\boldsymbol{w}_t)\} - F(\boldsymbol{w}^{\mathrm{o}}) \leq \frac{\rho_s}{2(t+\gamma)}\max\left(\frac{\rho_c^2+\tau^2 b}{\tau\rho_c}, \gamma\|\boldsymbol{w}_0-\boldsymbol{w}^{\mathrm{o}}\|^2\right), \quad (13)$$

where

$$b \triangleq \left(1+4M\zeta^2\bar{\sigma}_{\mathcal{L}}^2\tau^2\right)\sum_{k=1}^{K}\alpha_k^2\xi_k^2 + 6\rho_s\psi + 8(\tau-1)^2\sum_{k=1}^{K}\alpha_k\xi_k^2.$$

Proof: See Appendix C.

Theorem 3 implies that UVeQFed with local SGD, i.e., conventional federated averaging, converges at a rate of $\mathcal{O}(1/t)$. The physical meaning of this asymptotic convergence rate is that as the number of iterations t progresses, the learned model converges to the optimal one with a difference decaying as $1/t$. Specifically, the difference between the objective of the model learned in a federated manner and the optimal objective decays to zero at least as quickly as $1/t$ (up to some constant). This is the same order of convergence as FL without quantization constraints for i.i.d. [52] as well as heterogeneous data [25], [55]. Nonetheless, it is noted that the need to quantize the model updates yield an additive term in the coefficient b which grows with the number of parameter via M. This term adds to the linear dependence of b on the bound on the gradients norm ξ^2, which is expected to grow with the number of parameters, and also appears in the corresponding bounds for local SGD without quantization constraints. This implies that FL typically converges slower for larger models, i.e., the larger the dimensionality of the model updates which have to be quantized. A similar order of convergence was also reported for previous probabilistic quantization schemes which typically considered i.i.d. data, e.g., [17, Thm. 3.4].

While it is difficult to identify the convergence gains of UVeQFed over previously proposed FL quantizers, such as QSGD, by comparing Theorem 3 to their corresponding convergence bounds, in Section V we empirically demonstrate that UVeQFed converges to more accurate global models compared to FL with probabilistic scalar quantizers, when trained using i.i.d. as well as heterogeneous data sets. Additionally, we note that the communication load on the uplink channel induced by UVeQFed can be further reduced by allowing only part of the nodes to participate in each set of iterations [10], [19]. We leave the analysis of UVeQFed with partial node participation for future work.

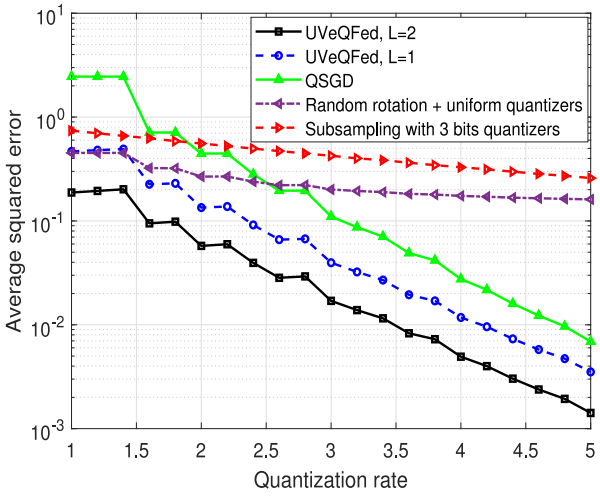


Fig. 4. Quantization distortion, i.i.d. data.

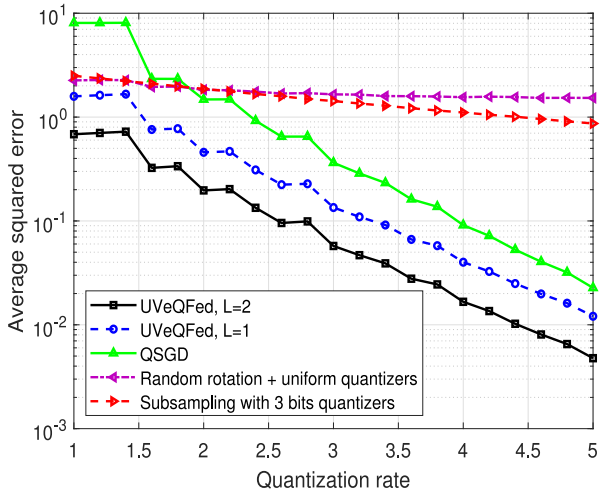


Fig. 5. Quantization distortion, correlated data.

V. Numerical Evaluations

In this section we numerically evaluate UVeQFed. We first compare the quantization error induced by UVeQFed to competing methods utilized in FL in Section V-A. Then, we numerically demonstrate how the reduced distortion is translated in FL performance gains using both MNIST and CIFAR-10 data sets.[1] in Section V-B.

A. Quantization Error

We begin by focusing only on the compression method, studying its accuracy using synthetic data. We evaluate the distortion induced in quantization of UVeQFed operating with a two-dimensional hexagonial lattice, i.e., $L = 2$ and $\boldsymbol{G} = [2, 0; 1, 1/\sqrt{3}]$ [33], as well as with scalar quantizers, namely, $L = 1$ and $\boldsymbol{G} = 1$. The normalization coefficient is set to $\zeta = \frac{2+R/5}{\sqrt{M}}$. As discussed in Subsection III-B, decreasing ζ results in higher overload probability, i.e., having more quantized sub-vectors lying outside the unit L-ball. Consequently, the setting used here results in having the quantized sub-vectors being spread in a more uniform manner inside the unit ball at lower quantization rates, where the lattice points are more distant from one another compared to higher quantization rates, at the cost of increased overload probability, thus balancing the overall distortion. The distortion of UVeQFed is compared to QSGD [17], as well as to uniform quantizers with random unitary rotation [12], and to subsampling by random masks followed by uniform three-bit quantizers [12], all operating with the same quantization rate, i.e., the same overall number of bits.

Let $\boldsymbol{H}$ be a 128×128 matrix with Gaussian i.i.d. entries, and let $\boldsymbol{\Sigma}$ be a 128×128 matrix whose entries are given by $(\boldsymbol{\Sigma})_{i,j} = e^{-0.2|i-j|}$, representing an exponentially decaying correlation. In Figs. 4-5 we depict the per-entry squared-error in quantizing $\boldsymbol{H}$ and $\boldsymbol{\Sigma}\boldsymbol{H}\boldsymbol{\Sigma}^T$, representing independent and correlated data, respectively, versus the quantization rate R, defined as the ratio of the number of bits to the number of entries of $\boldsymbol{H}$.

[1]The source code used in the numerical evaluations detailed in this section is available online at https://github.com/mzchen0/UVeQFed

TABLE I
Main Simulation Parameters

	MNIST		CIFAR-10
Users K	100	15	10
Samples n_k	500	1000	5000
Model	Two-layer fully connected		Five-layer convolutional [56]
Optimizer	Gradient descent		Mini-batch SGD
Local steps τ	1		17
Step-size η_1	10^{-2}		$5 \cdot 10^{-3}$

The distortion is averaged over 100 independent realizations of $\boldsymbol{H}$. To meet the bit rate constraint when using lattice quantizers we scaled $\boldsymbol{G}$ such that the resulting codewords use less than $128^2 R$ bits. For the scalar quantizers and subsampling-based scheme, the rate determines the quantization resolution and the subsampling ratio, respectively.

We observe in Figs. 4-5 that UVeQFed achieves a more accurate digital representation compared to previously proposed methods. It is also observed that UVeQFed with vector quantization, outperforms its scalar counterpart, and that the gain is more notable when the quantized entries are correlated. This demonstrates the improved accuracy of jointly encoding multiple samples via vector quantization as well as its ability to exploit statistical correlation in a universal manner by using fixed lattice-based quantization regions which do not depend on the underlying distribution.

B. FL Convergence

Next, we demonstrate that the reduced distortion of UVeQFed also translates into FL performance gains. To that aim, we evaluate its application for training neural networks using the MNIST and CIFAR-10 data sets, and compare its performance to that achievable using previous quantization methods for FL. The simulation settings are detailed below, with the main parameters summarized in Table I.

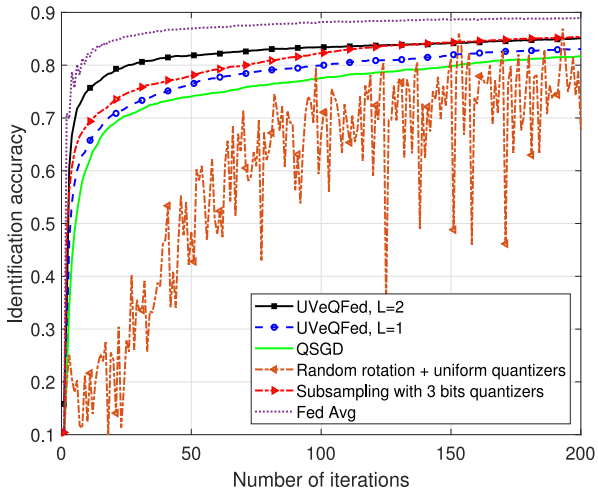


Fig. 6. Convergence profile, MNIST, $R = 2$, $K = 100$.

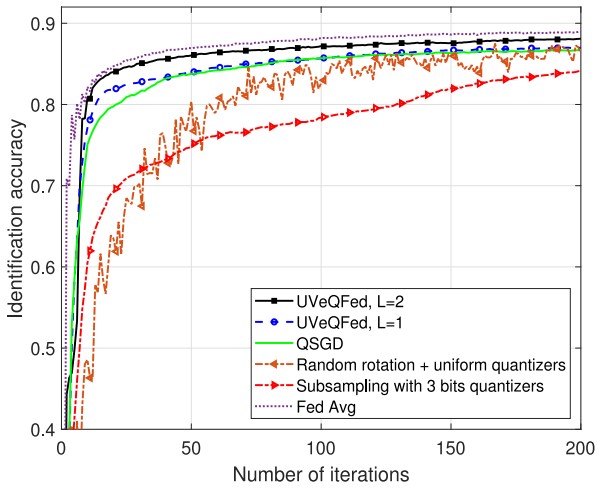


Fig. 7. Convergence profile, MNIST, $R = 4$, $K = 100$.

We first compare the accuracy of models trained using UVeQFed to those obtained using federated averaging combined with the quantization methods considered in Subsection V-A, i.e., QSGD [17] and the schemes proposed in [12] of uniform quantizers with random rotation as well as random subsampling followed by three-bit uniform quantizers. To that aim, we train a fully-connected network with a single hidden layer of 50 neurons and an intermediate sigmoid activation for detecting handwritten digits based on the MNIST data set. Training is carried out using $K = 100$ users, each has access to 500 training samples distributed in an i.i.d. fashion, such that each user has an identical number of images from each label. The users update their weights using gradient descent, where federated averaging is carried out on each iteration. The resulting accuracy versus the number of iterations of these quantized FL schemes compared to federated averaging without quantization is depicted in Figs. 6-7 for quantization rates $R = 2$ and $R = 4$, respectively.

Observing Figs. 6-7, we note that UVeQFed with vector quantization, i.e., $L = 2$, achieves the most rapid and accurate convergence among all considered schemes. In particular, for $R = 4$, UVeQFed with $L = 2$ achieves a convergence profile within

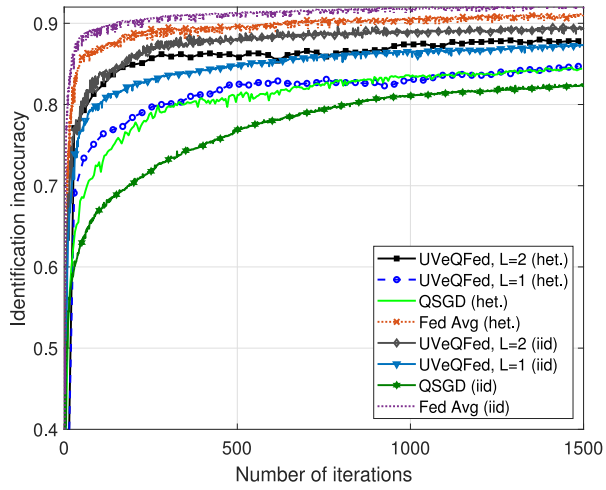


Fig. 8. Convergence profile, MNIST, $R = 2$, $K = 15$.

a minor gap from federated averaging without quantization constraints. Among the previous schemes, QSGD demonstrates steady accuracy improvements, though it is still outperformed by UVeQFed with $L = 1$, indicating that the reduced distortion achieved by using subtractive dithering is translated into improved trained models. The quantization methods proposed in [12] result in notable variations in the trained model accuracy and in slower convergence due to their increased error induced in quantization, as noted in Subsection V-A.

We next evaluate UVeQFed for both heterogeneous as well as i.i.d. distributions of the training data. Based on the results observed in Figs. 6–7 and to avoid cluttering, we compare UVeQFed only to QSGD and to the accuracy achieved using federated averaging without quantization. Here, we train neural classifiers for both the MNIST and the CIFAR-10 data sets, where for each data set we use both heterogeneous and i.i.d. division of the data.

For MNIST, we again use a fully-connected network with a single hidden layer of 50 neurons and an intermediate sigmoid activation with gradient descent optimization. Each of the $K = 15$ users has 1000 training samples. We consider the case where the samples are distributed sequentially among the users, i.e., the first user has the first 1000 samples in the data set, and so on, resulting in an uneven heterogeneous division of the labels of the users. We also train using an i.i.d. data division, where the labels are uniformly distributed among the users. The resulting accuracy versus the number of iterations is depicted in Figs. 8-9 for quantization rates $R = 2$ and $R = 4$, respectively.

For CIFAR-10, we train the deep convolutional neural network architecture used in [56], whose trainable parameters constitute three convolution layers and two fully-connected layers. Here, we consider two methods for distributing the 50000 training images of CIFAR-10 among the $K = 10$ users: An i.i.d. division, where each user has the same number samples from each of the 10 labels, and a heterogeneous division, in which at least 25% of the samples of each user correspond to a single distinct label. Each user completes a single epoch of SGD with mini-batch size 60 before the models are aggregated. The resulting accuracy versus the number of epochs is

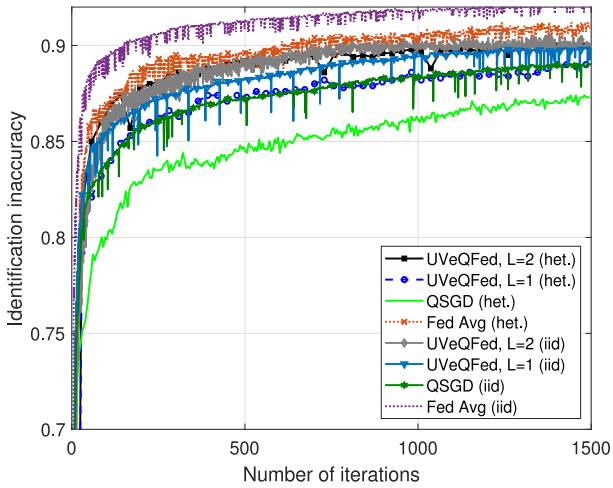


Fig. 9. Convergence profile, MNIST, $R = 4$, $K = 15$.

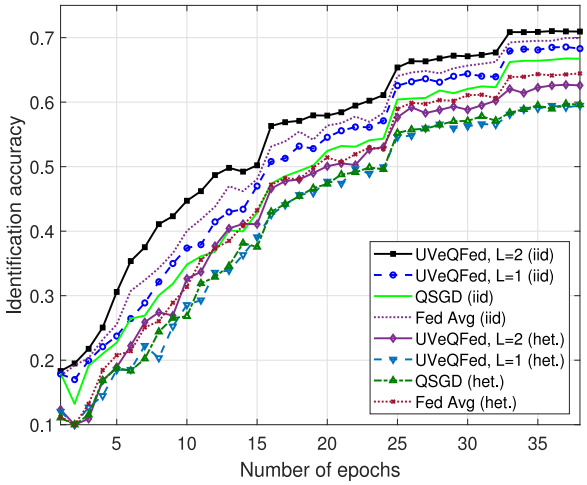


Fig. 10. Convergence profile, CIFAR-10, $R = 2$.

depicted in Figs. 10–11 for quantization rates $R = 2$ and $R = 4$, respectively.

We observe in Figs. 8–11 that UVeQFed with vector quantizer, i.e., $L = 2$, results in convergence to the most accurate model for all the considered scenarios. In fact, when training a deep convolutional network, for which the loss surface is extremely complex and non-convex, we observe in Figs. 10-11 that UVeQFed with $L = 2$ trained using i.i.d. data achieves improved accuracy over federated averaging without quantization. This follows from the fact that the stochastic nature of the quantization error in UVeQFed results in its implementing a noisy variant of local SGD, which is known to be capable of boosting convergence and avoid local minimas when training deep neural networks with non-convex loss surfaces [57], as also observed in [42].

The observed gains are more dominant for $R = 2$, implying that the usage of UVeQFed with multi-dimensional lattices can notably improve the performance over low rate channels. Particularly, we observe in Figs. 8-11 that similar gains of UVeQFed are noted for both i.i.d. as well as heterogeneous setups, while the heterogeneous division of the data degrades the accuracy of all considered schemes compared to the i.i.d division. It is

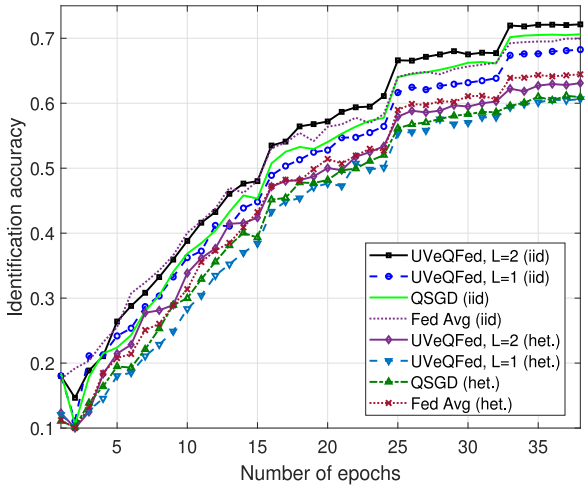


Fig. 11. Convergence profile, CIFAR-10, $R = 4$.

also observed that UVeQFed with scalar quantizers, i.e., $L = 1$, achieves improved convergence compared to QSGD for most considered setups, which stems from its reduced distortion.

The results presented in this section demonstrate that the theoretical benefits of UVeQFed, which rigorously hold under *AS*1–*AS*2, translate into improved convergence when operating under rate constraints with non-synthetic data.

VI. Conclusion

In this work we have proposed UVeQFed, which utilizes universal vector quantization methods to mitigate the effect of limited communication in FL. We first identified the specific requirements from quantization schemes used in FL setups. Then, we proposed an encoding-decoding strategy based on dithered lattice quantization. We analyzed UVeQFed, proving that its error term is mitigated by federated averaging. We also characterized its convergence profile, showing that its asymptotic decay rate is the same an unquantized local SGD. Our numerical study demonstrates that UVeQFed allows achieving more accurate recovery of model updates in each FL iteration compared to previously proposed schemes for the same number of bits, and that its reduced distortion is translated into improved convergence with the non-synthetic MNIST and CIFAR-10 data sets.

Appendix

A. Proof of Theorem 1

To prove the theorem, we note that by decoding step *D*3, the error vector $\boldsymbol{\epsilon}_{t+\tau}^{(k)}$ scaled by $\zeta\|\boldsymbol{h}_{t+\tau}^{(k)}\|$, consists of M vectors $\{\bar{\boldsymbol{\epsilon}}_i^{(k)}\}$. Each $\bar{\boldsymbol{\epsilon}}_i^{(k)}$ is an $L \times 1$ vector representing the ith subtractive dithered quantization error, defined as $\bar{\boldsymbol{\epsilon}}_i^{(k)} \triangleq Q_{\mathcal{L}}(\bar{\boldsymbol{h}}_i^{(k)} + \boldsymbol{z}_i^{(k)}) - \boldsymbol{z}_i^{(k)} - \bar{\boldsymbol{h}}_i^{(k)}$. The fact that we have used subtractive dithered quantization via encoding steps *E*2–*E*3 and decoding step *D*2, implies that, regardless of the statistical model of $\{\bar{\boldsymbol{h}}_i^{(k)}\}$, the quantization error vectors $\{\bar{\boldsymbol{\epsilon}}_i^{(k)}\}$ are zero-mean, i.i.d (over both i and k), and uniformly distributed over $\mathcal{P}_0$ [24].

Consequently,

$$\mathbb{E}\left\{\left\|\boldsymbol{\epsilon}_{t+\tau}^{(k)}\right\|^2\left|\boldsymbol{h}_{t+\tau}^{(k)}\right.\right\} = \zeta^2\|\boldsymbol{h}_{t+\tau}^{(k)}\|^2\sum_{i=1}^{M}\mathbb{E}\left\{\left\|\bar{\boldsymbol{\epsilon}}_i^{(k)}\right\|^2\right\} = \zeta^2\|\boldsymbol{h}_{t+\tau}^{(k)}\|^2 M\bar{\sigma}_{\mathcal{L}}^2,$$

thus proving the theorem. □

B. Proof of Theorem 2

To prove the the theorem, we first express the the updated global model using as a sum of the desired global model and the quantization noise. Then, we show that the distance between $\boldsymbol{w}_{t+\tau}$ and $\boldsymbol{w}_{t+\tau}^{\text{des}}$ can be bounded via (11) due to the statistical properties of subtractive dithered quantization error [24]. To formulate this distance between $\boldsymbol{w}_{t+\tau}$ and the desired $\boldsymbol{w}_{t+\tau}^{\text{des}}$, we use $\{\bar{\boldsymbol{w}}_{t,i}\}_{i=1}^M$, $\{\bar{\boldsymbol{w}}_{t+\tau,i}\}_{i=1}^M$, and $\{\bar{\boldsymbol{w}}_{t+\tau,i}^{\text{des}}\}_{i=1}^M$ to denote the partitions of $\boldsymbol{w}_t$, $\boldsymbol{w}_{t+\tau}$, and $\boldsymbol{w}_{t+\tau}^{\text{des}}$ into M distinct $L\times 1$ vectors, as done in Step *E*1. To formulate this distance, we use $\{\bar{\boldsymbol{w}}_{t,i}\}_{i=1}^M$ to denote the partition of $\boldsymbol{w}_t$ into M distinct $L\times 1$ vectors via step *E*1, similarly to the definitions of $\{\bar{\boldsymbol{w}}_{t+\tau,i}\}$ and $\{\bar{\boldsymbol{w}}_{t+\tau,i}^{\text{des}}\}$.

From the decoding and model recovery steps *D*3–*D*4 it follows that

$$\begin{aligned}\bar{\boldsymbol{w}}_{t+\tau,i} &= \bar{\boldsymbol{w}}_{t,i} + \sum_{k=1}^{K}\alpha_k\zeta\|\boldsymbol{h}_{t+\tau}^{(k)}\|\left(Q_{\mathcal{L}}(\bar{\boldsymbol{h}}_i^{(k)}+\boldsymbol{z}_i^{(k)})-\boldsymbol{z}_i^{(k)}\right)\\ &= \bar{\boldsymbol{w}}_{t,i} + \sum_{k=1}^{K}\alpha_k\zeta\|\boldsymbol{h}_{t+\tau}^{(k)}\|\bar{\boldsymbol{h}}_i^{(k)} + \sum_{k=1}^{K}\alpha_k\zeta\|\boldsymbol{h}_{t+\tau}^{(k)}\|\bar{\boldsymbol{\epsilon}}_i^{(k)},\end{aligned} \tag{B.1}$$

where $\bar{\boldsymbol{\epsilon}}_i^{(k)}$ is the subtractive dithered quantization error, defined in Appendix A. Now, since $\boldsymbol{h}_{t+\tau}^{(k)} = \tilde{\boldsymbol{w}}_{t+\tau}^{(k)} - \boldsymbol{w}_t$ combined with (5) and the fact that $\sum_{k=1}^K\alpha_k = 1$, it holds that $\bar{\boldsymbol{w}}_{t,i} + \sum_{k=1}^K\alpha_k\zeta\|\boldsymbol{h}_{t+\tau}^{(k)}\|\bar{\boldsymbol{h}}_i^{(k)} = \bar{\boldsymbol{w}}_{t+\tau,i}^{\text{des}}$. Substituting this into (B.1) yields

$$\bar{\boldsymbol{w}}_{t+\tau,i} - \bar{\boldsymbol{w}}_{t+\tau,i}^{\text{des}} = \sum_{k=1}^{K}\alpha_k\zeta\|\boldsymbol{h}_{t+\tau}^{(k)}\|\bar{\boldsymbol{\epsilon}}_i^{(k)}. \tag{B.2}$$

As discussed in Appendix A, $\{\bar{\boldsymbol{\epsilon}}_i^{(k)}\}$ are zero-mean, i.i.d (over both i and k), and independent of $\boldsymbol{h}_{t+\tau}^{(k)}$. Consequently, by the law of total expectation

$$\begin{aligned}\mathbb{E}\left\{\left\|\boldsymbol{w}_{t+\tau}-\boldsymbol{w}_{t+\tau}^{\text{des}}\right\|^2\right\} &= \mathbb{E}\left\{\left\|\sum_{i=1}^{M}\sum_{k=1}^{K}\alpha_k\zeta\|\boldsymbol{h}_{t+\tau}^{(k)}\|\bar{\boldsymbol{\epsilon}}_i^{(k)}\right\|^2\right\}\\ &\overset{(a)}{=} \mathbb{E}\left\{\mathbb{E}\left\{\left\|\sum_{i=1}^{M}\sum_{k=1}^{K}\alpha_k\zeta\|\boldsymbol{h}_{t+\tau}^{(k)}\|\bar{\boldsymbol{\epsilon}}_i^{(k)}\right\|^2\left|\boldsymbol{h}_{t+\tau}^{(k)}\right.\right\}\right\}\\ &\overset{(b)}{=} \mathbb{E}\left\{M\sum_{k=1}^{K}\alpha_k^2\zeta^2\bar{\sigma}_{\mathcal{L}}^2\|\boldsymbol{h}_{t+\tau}^{(k)}\|^2\right\}.\end{aligned} \tag{B.3}$$

where (a) follows from the law of total expectation, and (b) holds by (10).

Next, we note that by (9), the model update $\boldsymbol{h}_{t+\tau}^{(k)} = \tilde{\boldsymbol{w}}_{t+\tau}^{(k)} - \tilde{\boldsymbol{w}}_t^{(k)}$ can be written as the sum of the stochastic gradients $\boldsymbol{h}_{t+\tau}^{(k)} = \sum_{t'=t}^{t+\tau-1}\eta_{t'}\nabla F_k^{i_{t'}^{(k)}}(\tilde{\boldsymbol{w}}_{t'}^{(k)})$. Since the indices $\{i_t^{(k)}\}$ are i.i.d. over t and k, applying the law of total expectation to (B.3) yields

$$\begin{aligned}&\mathbb{E}\left\{\left\|\boldsymbol{w}_{t+\tau}-\boldsymbol{w}_{t+\tau}^{\text{des}}\right\|^2\right\}\\ &=\mathbb{E}\left\{M\zeta^2\bar{\sigma}_{\mathcal{L}}^2\sum_{k=1}^{K}\alpha_k^2\mathbb{E}\left\{\|\boldsymbol{h}_{t+\tau}^{(k)}\|^2\middle|\{\tilde{\boldsymbol{w}}_{t'}^{(k)}\}\right\}\right\}\\ &=\mathbb{E}\left\{M\zeta^2\bar{\sigma}_{\mathcal{L}}^2\sum_{k=1}^{K}\alpha_k^2\mathbb{E}\left\{\left\|\sum_{t'=t}^{t+\tau-1}\eta_{t'}\nabla F_k^{i_{t'}^{(k)}}\left(\tilde{\boldsymbol{w}}_{t'}^{(k)}\right)\right\|^2\middle|\{\tilde{\boldsymbol{w}}_{t'}^{(k)}\}\right\}\right\}\\ &\overset{(a)}{\leq}\mathbb{E}\left\{M\zeta^2\bar{\sigma}_{\mathcal{L}}^2\sum_{k=1}^{K}\alpha_k^2\tau\sum_{t'=t}^{t+\tau-1}\eta_{t'}^2\mathbb{E}\left\{\left\|\nabla F_k^{i_{t'}^{(k)}}\left(\tilde{\boldsymbol{w}}_{t'}^{(k)}\right)\right\|^2\middle|\{\tilde{\boldsymbol{w}}_{t'}^{(k)}\}\right\}\right\}\\ &\overset{(b)}{\leq} M\zeta^2\bar{\sigma}_{\mathcal{L}}^2\tau\left(\sum_{t'=t}^{t+\tau-1}\eta_{t'}^2\right)\sum_{k=1}^{K}\alpha_k^2\xi_k^2,\end{aligned} \tag{B.4}$$

where in (a) we used the inequality $\|\sum_{t'=t+1-\tau}^{t+1}\boldsymbol{r}_t\|^2 \leq \tau\sum_{t'=t+1-\tau}^{t+1}\|\boldsymbol{r}_t\|^2$, which holds for any multivariate sequence $\{\boldsymbol{r}_t\}$; and (b) holds since the uniform distribution of the random index i_k implies that the expected value of the stochastic gradient is the full gradient, i.e., $\mathbb{E}\{\nabla F_k^{i_t^{(k)}}(\boldsymbol{w})\} = \nabla F_k(\boldsymbol{w})$, and consequently, $\mathbb{E}\{\|\nabla F_k^{i_t^{(k)}}(\tilde{\boldsymbol{w}}_t^{(k)}) - \nabla F_k(\tilde{\boldsymbol{w}}_t^{(k)})\|^2\} \leq \mathbb{E}\{\|\nabla F_k^{i_t^{(k)}}(\tilde{\boldsymbol{w}}_t^{(k)})\|^2\} \leq \xi_k^2$ by *AS*1. Equation (B.4) proves the theorem. □

C. Proof of Theorem 3

Our proof follows a similar outline to that used in [25], [52], with the introduction of additional arguments for handling the quantization constraints. The unique characteristics of the quantization error which arise from the dithered strategy presented in Section III allow us to rigorously incorporate its contribution into the overall flow of the proof.

1) Recursive Bound on Weights Error: From [24] it follows that the effect of substractive dithered quantization can be modeled as additive noise, independent of the quantized value, whose distribution depends only on the properties of the lattice. In particular, it holds that the distortion induced in quantizing the model update $\boldsymbol{h}_t^{(k)}$, denoted $\boldsymbol{\epsilon}_t^{(k)}$, is an $m\times 1$ zero-mean additive noise vector *independent of* $\boldsymbol{h}_{t\tau}^{(k)}$, *and thus also of* $\tilde{\boldsymbol{w}}_t^{(k)}$ *and* $i_t^{(k)}$. Consequently, by defining the sequence $\boldsymbol{e}_t^{(k)}$ such that $\boldsymbol{e}_t^{(k)} = \boldsymbol{\epsilon}_t^{(k)}$ if t is an integer multiple of τ and $\boldsymbol{e}_t^{(k)} = \boldsymbol{0}$ otherwise, it follows that (9) can be written as

$$\tilde{\boldsymbol{w}}_{t+1}^{(k)} = \begin{cases}\tilde{\boldsymbol{w}}_t^{(k)} - \eta_t\nabla F_k^{i_t^{(k)}}\left(\tilde{\boldsymbol{w}}_t^{(k)}\right) + \boldsymbol{e}_{t+1}^{(k)} & t+1\notin\mathcal{T}_\tau,\\ \sum_{k'=1}^{K}\alpha_{k'}\left(\tilde{\boldsymbol{w}}_t^{(k')} - \eta_t\nabla F_k^{i_t^{(k')}}\left(\tilde{\boldsymbol{w}}_t^{(k')}\right) + \boldsymbol{e}_{t+1}^{(k')}\right) & t+1\in\mathcal{T}_\tau.\end{cases} \tag{C.1}$$

The equivalent model update representation (C.1) allows us to model the effect of subtractive dithered quantization on the overall FL procedure as additional noise corrupting the computation of the stochastic gradients. Building upon this representation, we now follow the strategy proposed in [52] and adapted to heterogeneous data in [25]. This is achieved by defining a virtual sequence $\{\boldsymbol{v}_t\}$ from $\{\tilde{\boldsymbol{w}}_t^{(k)}\}$ which can be shown to behave almost like mini-batch SGD with batch size τ, while being within a bounded distance of the FL model weights $\{\tilde{\boldsymbol{w}}_t^{(k)}\}$, by properly setting the step size η_t. In particular, we define the virtual sequence $\{\boldsymbol{v}_t\}$ via

$$\boldsymbol{v}_t \triangleq \sum_{k=1}^{K} \alpha_k \tilde{\boldsymbol{w}}_t^{(k)}, \tag{C.2}$$

which coincides with $\tilde{\boldsymbol{w}}_t^{(k)}$ when t is an integer multiple of τ. Further define the averaged noisy stochastic gradients and the averaged full gradients as

$$\tilde{\boldsymbol{g}}_t \triangleq \sum_{k=1}^{K} \alpha_k \left(\nabla F_k^{i_t^{(k)}} \left(\tilde{\boldsymbol{w}}_t^{(k)} \right) - \frac{1}{\eta_t} \boldsymbol{e}_{t+1}^{(k)} \right), \tag{C.3a}$$

$$\boldsymbol{g}_t \triangleq \sum_{k=1}^{K} \alpha_k \nabla F_k \left(\tilde{\boldsymbol{w}}_t^{(k)} \right), \tag{C.3b}$$

respectively. Note that since the quantization error is zero-mean and the sample indexes $\{i_t^{(k)}\}$ are independent and uniformly distributed, it holds that $\mathbb{E}\{\tilde{\boldsymbol{g}}_t\} = \boldsymbol{g}_t$. Additionally, the virtual sequence (C.2) satisfies $\boldsymbol{v}_{t+1} = \boldsymbol{v}_t - \eta_t \tilde{\boldsymbol{g}}_t$.

The resulting model is thus equivalent to that used in [25, App. A], and as a result, by assumptions *AS*2–*AS*3, it follows from [25, Lemma 1] that if $\eta_t \leq \frac{1}{4\rho_s}$ then

$$\mathbb{E}\left\{ \|\boldsymbol{v}_{t+1} - \boldsymbol{w}^{\mathrm{o}}\|^2 \right\} \leq (1-\eta_t\rho_c)\mathbb{E}\left\{ \|\boldsymbol{v}_t - \boldsymbol{w}^{\mathrm{o}}\|^2 \right\} + 6\rho_s \eta_t^2 \psi + \eta_t^2 \mathbb{E}\left\{ \|\tilde{\boldsymbol{g}}_t - \boldsymbol{g}_t\|^2 \right\} + 2\mathbb{E}\left\{ \sum_{k=1}^{K} \alpha_k \left\| \boldsymbol{v}_t - \tilde{\boldsymbol{w}}_t^{(k)} \right\|^2 \right\}. \tag{C.4}$$

Expression (C.4) bounds the expected distance between the virtual sequence $\{\boldsymbol{v}_t\}$ and the optimal weights $\boldsymbol{w}^{\mathrm{o}}$ in a recursive manner. We further bound the summands in (C.4), using the following lemmas:

Lemma C.1: If the step size η_t is non-increasing and satisfies $\eta_t \leq 2\eta_{t+\tau}$ for each $t \geq 0$, then, when assumption *AS*1 is satisfied, it holds that

$$\eta_t^2 \mathbb{E}\left\{ \|\tilde{\boldsymbol{g}}_t - \boldsymbol{g}_t\|^2 \right\} \leq \left(1 + 4\, M\zeta^2 \bar{\sigma}_{\mathcal{L}}^2 \tau^2\right) \eta_t^2 \sum_{k=1}^{K} \alpha_k^2 \xi_k^2. \tag{C.5}$$

Lemma C.2: If the step size η_t is non-increasing and satisfies $\eta_t \leq 2\eta_{t+\tau}$ for each $t \geq 0$, then, by *AS*1, it holds that

$$\mathbb{E}\left\{ \sum_{k=1}^{K} \alpha_k \left\| \boldsymbol{v}_t - \tilde{\boldsymbol{w}}_t^{(k)} \right\|^2 \right\} \leq 4(\tau-1)^2 \eta_t^2 \sum_{k=1}^{K} \alpha_k \xi_k^2. \tag{C.6}$$

Next, we define $\delta_t \triangleq \mathbb{E}\{\|\boldsymbol{v}_t - \boldsymbol{w}^{\mathrm{o}}\|^2\}$. When $t \in \mathcal{T}_\tau$, the term δ_t represents the ℓ_2 norm of the error in the weights of the global model. Using Lemmas C.1-C.2, while substituting (C.6) and (C.5) into (C.4), we obtain the following recursive relationship on the weights error:

$$\delta_{t+1} \leq (1 - \eta_t \rho_c)\delta_t + \eta_t^2 b, \tag{C.7}$$

where

$$b \triangleq \left(1 + 4\, M\zeta^2 \bar{\sigma}_{\mathcal{L}}^2 \tau^2\right) \sum_{k=1}^{K} \alpha_k^2 \xi_k^2 + 6\rho_s \psi + 8(\tau-1)^2 \sum_{k=1}^{K} \alpha_k \xi_k^2.$$

The relationship in (C.7) is used in the sequel to prove the FL convergence bound stated in Theorem 3.

2) FL Convergence Bound: Here, we prove Theorem 3 based on the recursive relationship in (C.7). This is achieved by properly setting the step-size and the FL systems parameters in (C.7) to bound $\delta_t = \mathbb{E}\{\|\boldsymbol{v}_t - \boldsymbol{w}^{\mathrm{o}}\|^2\}$, and combining the resulting bound with the strong convexity of the objective *AS*2 to prove (13).

In particular, we set the step size η_t to take the form $\eta_t = \frac{\beta}{t+\gamma}$ for some $\beta > 0$ and $\gamma \geq \max(4\rho_s\beta, \tau)$, for which $\eta_t \leq \frac{1}{4\rho_s}$ and $\eta_t \leq 2\eta_{t+\tau}$, implying that (C.4) and (C.6) hold.

Under such settings, we show that there exists a finite ν such that $\delta_t \leq \frac{\nu}{t+\gamma}$ for all integer $l \geq 0$. We prove this by induction, noting that setting $\nu \geq \gamma\delta_0$ guarantees that it holds for $t = 0$. Consequently, we next show that if $\delta_t \leq \frac{\nu}{t+\gamma}$, then $\delta_{t+1} \leq \frac{\nu}{t+1+\gamma}$. It follows from (C.9) that

$$\begin{aligned} \delta_{t+1} &\leq \left(1 - \frac{\beta}{t+\gamma}\rho_c\right) \frac{\nu}{t+\gamma} + \left(\frac{\beta}{t+\gamma}\right)^2 b \\ &= \frac{1}{t+\tau} \left(\left(1 - \frac{\beta}{t+\gamma}\rho_c\right) \nu + \frac{\beta^2}{t+\gamma} b \right). \end{aligned} \tag{C.8}$$

Consequently, $\delta_{t+1} \leq \frac{\nu}{t+1+\gamma}$ holds when

$$\frac{1}{t+\tau} \left(\left(1 - \frac{\beta}{t+\gamma}\rho_c\right) \nu + \frac{\beta^2}{t+\gamma} b \right) \leq \frac{\nu}{t+1+\gamma},$$

or, equivalently,

$$\left(1 - \frac{\beta}{t+\gamma}\rho_c\right) \nu + \frac{\beta^2}{t+\gamma} b \leq \frac{t+\gamma}{t+1+\gamma} \nu. \tag{C.9}$$

By setting $\nu \geq \frac{1+\beta^2 b}{\beta\rho_c}$, the left hand side of (C.9) satisfies

$$\begin{aligned} &\left(1 - \frac{\beta}{t+\gamma}\rho_c\right) \nu + \frac{\beta^2}{t+\gamma} b = \frac{t-1+\gamma}{t+\gamma}\nu + \left(\frac{1-\beta\rho_c}{t+\gamma}\nu + \frac{\beta^2}{t+\gamma} b \right) \\ &= \frac{t-1+\gamma}{t+\gamma}\nu + \frac{1}{t+\gamma}\left((1-\beta\rho_c)\nu + \beta^2 b \right) \\ &\overset{(a)}{\leq} \frac{t-1+\gamma}{t+\gamma}\nu, \end{aligned} \tag{C.10}$$

where (a) holds since $\nu \geq \frac{1+\beta^2 b}{\beta\rho_c}$. As the right hand side of (C.10) is not larger than that of (C.9), it follows that (C.9) holds for the current setting, which in turn proves that $\delta_{t+1} \leq \frac{\nu}{t+1+\gamma}$. Finally, the smoothness of the objective *AS*2 implies that

$$\mathbb{E}\{F(\boldsymbol{w}_t)\} - F(\boldsymbol{w}^{\mathrm{o}}) \leq \frac{\rho_s}{2}\delta_t \leq \frac{\rho_s \nu}{2(t+\gamma)}, \tag{C.11}$$

which, in light of the above setting, holds for $\nu \geq \max(\frac{1+\beta^2 b}{\beta\rho_c}, \gamma\delta_0)$, $\gamma \geq \max(\tau, 4\beta\rho_s)$, and $\beta > 0$. In particular, setting $\beta = \frac{\tau}{\rho_c}$ results in $\gamma \geq \tau\max(1, 4\rho_s/\rho_c)$ and $\nu \geq \max(\frac{\rho_c^2+\tau^2 b}{\tau\rho_c}, \gamma\delta_0)$, which, when substituted into (C.11), proves (13). □

3) Deferred Proofs: Here we detail the proofs of the intermediate lemmas used for obtaining the recursion (C.7).

a) Proof of Lemma C.1: To prove (C.5), we note that since the quantization noise and the stochastic gradients are mutually independent, it follows from the definition of the gradient vectors (C.3) that

$$\begin{aligned}\eta_t^2\mathbb{E}\left\{\|\tilde{\boldsymbol{g}}_t - \boldsymbol{g}_t\|^2\right\} &= \sum_{k=1}^K \alpha_k^2 \mathbb{E}\left\{\left\|\boldsymbol{e}_{t+1}^{(k)}\right\|^2\right\} \\ &\quad + \eta_t^2\sum_{k=1}^K \alpha_k^2\mathbb{E}\left\{\left\|\nabla F_k^{i_t^{(k)}}\left(\tilde{\boldsymbol{w}}_t^{(k)}\right) - \nabla F_k\left(\tilde{\boldsymbol{w}}_t^{(k)}\right)\right\|^2\right\} \\ &\overset{(a)}{\leq} \sum_{k=1}^K \alpha_k^2\mathbb{E}\left\{\left\|\boldsymbol{e}_{t+1}^{(k)}\right\|^2\right\} + \eta_t^2\sum_{k=1}^K \alpha_k^2\xi_k^2, \end{aligned} \tag{C.12}$$

where (a) holds since the uniform distribution of the random index i_k implies that the expected value of the stochastic gradient is the full gradient, i.e., $\mathbb{E}\{\nabla F_k^{i_t^{(k)}}(\boldsymbol{w})\} = \nabla F_k(\boldsymbol{w})$, and consequently, $\mathbb{E}\{\|\nabla F_k^{i_t^{(k)}}(\tilde{\boldsymbol{w}}_t^{(k)}) - \nabla F_k(\tilde{\boldsymbol{w}}_t^{(k)})\|^2\} \leq \mathbb{E}\{\|\nabla F_k^{i_t^{(k)}}(\tilde{\boldsymbol{w}}_t^{(k)})\|^2\} \leq \xi_k^2$ by *AS*1. Furthermore, the definition of $\boldsymbol{e}_{t+1}^{(k)}$ implies that $\mathbb{E}\{\|\boldsymbol{e}_{t+1}^{(k)}\|^2\} = 0$ for $t+1 \notin \mathcal{T}$, while for $t+1 \in \mathcal{T}$ it holds that $\mathbb{E}\{\|\boldsymbol{e}_{t+1}^{(k)}\|^2\} = \mathbb{E}\{\|\boldsymbol{\epsilon}_{t+1}^{(k)}\|^2\} = M\sigma^2_{\mathcal{L}_{t+1}^{(k)}}$. Now, similarly to the derivation in (B.4), the quantization error induced by UVeQFed satisfies

$$\begin{aligned}&\mathbb{E}\left\{\left\|\boldsymbol{e}_{t+1}^{(k)}\right\|^2\right\} \\ &\leq M\zeta^2\bar{\sigma}_{\mathcal{L}}^2\mathbb{E}\left\{\left\|\sum_{t'=t+1-\tau}^{t+1}\eta_{t'}\nabla F_k^{i_{t'}^{(k)}}\left(\tilde{\boldsymbol{w}}_{t'}^{(k)}\right)\right\|^2\right\} \\ &\overset{(a)}{\leq} M\zeta^2\bar{\sigma}_{\mathcal{L}}^2\tau\sum_{t'=t+1-\tau}^{t+1}\eta_{t'}^2\mathbb{E}\left\{\left\|\nabla F_k^{i_{t'}^{(k)}}\left(\tilde{\boldsymbol{w}}_{t'}^{(k)}\right)\right\|^2\right\} \\ &\overset{(b)}{\leq} M\zeta^2\bar{\sigma}_{\mathcal{L}}^2\tau^2\eta_{t+1-\tau}^2\xi_k^2 \overset{(c)}{\leq} 4\,M\zeta^2\bar{\sigma}_{\mathcal{L}}^2\tau^2\eta_t^2\xi_k^2, \end{aligned} \tag{C.13}$$

where in (a) we used the inequality $\|\sum_t' = t+1-\tau^{t+1}\boldsymbol{r}_t\|^2 \leq \tau\sum_{t'=t+1-\tau}^{t+1}\|\boldsymbol{r}_t\|^2$, which holds for any multivariate sequence $\{\boldsymbol{r}_t\}$; (b) is obtained from assumption *AS*1; and (c) follows since $\eta_{t+1-\tau} \leq 2\eta_{t+1} \leq 2\eta_t$. Substituting (C.13) into (C.12) proves the lemma. □

b) Proof of Lemma C.2. Note that for $t_0 = \lfloor t/\tau\rfloor\tau$, which is an integer multiple of τ, it holds that $\boldsymbol{v}_{t_0} = \tilde{\boldsymbol{w}}_{t_0}^{(k)}$. Since (C.6) trivially holds for $t = t_0$, we henceforth focus on the case where $t > t_0$. We now write

$$\begin{aligned}&\mathbb{E}\left\{\sum_{k=1}^K \alpha_k\left\|\tilde{\boldsymbol{w}}_t^{(k)} - \boldsymbol{v}_t\right\|^2\right\} \\ &= \mathbb{E}\left\{\sum_{k=1}^K \alpha_k\left\|\tilde{\boldsymbol{w}}_t^{(k)} - \tilde{\boldsymbol{w}}_{t_0}^{(k)} - (\boldsymbol{v}_t - \boldsymbol{v}_{t_0})\right\|^2\right\} \\ &\overset{(a)}{\leq} \mathbb{E}\left\{\sum_{k=1}^K \alpha_k\left\|\tilde{\boldsymbol{w}}_t^{(k)} - \tilde{\boldsymbol{w}}_{t_0}^{(k)}\right\|^2\right\} \\ &= \sum_{k=1}^K \alpha_k\mathbb{E}\left\{\left\|\tilde{\boldsymbol{w}}_t^{(k)} - \tilde{\boldsymbol{w}}_{t_0}^{(k)}\right\|^2\right\}, \end{aligned} \tag{C.14}$$

where in (a) we used the fact that for every set $\{\boldsymbol{r}^{(k)}\}$, one can define a random vector $\boldsymbol{r}$ such that $\Pr(\boldsymbol{r} = \boldsymbol{r}^{(k)}) = \alpha_k$, and thus

$$\begin{aligned}\sum_{k=1}^K \alpha_k\left\|\boldsymbol{r}^{(k)} - \sum_{l=1}^K \alpha_l\boldsymbol{r}^{(l)}\right\|^2 &= \mathbb{E}\{\|\boldsymbol{r} - \mathbb{E}\{\boldsymbol{r}\}\|^2\} \\ &\leq \mathbb{E}\{\|\boldsymbol{r}\|^2\} = \sum_{k=1}^K \alpha_k\|\boldsymbol{r}^{(k)}\|^2.\end{aligned}$$

Next, we recall that $\boldsymbol{e}_{t'} = \boldsymbol{0}$ for each $t' = t_0+1, \ldots, t$. Consequently, similarly to the derivation in (C.13),

$$\begin{aligned}\mathbb{E}\left\{\left\|\tilde{\boldsymbol{w}}_t^{(k)} - \tilde{\boldsymbol{w}}_{t_0}^{(k)}\right\|^2\right\} &= \mathbb{E}\left\{\left\|\sum_{t'=t_0}^{t-1}\eta_{t'}\nabla F_k^{i_{t'}^{(k)}}\left(\tilde{\boldsymbol{w}}_{t'}^{(k)}\right)\right\|^2\right\} \\ &\overset{(a)}{\leq} (\tau-1)\sum_{t'=t_0}^{t-1}\eta_{t'}^2\mathbb{E}\left\{\left\|\nabla F_k^{i_{t'}^{(k)}}\left(\tilde{\boldsymbol{w}}_{t'}^{(k)}\right)\right\|^2\right\} \\ &\overset{(b)}{\leq} (\tau-1)^2\eta_{t_0}^2\xi_k^2 \overset{(c)}{\leq} 4(\tau-1)^2\eta_t^2\xi_k^2, \end{aligned} \tag{C.15}$$

where in (a) we used the inequality $\|\sum_{t'=t_0}^{t-1}\boldsymbol{r}_t\|^2 \leq (t-1-t_0)\sum_{t'=t_0}^{t-1}\|\boldsymbol{r}_t\|^2 \leq (\tau-1)\sum_{t'=t_0}^{t-1}\|\boldsymbol{r}_t\|^2$, which holds for any multivariate sequence $\{\boldsymbol{r}_t\}$; (b) is obtained from assumption *AS*1; and (c) follows since $\eta_{t_0} \leq \eta_{t-\tau} \leq 2\eta_t$. Substituting (C.15) into (C.14) proves the lemma. □

REFERENCES

[1] N. Shlezinger, M. Chen, Y. C. Eldar, H. V. Poor, and S. Cui, "Federated learning with quantization constraints," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 8851–8855.

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

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

[4] J. Dean *et al.*, "Large scale distributed deep networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1223–1231.

[5] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Artif. Intell. Statist.*, 2017, pp. 1273–1282.

[6] K. Bonawitz *et al.*, "Towards federated learning at scale: System design," 2019, *arXiv:1902.01046*.

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

[8] 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," 2019, *arXiv:1909.07972*.

[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, 2020.

[10] H. H. Yang, Z. Liu, T. Q. Quek, and H. V. Poor, "Scheduling policies for federated learning in wireless networks," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 317–333, 2019.

[11] M. M. Amiri, D. Gunduz, S. R. Kulkarni, and H. V. Poor, "Update aware device scheduling for federated learning at the wireless edge," 2020, *arXiv:2001.10402*.

[12] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," 2016, *arXiv:1610.05492*.
[13] Y. Lin, S. Han, H. Mao, Y. Wang, and W. J. Dally, "Deep gradient compression: Reducing the communication bandwidth for distributed training," in *Proc. Int. Conf. Learn. Representations*, 2018.
[14] C. Hardy, E. Le Merrer, and B. Sericola, "Distributed deep learning on edge-devices: Feasibility via adaptive compression," in *Proc. IEEE Int. Symp. Netw. Comput. Appl.*, 2017.
[15] A. F. Aji and K. Heafield, "Sparse communication for distributed gradient descent," 2017, *arXiv:1704.05021*.
[16] W. Wen *et al.*, "Terngrad: Ternary gradients to reduce communication in distributed deep learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1509–1519.
[17] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "QSGD: Communication-efficient SGD via gradient quantization and encoding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1709–1720.
[18] S. Horvath, C.-Y. Ho, L. Horvath, A. N. Sahu, M. Canini, and P. Richtarik, "Natural compression for distributed deep learning," 2019, *arXiv:1905.10988*.
[19] A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie, and R. Pedarsani, "FedPAQ: A. communication-efficient federated learning method with periodic averaging and quantization," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2020, pp. 2021–2031.
[20] S. Horváth, D. Kovalev, K. Mishchenko, S. Stich, and P. Richtárik, "Stochastic distributed learning with gradient quantization and variance reduction," 2019, *arXiv:1904.05115*.
[21] J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar, "SignSGD: Compressed optimisation for non-convex problems," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 560–569.
[22] Y. Polyanskiy and Y. Wu, "Lecture notes on information theory," *Lecture Notes* 6.441 (MIT), ECE563 (University of Illinois Urbana-Champaign), and *STAT* 664 (Yale), 2012–2017.
[23] R. Zamir and M. Feder, "On universal quantization by randomized uniform/lattice quantizers," *IEEE Trans. Inf. Theory*, vol. 38, no. 2, pp. 428–436, Mar. 1992.
[24] R. Zamir and M. Feder, "On lattice quantization noise," *IEEE Trans. Inf. Theory*, vol. 42, no. 4, pp. 1152–1159, Jul. 1996.
[25] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-IID data," in *Proc. Int. Conf. Learn. Representations*, 2019.
[26] speedtest.net, "Speedtest United States Market Report," Jul. 9, 2019. [Online]. Available: https://www.speedtest.net/reports/united-states/
[27] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2325–2383, 1998.
[28] P. A. Chou, M. Effros, and R. M. Gray, "A vector quantization approach to universal noiseless coding and quantization," *IEEE Trans. Inf. Theory*, vol. 42, no. 4, pp. 1109–1138, 1996.
[29] J. Ziv, "On universal quantization," *IEEE Trans. Inf. Theory*, vol. IT-31, no. 3, pp. 344–347, May 1985.
[30] R. M. Gray and T. G. Stockham, "Dithered quantizers," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 805–812, May 1993.
[31] S. P. Lipshitz, R. A. Wannamaker, and J. Vanderkooy, "Quantization and dither: A theoretical survey," *J. Audio Eng. Soc.*, vol. 40, no. 5, pp. 355–375, May 1992.
[32] R. Zamir and T. Berger, "Multiterminal source coding with high resolution," *IEEE Trans. Inf. Theory*, vol. 45, no. 1, pp. 106–117, Jan. 1999.
[33] A. Kirac and P. Vaidyanathan, "Results on lattice vector quantization with dithering," *IEEE Trans. Circuits Syst. II Analog Digit. Signal Process.*, vol. 43, no. 12, pp. 811–826, Dec. 1996.
[34] J. H. Conway and N. J. A. Sloane, *Sphere Packings, Lattices and Groups*. vol. 290, Berlin, Germany: Springer Science Business Media, 2013.
[35] R. Rubinstein, "Generating random vectors uniformly distributed inside and on the surface of different regions," *Eur. J. Oper. Res.*, vol. 10, no. 2, pp. 205–209, Jun. 1982.
[36] T. C. Aysal, M. J. Coates, and M. G. Rabbat, "Distributed average consensus with dithered quantization," *IEEE Trans. Signal Process.*, vol. 56, no. 10, pp. 4905–4918, Oct. 2008.
[37] N. Shlezinger and Y. C. Eldar, "Task-based quantization with application to MIMO receivers," 2020, *arXiv:2002.04290*.
[38] N. Shlezinger, Y. C. Eldar, and M. R. Rodrigues, "Hardware-limited task-based quantization," *IEEE Trans. Signal Process.*, vol. 67, no. 20, pp. 5223–5238, Oct. 2019.
[39] N. Shlezinger, Y. C. Eldar, and M. R. Rodrigues, "Asymptotic task-based quantization with application to massive MIMO," *IEEE Trans. Signal Process.*, vol. 67, no. 15, pp. 3995–4012, Aug. 2019.
[40] S. Salamtian, N. Shlezinger, Y. C. Eldar, and M. Médard, "Task-based quantization for recovering quadratic functions using principal inertia components," in *Proc. IEEE Int. Symp. Inf. Theory*, 2019, pp. 390–394.
[41] 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.
[42] T. Sery, N. Shlezinger, K. Cohen, and Y. C. Eldar, "Over-the-air federated learning from heterogeneous data," 2020, *arXiv:2009.12787*.
[43] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y.-J. A. Zhang, "The roadmap to 6G: AI empowered wireless networks," *IEEE Commun. Mag.*, vol. 57, no. 8, pp. 84–90, Aug. 2019.
[44] J. Kang, Z. Xiong, D. Niyato, S. Xie, and J. Zhang, "Incentive mechanism for reliable federated learning: A joint optimization approach to combining reputation and contract theory," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 10 700–10 714, Dec. 2019.
[45] 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.
[46] J. Kang, Z. Xiong, D. Niyato, Y. Zou, Y. Zhang, and M. Guizani, "Reliable federated learning for mobile networks," *IEEE Wireless Commun.*, vol. 27, no. 2, pp. 72–80, Apr. 2020.
[47] M. Chen, U. Challita, W. Saad, C. Yin, and M. Debbah, "Artificial neural networks-based machine learning for wireless networks: A tutorial," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3039–3071, Oct./Dec. 2019.
[48] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 2012.
[49] A. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inf. Theory*, vol. IT-22, no. 1, pp. 1–10, Jan. 1976.
[50] E. Agrell and T. Eriksson, "Optimization of lattices for quantization," *IEEE Trans. Inf. Theory*, vol. 44, no. 5, pp. 1814–1828, 1998.
[51] K. Ferentios, "On Tchebycheff's type inequalities," *Trabajos De Estadistica Y De Investigacion Operativa*, vol. 33, no. 1, p. 125, 1982.
[52] S. U. Stich, "Local SGD converges fast and communicates little," in *Proc. Int. Conf. Learn. Representations*, 2019.
[53] J. Conway and N. Sloane, "Voronoi regions of lattices, second moments of polytopes, and quantization," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 211–226, Mar. 1982.
[54] Y. Zhang, J. C. Duchi, and M. J. Wainwright, "Communication-efficient algorithms for statistical optimization," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 3321–3363, Jan. 2013.
[55] A. Koloskova, N. Loizou, S. Boreiri, M. Jaggi, and S. U. Stich, "A unified theory of decentralized SGD with changing topology and local updates," 2020, *arXiv:2003.10422*.
[56] *MathWorks Deep Learning Toolbox Team*, "Deep learning tutorial series," MATLAB central file exchange, Jun. 9, 2020. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/62990-deep-learning-tutorial-series
[57] G. An, "The effects of adding noise during backpropagation training on a generalization performance," *Neural Comput.*, vol. 8, no. 3, pp. 643–674, Apr. 1996.

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

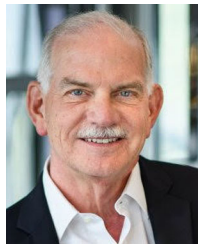
Mingzhe Chen received the Ph.D. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2019. From 2016 to 2019, he was a Visiting Researcher with the Department of Electrical and Computer Engineering, Virginia Tech., Blacksburg, VA, USA. He is currently a Postdoctoral Researcher with the Electrical Engineering Department, Princeton University, Princeton, NJ, USA. His research interests include federated learning, reinforcement learning, virtual reality, unmanned aerial vehicles, and wireless networks. He was the recipient of the 2020 IEEE International Conference on Communications (ICC) Best Paper Award and the 2020 IEEE Global Communications Conference (GLOBECOM) Best Paper Award, and an Exemplary Reviewer for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS in 2018 and the IEEE TRANSACTIONS ON COMMUNICATIONS in 2018 and 2019, respectively. He was a Co-Chair for workshops on edge learning and wireless communications in several conferences including the IEEE ICC, the IEEE GLOBECOM, and the IEEE Wireless Communications and Networking Conference. He is currently an Associate Editor for *Frontiers in Communications and Networks* Specialty Section on Data Science for Communications, and a Guest Editor for the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS Special Issue on Distributed Learning over Wireless Edge Networks.

Yonina C. Eldar (Fellow, IEEE) received the B.Sc. degree in physics in 1995 and the B.Sc. degree in electrical engineering in 1996 both from Tel-Aviv University (TAU), 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 a Professor with the Department of Electrical Engineering, Technion, Haifa, Israel, where she held the Edwards Chair in engineering. She is also a Visiting Professor with MIT, a Visiting Scientist with the Broad Institute, and an Adjunct Professor with Duke University, and was a Visiting Professor at Stanford. She is the author of the book *Sampling Theory: Beyond Bandlimited Systems* and coauthor of five 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.

Dr. Eldar has received many awards for excellence in research and teaching, including the IEEE Signal Processing Society Technical Achievement Award in 2013, the IEEE/AESS Fred Nathanson Memorial Radar Award in 2014, and the IEEE Kiyo Tomiyasu Award in 2016. She was a Horev Fellow of the Leaders in Science and Technology program with the Technion and an Alon Fellow. She was the recipient of 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 with the Technion, the Muriel David Jacknow Award for Excellence in Teaching, and the Technions Award for Excellence in Teaching (two times). She was the recipient of several 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. She 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 also a EURASIP Fellow. 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 several 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 a Technical Co-Chair of several international conferences and workshops.



H. Vincent Poor (Fellow, IEEE) received the Ph.D. degree in EECS from Princeton University, Princeton, NJ, USA, in 1977. From 1977 to 1990, he was with the Faculty of the University of Illinois at Urbana-Champaign, Champaign, IL, USA. Since 1990, he has been with the Princeton University, where he is the Michael Henry Strater University Professor of electrical engineering. From 2006 to 2016, he was the Dean of School of Engineering and Applied Science, Princeton University. He has also held visiting appointments with several other institutions, including most recently at Berkeley and Cambridge. Among his publications is the forthcoming book *Advanced Data Analytics for Power Systems* (Cambridge University Press, 2021). His research interests include information theory, machine learning, and network science, and their applications in wireless networks, energy systems, and related fields. He is a Member of the National Academy of Engineering and the National Academy of Sciences, and is a Foreign Member of the Chinese Academy of Sciences, the Royal Society, and other national and international academies. He was the recipient of the Technical Achievement and Society Awards of the IEEE Signal Processing Society in 2007 and 2011, respectively. Recent recognition of his work includes the 2017 IEEE Alexander Graham Bell Medal, the 2019 ASEE Benjamin Garver Lamme Award, a D.Sc. honoris causa from Syracuse University, awarded in 2017, and a D.Eng. honoris causa from the University of Waterloo, awarded in 2019.

Shuguang Cui (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, USA, in 2005. Afterwards, he has been an Assistant, Associate, Full, Chair Professor in electrical and computer engineering with the University of Arizona, Tucson, AZ, USA, Texas AM University, College Station, TX, USA, University of California, Davis, Davis, CA, USA, and Chinese University of Hong Kong (CUHK), Shenzhen, China He has also been the Executive Dean for the School of Science and Engineering, CUHK, and the Executive Vice Director with Shenzhen Research Institute of Big Data. His current research interests include data-driven large-scale system control and resource management, large data set analysis, IoT system design, energy harvesting based communication system design, and cognitive network optimization. He was selected as the Thomson Reuters Highly Cited Researcher and listed in the Worlds Most Influential Scientific Minds by ScienceWatch in 2014. He was the recipient of the IEEE Signal Processing Society 2012 Best Paper Award. He was the General Co-Chair and TPC Co-Chairs for many IEEE conferences. He has also been the Area Editor of the IEEE Signal Processing Magazine, and an Associate Editor for the IEEE TRANSACTIONS ON BIG DATA, IEEE TRANSACTIONS ON SIGNAL PROCESSING, IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS Series on Green Communications and Networking, and IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. He has been the Elected Member for the IEEE Signal Processing Society SPCOM Technical Committee (2009-2014) and the elected Chair for the IEEE ComSoc Wireless Technical Committee (2017-2018). He is a Member of the Steering Committee for IEEE TRANSACTIONS ON BIG DATA and the Chair of the Steering Committee for IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING. He was also a Member of the IEEE ComSoc Emerging Technology Committee. He was elected as an IEEE Fellow in 2013, an IEEE ComSoc Distinguished Lecturer in 2014, and an IEEE VT Society Distinguished Lecturer in 2019. He was the recipient of the IEEE ICC Best Paper Award, the ICIP Best Paper finalist, the IEEE GLOBECOM Best Paper Award, the First Class Prize in Natural Science from Chinese Institute of Electronics, and the First Class Prize in Technology Invention from China Institute of Communications, all in 2020.