Rate Region for Indirect Multiterminal Source Coding in Federated Learning

Naifu Zhang, Meixia Tao and Jia Wang

Abstract

One of the main focus in federated learning (FL) is the communication efficiency since a large number of participating edge devices send their updates to the edge server at each round of the model training. Existing works reconstruct each model update from edge devices and implicitly assume that the local model updates are independent over edge device. In FL, however, the model update is an indirect multi-terminal source coding problem where each edge device cannot observe directly the source that is to be reconstructed at the decoder, but is rather provided only with a noisy version. The existing works do not leverage the redundancy in the information transmitted by different edges. This paper studies the rate region for the indirect multiterminal source coding problem in FL. The goal is to obtain the minimum achievable rate at a particular upper bound of gradient variance. We obtain the rate region for multiple edge devices in general case and derive an explicit formula of the sum-rate distortion function in the special case where gradient are identical over edge device and dimension. Finally, we analysis communication efficiency of convex Mini-batched SGD and non-convex Minibatched SGD based on the sum-rate distortion function, respectively.

Index Terms

Federated learning, multiterminal rate-distortion theory, quadratic vector Gaussian CEO problem, communication efficiency.

I. Introduction

Federated learning (FL) [1]–[4] is a new edge learning framework that enables many edge devices to collaboratively train a machine learning model without exchanging datasets under the coordination of an edge server. In FL, each edge device downloads a shared model from the edge server, computes an update to the current model by learning from its local dataset, then

N. Zhang, M. Tao and J. Wang are with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, 200240, P. R. China. (email: arthaslery@sjtu.edu.cn; mxtao@sjtu.edu.cn; jiawang@sjtu.edu.cn.).

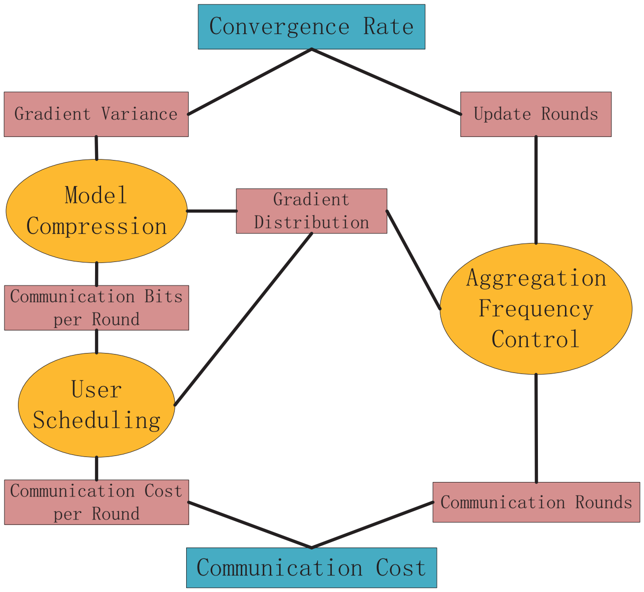


Fig. 1. Illustration of technologies that improve communication efficiency in FL.

sends this update to the edge server. Therein, the updates are averaged to improve the shared model. Compared with traditional learning at a centralized data center, FL offers several distinct advantages, such as preserving privacy, reducing network congestion, and leveraging distributed on-device computation. FL has recently attracted significant attention from both academia and industry, such as [5]–[8].

The main focus in the research area is communication-efficient FL. Specifically, the communication efficient FL is to achieve better convergence rate (high model accuracy) with lower communication costs. The works of SGD convergence analysis [9], [10] state that the convergence rate mainly depends on the variance bound of gradient and the number of updates. The communication cost depends on the communication cost per round and the communication rounds. Fig. 1 shows three types of technologies that improve communication efficiency in FL, including aggregation frequency control, compression schemes and user scheduling.

1) *Aggregation Frequency Control:* Several recent methods have been proposed to improve communication-efficiency in federated settings by allowing for a variable aggregation frequency to be applied on each edge device in parallel at each communication round. Reducing the aggregate frequency will reduce the communication cost by decreasing the number of communication rounds, while increase the convergence rate by increasing the gradient variance. Hence, the

control of aggregation frequency significantly influences the communication-efficiency. The work [1] proposed a framework named ”FedAvg” by updating local model at each device with a given number of SGD iterations and model averaging. The work [11] studied the distributed machine learning across multiple datacenters in different geographical locations, where a threshold-based approach toreduce the communication among different datacenters is proposed. The work [12] propose a control algorithm that determines the best tradeoff between the local update and global parameter aggregation under a given resource budget.

2) *Model Compression:* While aggregation frequency control can reduce the number of communication rounds, model compression schemes such as sparsification and quantization can significantly reduce the communication bits per round with increasing gradient variance. Partially initiated by a 1-bit implementation of SGD by Microsoft in [13], a large number of recent studies revisited the idea of low-precision training as a means to reduce communication [14]–[18]. Other approaches for low-precision training focus on sparsification of gradients, either by thresholding small entries or by random sampling [19]–[22]. Several approaches, including QSGD [14] and TernGrad [16], implicitly combine quantization and sparsification to maximize communication efficiency.

3) *User Scheduling:* In federated learning, edge devices may differ in terms of computation capacity, communication quality and data distribution. The goal of user scheduling is to minimize the communication cost per round and transmit the bits required by quantization, by selecting a subset of devices to upload model update in each round. A host of studies propose various scheduling policy for FL, ranging from minimizing the transmission latency [23], local model importance aware user scheduling [24], [25], to accounting for both the staleness and the communication quality to improve the communication efficiency [26].

All the existing researches aim to reconstruct each model update from edge devices and implicit assume that the local model updates are independent over edge device. For example, the model compression schemes quantize and/or compress the models or stochastic gradients at each individual edge device, and the edge server recovers the exact value of the stochastic gradients of each individual edge device. The user scheduling works and aggregation frequency control works also assume that the communication cost increases linearly with the number of edge devices. However, we observe that the main objective in FL is a good estimate of a model update at the edge server by using the information received from edge devices, rather than the exact recovery of each model update from each devices. It is an indirect multi-terminal source coding problem, where each edge device cannot observe directly the source that is to be reconstructed at the decoder, but is rather provided only with a noisy version. Specifically, in FL, the objective is estimating the global model update computed by gradient decent (GD) on global dataset, while the local model update computed by each edge device is a noisy version of the global model update. Furthermore, the local gradients are highly correlated among different workers, providing opportunity for correlated source coding. Hence, the existing works do not leverage the redundancy in the information transmitted by different edges.

In addition, although the model compression schemes have provided the tradeoff between the variance of received gradients at edge server and the achievable communication bits, their proposed achievable communication bits are inconsistent. It is hard to combine the existing model compression schemes with the communication efficient FL works such as the user aggregation frequency control and user scheduling. Therefore, the transmission model of communication efficient FL urgently requires a theoretical limits of communication bits at a particular gradient variance instead of simply assuming that the gradient parameters are transmitted perfectly with a constant communication bits.

Motivated by the above issue, in this paper, we derive the rate region for the indirect multiterminal source coding problem in FL referred to as the quadratic vector Gaussian CEO problem. Our goal is to obtain the minimum achievable rate at a particular upper bound of gradient variance. We formulate the indirect multiterminal source coding problem based on the observation of the gradient distribution and solve it from the standpoint of multiterminal rate-distortion theory. Our result can be regarded as a tool to analysis the communication efficiency in a certain FL system. The main contributions of this work are outlined below:

- *Gradient distribution model:* We have made assumptions of the distribution of global gradient and local gradient in federated learning, respectively and they are justified by the experiment results on MNIST dataset. We reveal that the multiterminal source coding problem in FL is the quadratic vectored Gaussian CEO problem base on a thorough understanding of gradient distributions.
- *Rate region results:* We derive the rate region for the K edge devices quadratic Gaussian vectored CEO problem. We show that the Berger-Tung achievable region is indeed tight. Our converse proof is inspired by the converse in work [27]. But unlike there, the distortion

of our problem is the variance of estimator instead of MSE.

- *Explicit rate-distortion function* We derive an explicit formula of the rate-distortion function in the special case where gradient are identical over edge device and dimension. The derived function has the form of a sum of two nonnegative functions. One is a classical rate-distortion function for single Gaussian source and the other is a new rate-distortion function which dominates the performance of the system for a relatively small distortion.
- *Communication efficiency analysis:* we analysis communication efficiency of convex Minibatched SGD and non-convex Minibatched SGD based on the sum-rate distortion function, respectively. We provide an inherent trade-off between communication cost and convergence guarantees.

The rest of this paper is organized as follows. The FL system is modeled in Section II. Section III formulates the multi-terminal source coding problem in FL. In Section IV, we give the rate region results. Section V we prove the direct and converse the rate region results. Section VI derives the explicit rate region. Section VII provides the communication efficiency in FL. Finally we conclude the paper in Section VIII.

II. Federated Learning

There are many variants in FL, which have different preconditions and convergence guarantees. Our proposed rate region results are very adaptable. They can be applied to any FL whose local gradient distribution meets our assumed Gaussian distribution. To facilitate the presentation, we only focus on a basic FL setup. However, the results can be extended to cases with gradient distribution under our assumptions. In this section, we introduce the system model and the convergence rate of federated learning in error-free communication.

A. System Model

We consider a FL framework as illustrated in Fig. 2, where a shared AI model (e.g., a classifier) is trained collaboratively across K edge devices via the coordination of an edge server. Let $\mathcal{K} = \{1, ..., K\}$ denote the set of edge devices. Each device $k \in \mathcal{K}$ collects a fraction of labelled training data via interaction with its own users, constituting a local dataset, denoted as $\mathcal{S}_k$. Let $\boldsymbol{w} \in \mathbb{R}^P$ denote the P-dimensional model parameter to be learned. The loss function measuring

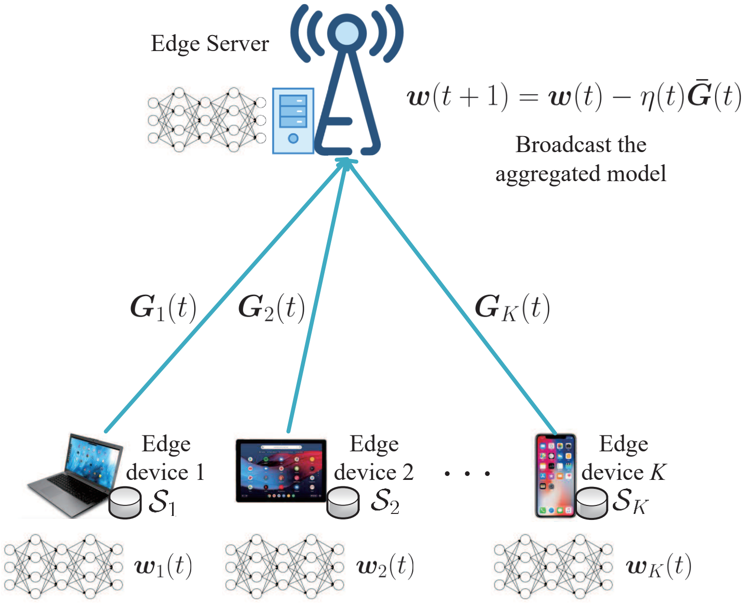


Fig. 2. Illustration of federated learning system in error-free communication.

the model error is defined as

$$F(\boldsymbol{w})=\sum_{k\in\mathcal{K}}\frac{|\mathcal{S}_k|}{|\mathcal{S}|}F_k(\boldsymbol{w}),\tag{1}$$

where $F_k(\boldsymbol{w})=\frac{1}{|\mathcal{S}_k|}\sum_{i\in\mathcal{S}_k}f_i(\boldsymbol{w})$ is the loss function of device k quantifying the prediction error of the model $\boldsymbol{w}$ on the local dataset collected at the k-th device, with $f_i(\boldsymbol{w})$ being the sample-wise loss function, and $\mathcal{S}=\bigcup_{k\in\mathcal{K}}\mathcal{S}_k$ is the union of all datasets. Let $\boldsymbol{G}(t)\triangleq\nabla F(\boldsymbol{w}(t))\in\mathbb{R}^P$ denote the gradient vector calculated through the gradient descent (GD) algorithm at iteration t. The minimization of $F(\boldsymbol{w})$ is typically carried out through the Minibatched stochastic gradient descent (Minibatched SGD) algorithm, where device k's local dataset $\mathcal{S}_k$ is split into mini-batches of size B_k and at each iteration $t=1,2,...$, we draw one mini-batch $\mathcal{B}_k(t)$ randomly and calculate the local gradient vector as

$$\boldsymbol{G}_k(t)=\nabla\frac{1}{B_k}\sum_{i\in\mathcal{B}_k(t)}f_i(\boldsymbol{w}).\tag{2}$$

When the mini-batch size $B_k=1$, the Minibatched SGD algorithm reduces to SGD algorithm. In this case, we say the local gradient $\boldsymbol{G}_k(t)$ has variance $\sigma^2(t)$ at iteration t, i.e., $\mathbb{E}[\|\boldsymbol{G}_k(t)-\boldsymbol{G}(t)\|^2]=\sigma^2(t)$. In general case, local gradient $\boldsymbol{G}_k(t)$ has variance $\frac{\sigma^2(t)}{B_k(t)}$ at iteration t. If all the local gradients $\{\boldsymbol{G}_k(t)\}_{k=1}^K$ are available at edge server through error-free transmission, the optimal estimator of $\boldsymbol{G}(t)$ is the Sample Mean Estimator, i.e., $\bar{\boldsymbol{G}}(t)=\sum_{k=1}^K\frac{B_k(t)}{B(t)}\boldsymbol{G}_k(t)$, where

$B(t) \triangleq \sum_{k=1}^{K} B_k(t)$ is global batch size. It is not hard to see that the variance of the optimal estimator $\bar{\boldsymbol{G}}(t)$ is $\frac{\sigma^2(t)}{B(t)}$. Then the edge server update the model parameter as

$$\boldsymbol{w}(t+1) = \boldsymbol{w}(t) - \eta(t)\bar{\boldsymbol{G}}(t), \tag{3}$$

with $\eta(t)$ being the learning rate at iteration t.

B. Convergence rate

Given access to local gradients, and a starting point $\boldsymbol{w}(0)$, Minibatched SGD builds iterates $\boldsymbol{w}(t)$ given by Equation (3), projected onto $\mathbb{R}^P$, where $\{\eta\}_{t\geq 0}$ is a sequence of learning rate. In this setting, one can show:

Theorem 1 ([9], Theorem 6.3): Let $F : \mathbb{R}^P \to \mathbb{R}$ be unknown, convex and L-smooth. Let $\boldsymbol{w}(0)$ be given, and let $A^2 = \sup_{\boldsymbol{w}\in\mathbb{R}^P} \|\boldsymbol{w} - \boldsymbol{w}(0)\|^2$. Let $T > 0$ be fixed. Given repeated, independent access to the estimator of gradients with variance bound D for loss function F, i.e., $\mathbb{E}[\|\bar{\boldsymbol{G}}(t) - \boldsymbol{G}(t)\|^2] \leq D$ for all $t = 1, 2, ..., T$, training with initial point $\boldsymbol{w}(0)$ and constant step sizes $\eta(t) = \frac{\gamma}{L+1}$, where $\gamma = A\sqrt{\frac{2}{DT}}$, achieves

$$\mathbb{E}\left[L\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{w}(t)\right)\right] - \min_{\boldsymbol{w}\in\mathbb{R}^P} L(\boldsymbol{w}) \leq A\sqrt{\frac{2D}{T}} + \frac{LA^2}{T}. \tag{4}$$

Remark 1: When the gradient vector calculated through the gradient descent (GD) algorithm and the model is updated with gradient sequence $\{\boldsymbol{G}(t)\}_{t=1}^{T}$ in error-free transmission, the variance bound $D = 0$, the convergence rate achieve the best $\frac{LA^2}{T}$, thus the edge server is interested in this sequence. When the gradient vector calculated through minibatched SGD algorithm, and the model is updated with the unbiased estimator $\{\bar{\boldsymbol{G}}(t)\}_{t=1}^{T}$ in error-free transmission, the variance bound $D = \max\limits_{t=1,2,...,T} \frac{\sigma^2(t)}{B(t)}$ given the global batch size $B(t)$. However, the error-free transmission is infeasible in practical due to communication resource limitations. The local gradients have to be quantized and the unbiased estimation can only be based on these quantized values. Hence, the variance of the unbiased estimator based on quantized gradients must be larger than $\frac{\sigma^2(t)}{B(t)}$ at each iteration t.

In general, convergence rate increases with the variance bound of gradient estimator given the number of the iterations while the communication bits per iteration decreases with the variance bound of gradient estimator. To obtain the trade-off between the convergence rate and the communication cost, we need to find out what is the minimum achievable rate at a particular variance upper bound. It is a basic problem in rate distortion theory.

III. Multi-terminal source coding problem in FL

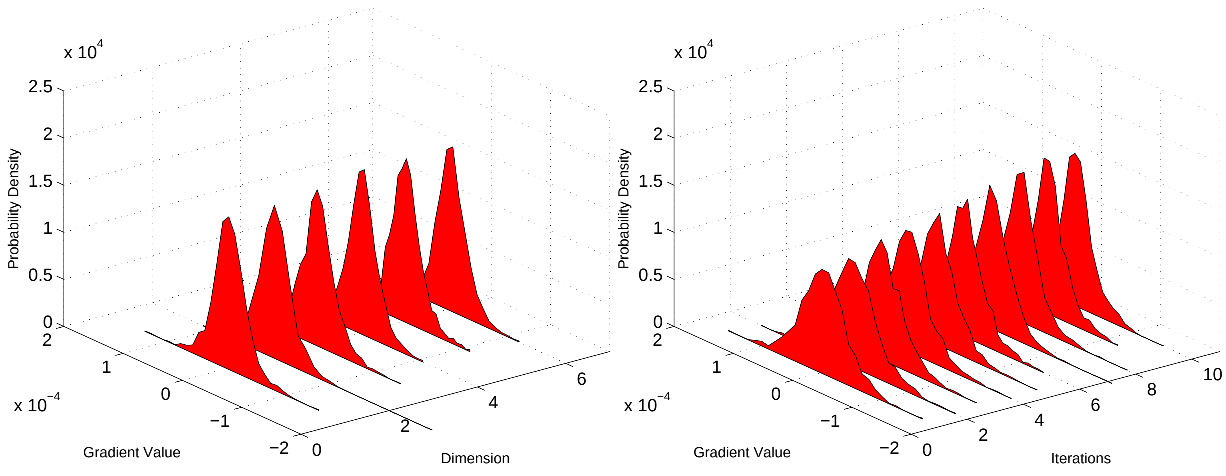
In this section, in order to accurately formulate the indirect multiterminal source coding problem in federated learning, we first study the distribution of the global gradients and the local gradients in FL. Then we formulate a quadratic vector Gaussian CEO problem based on a thorough understanding of gradient distributions.

A. Gradient distribution

1) Global Gradient: Recall that the edge server is interested in this sequence $\{\boldsymbol{G}(t)\}_{t=1}^{\infty}$, which is the global gradient vector sequence calculated through the gradient descent (GD) algorithm. The following are key assumptions on the distribution of $\{\boldsymbol{G}(t)\}_{t=1}^{\infty}$

- The gradient $\{\boldsymbol{G}(t)\}_{t=1}^{\infty}$ is independent distributed vector Gaussian sequence with zero mean. The Gaussian distribution is valid since the field of probabilistic modeling uses Gaussian distributions to model the gradient [28]–[31]. The assumption of independence is valid when the learning rate is large enough.
- The gradient $\boldsymbol{G}(t)$ are independent over gradient vector dimension p's. This assumption is valid as long as the features in a data sample are independent but non-identically distributed, which is typically the case. Even if the gradient are strongly correlated over dimension p's, the gradient can be de-correlated by the regularization methods such as sparsity-inducing regularization [32], [33] and parameter sharing/tying [34], [35].

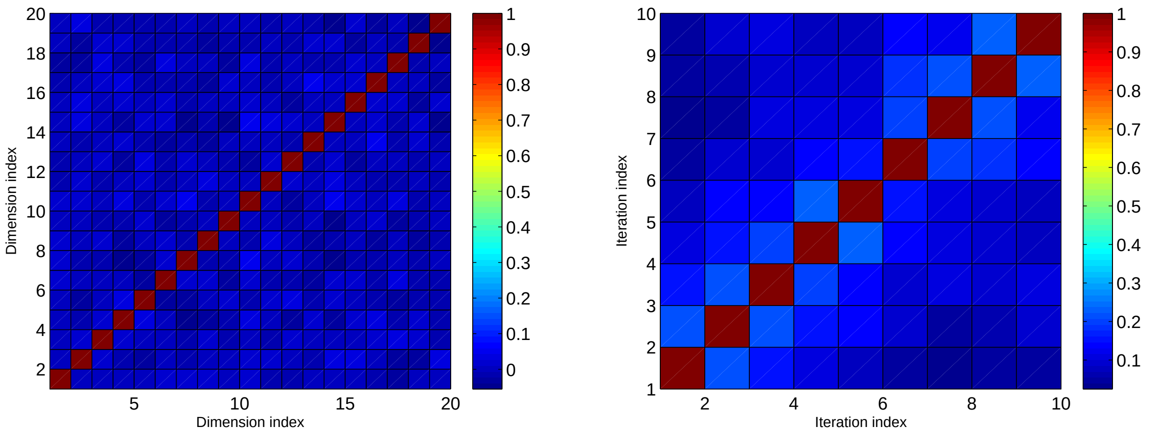
We perform experiments on dataset MNIST to justify the assumption of global gradients $\boldsymbol{G}(t)$. We evenly sampled 25000 gradients in iterations [1, 10]. Fig. 3 illustrates the experimental results of the distribution of global gradient of dataset MNIST. It is observed that the global gradient $\boldsymbol{G}(t)$ follow a Gaussian distribution with mean zero. Fig. 3(a) shows 6 dimensions on two layers of global gradient $\boldsymbol{G}(t)$ in iteration 10. The variance of global gradient on the same layer are similar and the variance of global gradient on the different layer are different. Fig. 3(b) shows one dimension of global gradient $\boldsymbol{G}(t)$ in iteration [1, 10]. It is observed that the variance of the global gradient gradually decreases with the model converges.



(a) The distribution of global gradient at different dimension.

(b) The distribution of global gradient at different iteration.

Fig. 3. The distribution of global gradient.



(a) Correlation coefficient of global gradient over dimension p's.

(b) Correlation coefficient of global gradient over iteration t's.

Fig. 4. Correlation coefficient of global gradient.

The correlation coefficient of Gaussian variables can indicate the independence of them. Fig. 4 shows the correlation coefficient of global gradient over iteration t's and dimension p's. Fig. 4(a) shows the correlation coefficient of global gradient over dimension p's. It is observed that the correlation coefficient of gradients on two different dimension are almost zero. The global gradient sequence $\{\boldsymbol{G}(t)\}_{t=1}^{\infty}$ are almost independent over dimension p's. Fig. 4(b) shows the correlation coefficient of global gradient over iteartion t's. It is observed that the correlation coefficient of gradients on two adjacent iterations are large at the beginning of the training and are almost zero as the training progresses. The global gradient sequence $\{\boldsymbol{G}(t)\}_{t=1}^{\infty}$ are almost independent over iteration t's.

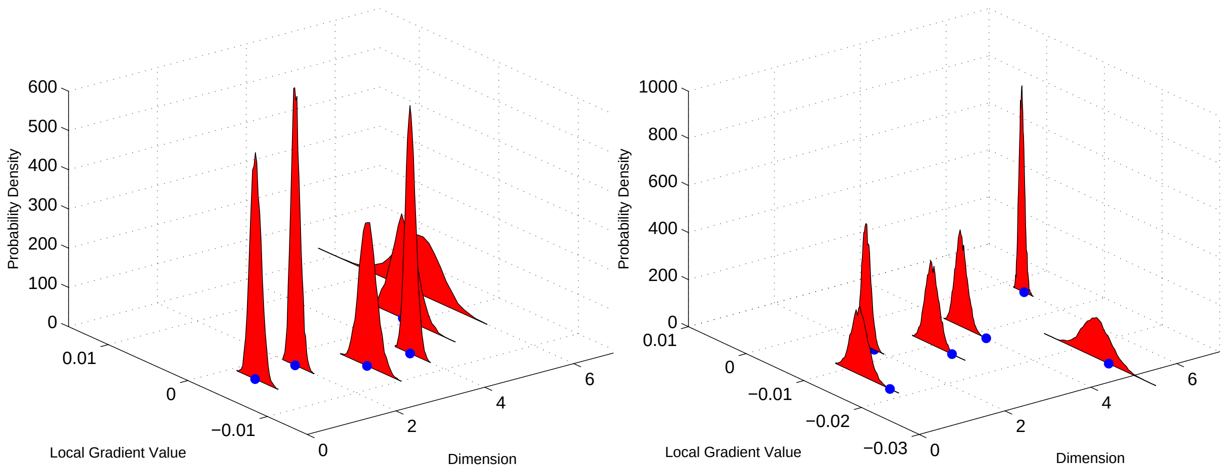
2) Local Gradients: For $k = 1, 2, ..., K$, edge device k carries out Minibatched SGD in FL. Recall that $\{\boldsymbol{G}_k(t)\}_{t=1}^{\infty}$ is the local gradient vector sequence calculated through the Minibatched SGD algorithm at device k. The local gradient vector sequence $\{\boldsymbol{G}_k(t)\}_{t=1}^{\infty}$ can be viewed as noisy versions of $\{\boldsymbol{G}(t)\}_{t=1}^{\infty}$ and corrupted by additive noise, i.e., $\boldsymbol{G}_k(t) = \boldsymbol{G}(t) + \boldsymbol{N}_k(t)$. The following are key assumptions on the distribution of $\{\boldsymbol{N}_k(t)\}_{t=1}^{\infty}$ for $k \in \mathcal{K}$:

- The gradient noise $\boldsymbol{N}_k(t)$ are Gaussian random vectors independent of the $\boldsymbol{G}$ process. The mean of $\boldsymbol{N}_k(t)$ is zero in IID data setting and non-zero in non-IID data setting. Note that the gradient noise the $\boldsymbol{N}_k(t)$ depends on the selection of local batch at edge device k. The assumption of independence is valid since the selection of local batch is independence of the global gradient $\boldsymbol{G}$.
- The gradient noise $\boldsymbol{N}_k(t)$ are independent and non-identical distributed over devices k's. This assumption is valid as long as the selection of the local batch are independent and non-identical over edge device k.
- The gradient noise $\boldsymbol{N}_k(t)$ are independent and non-identical distributed over dimension p's. The reason for this assumption is similar to that of gradient $\boldsymbol{G}(t)$.

We perform experiments to justify the assumption that gradients noises $\boldsymbol{N}_k(t)$ are Gaussian distribution and independent over iteration t's, devices k's and dimension p's.

Fig. 5 illustrates the experimental results of the conditional distribution of local gradient $\boldsymbol{G}_k(t)$ given global gradient $\boldsymbol{G}(t)$ in IID data setting and non-IID data setting. We sampled 10000 local gradients from one edge device with IID dataset and one edge device with non-IID dataset within one iteration, respectively. It is observed that the conditional local gradient $\boldsymbol{G}_k(t)$ given $\boldsymbol{G}(t)$ follows a Gaussian distribution, i.e., the gradient noise $\boldsymbol{N}_k(t)$ follow a Gaussian distribution. From Fig. 5(a), it is observed that the conditional expectation of local gradient is equal to global gradient, i.e.,

$$\mathbb{E}[\boldsymbol{G}_k(t)|\boldsymbol{G}(t) = \boldsymbol{g}(t)] = \boldsymbol{g}(t). \tag{5}$$

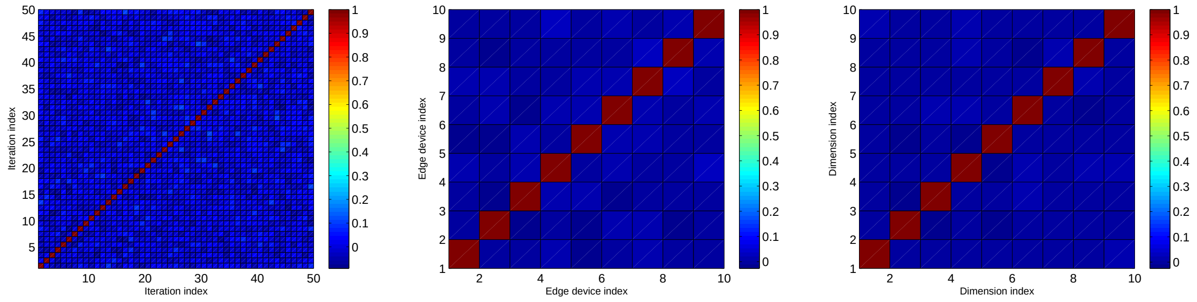


(a) IID data setting.

(b) Non-IID data setting.

Fig. 5. The conditional distribution of local gradient $\boldsymbol{G}_k(t)$ given global gradient $\boldsymbol{G}(t)$ in IID data setting and non-IID data setting in IID data setting and non-IID data setting, where the blue dot represents the given global gradient $\boldsymbol{G}(t)$.

Hence, we have justified that the mean of noise $\boldsymbol{N}_k(t)$ is zero in IID data setting. From Fig. 5(b), it is observed that $\mathbb{E}[\boldsymbol{G}_k(t)|\boldsymbol{G}(t)=\boldsymbol{g}(t)] \neq \boldsymbol{g}(t)$, similarly we have justified that the mean of noise $\boldsymbol{N}_k(t)$ is non-zero in non-IID data setting.



(a) Correlation coefficient of noise over iteration t's.

(b) Correlation coefficient of noise over device k's.

(c) Correlation coefficient of noise over dimension p's.

Fig. 6. Correlation coefficient of gradient noise over iteration, device and dimension.

Fig. 6 illustrates the correlation coefficient of gradient noise over iteration t, device k and dimension p. We sample 1000 times in each gradient noise $\boldsymbol{N}_k(t)$ for $t=1,2,...,50$ and $k=1,2,...,10$. We randomly select 10-dimensional in the gradient noise $\boldsymbol{N}_k(t)$ to show the correlation coefficient over dimension p. It is observed that the correlation coefficient of gradient

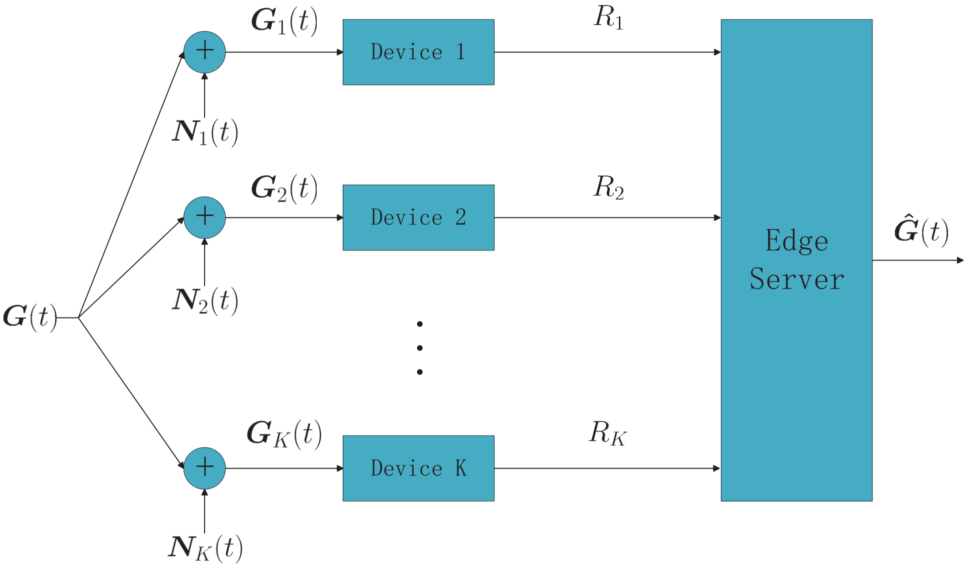


Fig. 7. The Gaussian vector CEO problem in FL.

noise are almost zero over dimension p's, device k's and iteration t's.. Hence, the noise sequence $\{\boldsymbol{N}_k(t)\}_{t=1}^{\infty}$ are independent over dimension p's, device k's and iteration t's.

B. Problem Formulation

In this subsection, we formulate the quadratic vector gaussian CEO problem based on the observation of the distribution of the global gradient $\boldsymbol{G}(t)$ and the local gradient $\boldsymbol{G}_k(t)$. Let the global gradient $\{\boldsymbol{G}(t)\}_{t=1}^{\infty}$ be an independent Gaussian vector sequence with mean 0 and variance $\text{diag}(\sigma_{X_1}^2(t), \sigma_{X_2}^2(t), ..., \sigma_{X_P}^2(t))$. Each $\boldsymbol{G}(t), t = 1, 2, ...$ takes value in real space $\mathbb{R}^P$. For $k = 1, ..., K$, Let the local gradient $\boldsymbol{G}_k(t)$ be noisy version of $\{\boldsymbol{G}(t)\}_{t=1}^{\infty}$, each taking value in real space $\mathbb{R}^P$ and corrupted by independent additive white Gaussian noise, i.e.,

$$\boldsymbol{G}_k(t) = \boldsymbol{G}(t) + \boldsymbol{N}_k(t), \tag{6}$$

where $\boldsymbol{N}_k(t)$ are Gaussian random vectors independent over device k, dimension p and iteration t. For $k = 1, 2, ..., K$, $p = 1, 2, ..., P$ and $t = 1, 2, ...$, we assume that $\boldsymbol{N}_k(t)$ is a centralized Gaussian variable with mean 0 and variance $\text{diag}(\sigma_{N_{k,1}}^2(t), \sigma_{N_{k,2}}^2(t), ..., \sigma_{N_{k,P}}^2(t))$.

Fig. 7 shows the CEO Problem in FL. The edge server (CEO) is interested in the sequence $\{\boldsymbol{G}(t)\}_{t=1}^{\infty}$ that cannot be observed directly. The edge server employs a team of K edge devices (agents) who observe independently corrupted versions $\{\boldsymbol{G}_k(t)\}_{t=1}^{\infty}, k = 1, 2, ..., K$ of $\{\boldsymbol{G}(t)\}_{t=1}^{\infty}$.

We write n independent copies of $\boldsymbol{G}(t)$ and $\boldsymbol{G}_k(t)$ as $\boldsymbol{G}^n(t)$ and $\boldsymbol{G}_k^n(t)$, respectively. To facilitate the following derivation, we omit iteration index t. For $k = 1, 2, ..., K$, each local gradient sequence $\boldsymbol{G}_k^n$ observed by edge device k is separately encoded to $\phi_k(\boldsymbol{G}_k^n)$, and those are sent to the information processing center, where the edge server observes $\phi_k(\boldsymbol{G}_k^n), k = 1, 2, ..., K$ and outputs the estimation $\hat{\boldsymbol{g}}^n$ of $\boldsymbol{G}^n$ by using the decoder function ψ_K. The encoder function $\phi_k, k = 1, 2, ..., K$ are defined by

$$\phi_k : \mathbb{R}^{nP} \rightarrow \mathcal{C}_k = \{1, 2, ..., |\mathcal{C}_k|\}, \tag{7}$$

and satisfy the total rate constraint

$$\frac{1}{n} \log |\mathcal{C}_k| \leq R_k, k = 1, 2, ..., K. \tag{8}$$

We write an K-tuple of encoder functions $\phi_k, k = 1, 2, ..., K$ as

$$\phi^K = (\phi_1, \phi_2, ..., \phi_K). \tag{9}$$

Similarly, we write

$$\phi^K(\boldsymbol{G}^{nK}) = (\phi_1(\boldsymbol{G}_1^n), \phi_2(\boldsymbol{G}_2^n), ..., \phi_K(\boldsymbol{G}_K^n)). \tag{10}$$

The decoder function ψ_K is defined by

$$\psi_K : \mathcal{C}_1 \times \mathcal{C}_2 \times ... \times \mathcal{C}_K \rightarrow \mathbb{R}^{nP}. \tag{11}$$

For $\hat{\boldsymbol{G}}^n = \psi_K(\phi^K(\boldsymbol{G}^{nK}))$, define the average mean squared error (MSE) distortion by

$$D^n(\boldsymbol{G}^n, \hat{\boldsymbol{G}}^n) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \|\boldsymbol{G}[i] - \hat{\boldsymbol{G}}[i]\|^2. \tag{12}$$

For a target distortion D, a rate K-tuple $(R_1, R_2, ..., R_K)$ is said to be achievable if there are encoders ϕ^K satisfying (8) and decoder ψ_K such that $\hat{\boldsymbol{G}}^n$ is unbiased estimator of $\boldsymbol{G}^n$, i.e., $\mathbb{E}[\hat{\boldsymbol{G}}^n] = \boldsymbol{G}^n$ and $D^n(\boldsymbol{G}^n, \hat{\boldsymbol{G}}^n) \leq D$ for some n. The closure of the set of all achievable rate K-tuples is called the rate-region and we denote it by $\mathcal{R}_\star \subseteq \mathbb{R}_+^K$. Our aim is to characterize the region $\mathcal{R}_\star$ in an explicit form.

Remark 2: This problem is difference from the existing Quadratic Gaussian CEO works [27], [36], [37] in the following aspects:

- The estimator of our CEO Problem must be unbiased and the distortion is the variance of estimator while existing Quadratic Vectored Gaussian CEO Problem use MSE as distortion.We choose this distortion function is because MSE cannot guarantee the convergence of federated learning. Recall that Theorem 1 require that the gradient estimator is unbiased.Intuitively, the effective information and noise in the gradient estimator will be enlarged or reduced in proportion to the learning rate η, and the learning rate η can be adjusted. Hence, the convergence rate depends on the ratio between the effective information and noise in the gradient estimator and not on the absolute noise in the gradient estimator. The MSE can only measure the absolute noise in the gradient estimator, but cannot measure the effective information in the gradient estimator. Our unbiased gradient estimator fixes the effective information in the gradient estimator and the variance measures the noise in the gradient estimator.

- Second, existing works assume that the sequence length n is large enough when deriving the achievability of the rate region. However, the local gradient sequence $\{\boldsymbol{G}_k(t)\}_{t=1}^{\infty}$ cannot jointly encode, i.e, the sequence length $n = 1$ in FL due to the fact that the local gradient $\boldsymbol{G}_k(t-1)$ must be transmitted to edge server before the local gradient $\boldsymbol{G}_k(t)$ is available. Therefore, the rate distortion function results may not be achievable in practice and is actually a lower bound of achievable rate. Specifically, see Theorem 3 in section VII.

IV. Rate Region Results

Let $\mathcal{R}_N(D)$ be the Berger-Tung achievable region using Gaussian auxiliary random variables for CEO problem in FL. We show that

$$\mathcal{R}_N(D) = \Bigg\{(R_1, ..., R_K) \in \mathbb{R}_+^K : R_k \geq \sum_{p=1}^{P} R_{k,p}, k = 1, 2, ..., K \tag{13}$$

$$\forall (R_{1,p}, R_{2,p}, ..., R_{K,p}) \in \mathcal{R}_p(r_{1,p}, r_{2,p}, ..., r_{K,p}), p = 1, 2, ..., P \tag{14}$$

$$\forall (r_{1,p}, r_{2,p}, ..., r_{K,p}) \in \mathcal{F}_p(D_p), p = 1, 2, ..., P \tag{15}$$

$$\forall (D_1, D_2, ..., D_P) \in \mathbb{R}_+^P, \sum_{p=1}^{P} D_p = D\Bigg\} \tag{16}$$

where

$$\mathcal{R}_p(r_{1,p}, r_{2,p}, ..., r_{K,p}) = \Bigg\{(R_{1,p}, R_{2,p}, ..., R_{K,p}) :$$

$$\sum_{k\in\mathcal{A}} R_{k,p} \geq \sum_{k\in\mathcal{A}} r_{k,p} + \frac{1}{2}\log\left(\frac{1}{\sigma^2_{X_p}} + \frac{1}{D_p}\right) - \frac{1}{2}\log\left(\frac{1}{\sigma^2_{X_p}} + \sum_{k\in\mathcal{A}^c}\frac{1-\exp(-2r_{k,p})}{\sigma^2_{N_{k,p}}}\right), \forall\mathcal{A}\subseteq\mathcal{K}\Bigg\}, \tag{17}$$

and

$$\mathcal{F}_p(D_p) = \left\{(r_{1,p}, r_{2,p}, ..., r_{K,p}) \in \mathbb{R}^K_+ : \sum_{k=1}^{K}\frac{1-\exp(-2r_{k,p})}{\sigma^2_{N_{k,p}}} = \frac{1}{D_p}\right\}, \tag{18}$$

where $\text{diag}(\sigma^2_{X_1}(t), ..., \sigma^2_{X_P}(t))$ is the variance of global gradient $\boldsymbol{G}(t)$ and $\text{diag}(\sigma^2_{N_{k,1}}(t), ..., \sigma^2_{N_{k,P}}(t))$ is the variance of the gradient noise $\boldsymbol{N}_k(t)$.

Our main result is

Theorem 2:

$$\mathcal{R}_\star(D) = \mathcal{R}_N(D). \tag{19}$$

Remark 3: As will become clear in the next section, parameter D_p can be interpreted as the distortion of gradient estimator at dimension p, $r_{k,p}$ can be interpreted as the rate the k-th edge device spends in quantizing its observation noise at dimension p and $R_{k,p}$ can be interpreted as the rate contributed by edge device k on dimension p. The set $\mathcal{F}_p(D_p)$ contains all feasible $r_{k,p}, k = 1, 2, ..., K$ such that the distortion at dimension p satisfying D_p. The set $\mathcal{F}_p(D_p)$ is empty when $D_p \leq (\sum_{k\in\mathcal{K}}\frac{1}{\sigma^2_{N_{k,p}}})^{-1}$ and $\mathcal{R}_N(D)$ is empty when $D \leq \sum_{p=1}^{P}(\sum_{k\in\mathcal{K}}\frac{1}{\sigma^2_{N_{k,p}}})^{-1}$. Moreover, $\mathcal{F}_p(D_p)$ and $\mathcal{R}_N(D)$ contains more elements as D_p and D increase, respectively. Given the $(r_{1,p}, r_{2,p}, ..., r_{K,p}) \in \mathcal{F}_p(D_p), p = 1, 2, ..., P$, the set $\mathcal{R}_p(r_{1,p}, r_{2,p}, ..., r_{K,p})$ can be interpreted as the sub-rate region of dimension p. The p-th sub-rate region $\mathcal{R}_p(r_{1,p}, r_{2,p}, ..., r_{K,p})$ contains less elements as $r_{k,p}$ increases. For $\forall(R_{1,p}, R_{2,p}, ..., R_{K,p}) \in \mathcal{R}_p(r_{1,p}, r_{2,p}, ..., r_{K,p}), p = 1, 2, ..., P$, there is an element in rate region $\mathcal{R}_N(D)$ corresponding to it, satisfying $R_k = \sum_{p=1}^{P} R_{k,p}, k = 1, 2, ..., K$. Since $\mathcal{R}_\star(D)$ must be convex, we can conclude that $\mathcal{R}_N(D)$ is also convex. This can also be directly inferred by noting that given the distortion $(D_1, D_2, ..., D_P)$, the rate region $\mathcal{R}_N(D)$ is a convex set.

The converse and direct parts of the proof of this theorem will be given in Sections V. We prove this theorem from the standpoint of multiterminal rate-distortion theory.

V. Proof

A. Achievability of Theorem 2

The achievable proof is based on the Berger-Tung inner bound developed by Berger [38].

Lemma 1 (Berger-Tung inner bound): If we can find auxiliary random variables $\boldsymbol{U}_1, \boldsymbol{U}_2, ..., \boldsymbol{U}_K$ such that

$$\boldsymbol{U}_1, \boldsymbol{U}_2, ..., \boldsymbol{U}_K | \boldsymbol{G}_1, \boldsymbol{G}_2, ..., \boldsymbol{G}_K \sim p(u_1|g_1)p(u_2|g_2)...p(u_K|g_K) \tag{20}$$

and decoding function $\hat{\boldsymbol{G}}(\boldsymbol{U}_1, \boldsymbol{U}_2, ..., \boldsymbol{U}_K)$ such that

$$\mathbb{E}\|\boldsymbol{G} - \hat{\boldsymbol{G}}(\boldsymbol{U}^K)\|^2 \leq D, \tag{21}$$

then the following rate region is achievable

$$\sum_{k \in \mathcal{A}} R_k \geq I(\boldsymbol{G}_{\mathcal{A}}; \boldsymbol{U}_{\mathcal{A}} | \boldsymbol{U}_{\mathcal{A}^c}), \forall \mathcal{A} \subseteq \{1, 2, ..., K\}. \tag{22}$$

Now, let us consider the Berger-Tung scheme in FL. For each edge device, we define the auxiliary random variable $\boldsymbol{U}_k = \boldsymbol{G}_k + \boldsymbol{V}_k$, where $\boldsymbol{V}_k \sim \mathcal{N}(0, \mathrm{diag}(\sigma^2_{V_{k,1}}, ..., \sigma^2_{V_{k,P}}))$ are independently distributed and independent of $\boldsymbol{G}_k$ and $\boldsymbol{G}$. Parameters $\{\sigma^2_{V_{k,p}}\}$ are determined in terms of the target distortion D. After recovering $\boldsymbol{U}^{nK}$, the decoder reconstructs $\hat{\boldsymbol{G}}^n$ by applying the following weighted averaging function component-wise

$$\hat{g} = \psi_K(\boldsymbol{U}^K) \triangleq [\sum_{k=1}^{K} \alpha_{k,1} u_{k,1}, \sum_{k=1}^{K} \alpha_{k,2} u_{k,2}, ..., \sum_{k=1}^{K} \alpha_{k,P} u_{k,P}], \tag{23}$$

where $\alpha_{k,p} = \frac{(\sigma^2_{N_{k,p}} + \sigma^2_{V_{k,p}})^{-1}}{\sum_{k=1}^{K} (\sigma^2_{N_{k,p}} + \sigma^2_{V_{k,p}})^{-1}}$. Note that $\hat{\boldsymbol{G}}$ is unbiased estimator of $\boldsymbol{G}$ and the variance of $\hat{\boldsymbol{G}}$ is $\mathbb{E}\|\boldsymbol{G} - \hat{\boldsymbol{G}}(\boldsymbol{U}^K)\|^2$. We set this equal to the target distortion D, which is given by

$$D = \sum_{p=1}^{P} D_p, \tag{24}$$

where

$$D_p = \left(\sum_{k=1}^{K} \frac{1}{\sigma^2_{N_{k,p}} + \sigma^2_{V_{k,p}}} \right)^{-1}, p = 1, 2, ..., P. \tag{25}$$

Let us define

$$r_{k,p} \triangleq I(G_{k,p}; U_{k,p} | G_p) = \frac{1}{2} \log{(1 + \frac{\sigma^2_{N_{k,p}}}{\sigma^2_{V_{k,p}}})}, k = 1, 2, ..., K, p = 1, 2, ..., P. \tag{26}$$

We can interpret $r_{k,p}$ as the rate the k-th device spends in quantizing its p-th dimension of the observation noise. We will use $r_{k,p}$'s as the parameters instead of $\sigma^2_{V_{k,p}}$. Note that for any choice of $(r_{k,p} \geq 0, k = 1, 2, ..., K, p = 1, 2, ..., P)$, we can find a corresponding $(\sigma^2_{V_{k,p}} \geq 0, k =$

$1, 2, ..., K, p = 1, 2, ..., P)$ and therefore, a set of auxiliary random variables. Then we can rewrite (25) in terms of $r_{k,p}$'s as

$$\frac{1}{D_p} = \sum_{k=1}^{K} \frac{1 - \exp(-2r_{k,p})}{\sigma^2_{N_{k,p}}}, p = 1, 2, ..., P \tag{27}$$

as desired.

From the Berger-Tung inner bound, $(R_1, ..., R_K)$ are achievable if for all non-empty $\mathcal{A} \subseteq \{1, 2, ..., K\}$

$$\sum_{k\in\mathcal{A}} R_k \geq I(\boldsymbol{G}_{\mathcal{A}}; \boldsymbol{U}_{\mathcal{A}}|\boldsymbol{U}_{\mathcal{A}^c}) \tag{28}$$

$$= I(\boldsymbol{G}_{\mathcal{A}}, \boldsymbol{G}; \boldsymbol{U}_{\mathcal{A}}|\boldsymbol{U}_{\mathcal{A}^c}) \tag{29}$$

$$= I(\boldsymbol{G}; \boldsymbol{U}_{\mathcal{A}}|\boldsymbol{U}_{\mathcal{A}^c}) + I(\boldsymbol{G}_{\mathcal{A}}; \boldsymbol{U}_{\mathcal{A}}|\boldsymbol{G}) \tag{30}$$

$$= I(\boldsymbol{G}; \boldsymbol{U}_{\mathcal{A}}|\boldsymbol{U}_{\mathcal{A}^c}) + \sum_{p=1}^{P}\sum_{k\in\mathcal{A}} r_{k,p} \tag{31}$$

$$= h(\boldsymbol{G}|\boldsymbol{U}_{\mathcal{A}^c}) - h(\boldsymbol{G}|\boldsymbol{U}) + \sum_{p=1}^{P}\sum_{k\in\mathcal{A}} r_{k,p} \tag{32}$$

$$= \sum_{p=1}^{P} \left[-\frac{1}{2} \log \left(\frac{1}{\sigma^2_{X_p}} + \sum_{k\in\mathcal{A}^c} (\frac{1}{\sigma^2_{N_{k,p}} + \sigma^2_{V_{k,p}}})^{-1} \right) + \frac{1}{2} \log (\frac{1}{\sigma^2_{X_p}} + \sum_{k\in\mathcal{K}} (\frac{1}{\sigma^2_{N_{k,p}} + \sigma^2_{V_{k,p}}})^{-1}) + \sum_{k\in\mathcal{A}} r_{k,p} \right] \tag{33}$$

$$= \sum_{p=1}^{P} \left(-\frac{1}{2} \log \left(\frac{1}{\sigma^2_{X_p}} + \sum_{k\in\mathcal{A}^c} \frac{1 - \exp(-2r_{k,p})}{\sigma^2_{N_{k,p}}} \right) + \frac{1}{2} \log (\frac{1}{\sigma^2_{X_p}} + \frac{1}{D}) + \sum_{k\in\mathcal{A}} r_{k,p} \right) \tag{34}$$

$$= \sum_{k\in\mathcal{A}} \sum_{p=1}^{P} R_{k,p}, \tag{35}$$

Equations (27) and (35) together concludes the achievable proof.

B. Converse of Theorem 2

Our converse proof is inspired by Oohama's converse in [27]. Suppose we achieve distortion $D = \sum_{p=1}^{P} D_p$, where $D_p = D^n(G_p^n, \hat{G}_p^n)$. Let $\boldsymbol{C}_{\mathcal{K}} = (C_1, C_2, ..., C_K)$ denote all the messages produced by the edge devices after observing an n-block. Let us define

$$r_{k,p} \triangleq \frac{1}{n} I(G^n_{k,p}; C_k|G^n_p). \tag{36}$$

For any $\mathcal{A} \subseteq \{1, 2, ..., K\}$,

$$\begin{aligned}
\sum_{k\in\mathcal{A}} R_k \geq \sum_{k\in\mathcal{A}} \frac{1}{n} H(C_k) &\geq \frac{1}{n} H(\boldsymbol{C}_{\mathcal{A}}) && (37)\\
&\geq \frac{1}{n} H(\boldsymbol{C}_{\mathcal{A}} | \boldsymbol{C}_{\mathcal{A}^c}) && (38)\\
&\geq \frac{1}{n} I(\boldsymbol{G}^n_{\mathcal{A}}; \boldsymbol{C}_{\mathcal{A}} | \boldsymbol{C}_{\mathcal{A}^c}) && (39)\\
&= \frac{1}{n} I(\boldsymbol{G}^n, \boldsymbol{G}^n_{\mathcal{A}}; \boldsymbol{C}_{\mathcal{A}} | \boldsymbol{C}_{\mathcal{A}^c}) && (40)\\
&= \frac{1}{n} I(\boldsymbol{G}^n; \boldsymbol{C}_{\mathcal{A}} | \boldsymbol{C}_{\mathcal{A}^c}) + \frac{1}{n} I(\boldsymbol{G}^n_{\mathcal{A}}; \boldsymbol{C}_{\mathcal{A}} | \boldsymbol{C}_{\mathcal{A}^c}, \boldsymbol{G}^n) && (41)\\
&= \frac{1}{n} I(\boldsymbol{G}^n; \boldsymbol{C}_{\mathcal{A}} | \boldsymbol{C}_{\mathcal{A}^c}) + \frac{1}{n} I(\boldsymbol{G}^n; \boldsymbol{C}_{\mathcal{A}^c}) - \frac{1}{n} I(\boldsymbol{G}^n; \boldsymbol{C}_{\mathcal{A}^c}) + \frac{1}{n} I(\boldsymbol{G}^n_{\mathcal{A}}; \boldsymbol{C}_{\mathcal{A}} | \boldsymbol{G}^n) && (42)\\
&= \frac{1}{n} I(\boldsymbol{G}^n; \boldsymbol{C}_{\mathcal{K}}) - \frac{1}{n} I(\boldsymbol{G}^n; \boldsymbol{C}_{\mathcal{A}^c}) + \sum_{k\in\mathcal{A}} \frac{1}{n} I(\boldsymbol{G}^n_k; C_k | \boldsymbol{G}^n) && (43)\\
&= \sum_{p=1}^{P} \left(\frac{1}{n} I(G^n_p; \boldsymbol{C}_{\mathcal{K}}) - \frac{1}{n} I(G^n_p; \boldsymbol{C}_{\mathcal{A}^c}) + \sum_{k\in\mathcal{A}} r_{k,p} \right), && (44)
\end{aligned}$$

where (38) follows from the fact that conditioning reduces entropy, (40) follows from the fact that $\boldsymbol{G}^n - \boldsymbol{G}^n_{\mathcal{A}} - \boldsymbol{C}_{\mathcal{A}}$ are Markov chain, (41) follows from the chain rule for mutual information, (42) follows from the fact that $(\boldsymbol{G}^n_{\mathcal{A}}, \boldsymbol{C}_{\mathcal{A}}) - \boldsymbol{G}^n - \boldsymbol{C}_{\mathcal{A}^c}$ are Markov chain, (43) follows from the chain rule for mutual information and the fact that $\boldsymbol{G}^n_k - \boldsymbol{G}^n - \boldsymbol{C}_{\mathcal{K}\setminus\{k\}}$ are Markov chain.

Lemma 2: Let $\hat{X}^n$ be any unbiased estimator of X^n with distortion $D^n(X^n, \hat{X}^n) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X[i] - \hat{X}[i])^2$. We have

$$\frac{1}{n} I(X^n; \hat{X}^n) \geq \frac{1}{2} \log(1 + \frac{\sigma_X^2}{D^n(X^n, \hat{X}^n)}). \quad (45)$$

Proof:

$$\begin{aligned}
\frac{1}{n} I(X^n; \hat{X}^n) &= \frac{1}{n} h(X^n) - \frac{1}{n} h(X^n | \hat{X}^n) && (46)\\
&= \frac{1}{n} h(X^n) - \frac{1}{n} \sum_{i=1}^{n} h(X[i] | X^{i-1} \hat{X}^n) && (47)\\
&\geq \frac{1}{n} h(X^n) - \frac{1}{n} \sum_{i=1}^{n} h(X[i] | \hat{X}[i]) && (48)\\
&\geq \frac{1}{2} \log 2\pi e \sigma_X^2 - \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \log 2\pi e \sigma^2_{X[i]|\hat{X}[i]}. && (49)
\end{aligned}$$

The conditional variance of variable X given the unbiased estimator $\hat{X}$ is given by

$$\sigma^2_{X|\hat{X}} = \mathbb{E}\left[X - \mathbb{E}(X|\hat{X})\right]^2 \tag{50}$$

$$\leq \min_a \mathbb{E}(X - a\hat{X})^2 \tag{51}$$

$$= \min_a \mathbb{E}\left[\mathbb{E}\left((X - a\hat{X})^2|X\right)\right] \tag{52}$$

$$= \min_a \mathbb{E}\left[\mathbb{E}^2(X - a\hat{X}|X) + \text{Var}(X - a\hat{X}|X)\right] \tag{53}$$

$$= \min_a \mathbb{E}\left\{(X - aX)^2 + \mathbb{E}\left[(X - a\hat{X}) - \mathbb{E}(X - a\hat{X}|X)\right]\right\} \tag{54}$$

$$= \min_a \left[(1-a)^2\sigma^2_X + a^2\mathbb{E}(X - \hat{X})^2\right] \tag{55}$$

$$= (\frac{1}{\sigma^2_X} + \frac{1}{\mathbb{E}(X - \hat{X})^2})^{-1}. \tag{56}$$

where (54) and (55) follows from the fact that $\mathbb{E}(\hat{X}|X) = X$. Substituting (56) to (49), we have

$$\frac{1}{n}I(X^n; \hat{X}^n) \geq \frac{1}{2}\log 2\pi e\sigma^2_X + \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\log\frac{1}{2\pi e}(\frac{1}{\sigma^2_X} + \frac{1}{\mathbb{E}(X[i] - \hat{X}[i])^2}) \tag{57}$$

$$\geq \frac{1}{2}\log 2\pi e\sigma^2_X + \frac{1}{2}\log\frac{1}{2\pi e}(\frac{1}{\sigma^2_X} + \frac{1}{D^n(X^n, \hat{X}^n)}) \tag{58}$$

$$= \frac{1}{2}\log(1 + \frac{\sigma^2_X}{D^n(X^n, \hat{X}^n)}), \tag{59}$$

where (58) follows from Jensens inequality. The proof of Lemma 2 is completed. ■

We have a simple lower-bound for the first term $\frac{1}{n}I(G^n_p; \boldsymbol{C}_{\mathcal{K}})$ based on Lemma 2

$$\frac{1}{n}I(G^n_p; \boldsymbol{C}_{\mathcal{K}}) \geq \frac{1}{n}I(G^n_p; \hat{G}^n_p) \tag{60}$$

$$\geq \frac{1}{2}\log(1 + \frac{\sigma^2_{X_p}}{D^n(G^n_p, \hat{G}^n_p)}) \tag{61}$$

$$= \frac{1}{2}\log(1 + \frac{\sigma^2_{X_p}}{D_p}), \tag{62}$$

where (60) follows from the data-processing inequality and (61) follows from Lemma 2.

To bound the second term, we will need the following lemma

Lemma 3: Let $\mathcal{A} \subseteq \{1, 2, ..., K\}$. Then

$$\frac{1}{\sigma^2_{X_p}}\exp\left(\frac{2}{n}I(G^n_p; \boldsymbol{C}_{\mathcal{A}})\right) \leq \frac{1}{\sigma^2_{X_p}} + \sum_{k\in\mathcal{A}}\frac{1 - \exp(-2r_{k,p})}{\sigma^2_{N_{k,p}}}, p = 1, 2, ..., P. \tag{63}$$

Proof: see Lemma 3.1 in [36]. ■

Using the bounds from (62) and (63) in (44), for all $\mathcal{A} \subseteq \{1, 2, ..., K\}$

$$\sum_{k\in\mathcal{A}} R_k \geq \sum_{p=1}^{P} \left(\frac{1}{2}\log(1+\frac{\sigma^2_{X_p}}{D_p}) - \frac{1}{2}\log\left(1+\sigma^2_{X_p}\sum_{k\in\mathcal{A}^c}\frac{1-\exp(-2r_{k,p})}{\sigma^2_{N_{k,p}}}\right) + \sum_{k\in\mathcal{A}} r_{k,p} \right) \tag{64}$$

$$= \sum_{p=1}^{P} \left(-\frac{1}{2}\log\left(\frac{1}{\sigma^2_{X_p}} + \sum_{k\in\mathcal{A}^c}\frac{1-\exp(-2r_{k,p})}{\sigma^2_{N_{k,p}}}\right) + \frac{1}{2}\log\,(\frac{1}{\sigma^2_{X_p}}+\frac{1}{D_p}) + \sum_{k\in\mathcal{A}} r_{k,p} \right). \tag{65}$$

Substituting (62) to (63) with $\mathcal{A} = \mathcal{K}$, we have the condition

$$\sum_{k=1}^{K} \frac{1-\exp(-2r_{k,p})}{\sigma^2_{N_{k,p}}} \geq \frac{1}{D_p}, p = 1, 2, ..., P. \tag{66}$$

Along with the non-negativity constraints on $r_{k,p}$'s, (65) and (66) define an outer bound for $\mathcal{R}_\star(D)$. It is easy to show that replacing the inequality in (66) with an equality will not change this outer bound. Thus we have $\mathcal{R}_\star(D) \subseteq \mathcal{R}_N(D)$.

VI. Explicit rate region

The rate region in section IV is implicit. To facilitate analysis the communication efficiency of FL, we will derive the explicit rate region in this section.

A. Sum-rate distortion function

In most cases of analyzing communication efficiency of FL, we only care about the sum-rate distortion function. Sum-rate distortion function are defined by

$$R_{sum}(D) \triangleq \min_{(R_1,R_2,...,R_K)\in\mathcal{R}_\star(D)} \sum_{k\in\mathcal{K}} R_k. \tag{67}$$

Recall that the rate region $\mathcal{R}_\star(D)$ in (16), we have

$$\min_{(R_1,R_2,...,R_K)\in\mathcal{R}_\star(D)} \sum_{k\in\mathcal{K}} R_k = \min_{D_p}\min_{r_{k,p}}\min_{R_{k,p}} \sum_{k\in\mathcal{K}}\sum_{p=1}^{P} R_{k,p}. \tag{68}$$

Note that $\sum_{k\in\mathcal{K}} R_{k,p} \geq \sum_{k\in\mathcal{K}} r_{k,p} + \frac{1}{2}\log(1+\frac{\sigma^2_{X_p}}{D_p})$, therefore we can explicitly compute the sum-rate distortion function by solving the following optimization problem

$$\mathcal{P}_1 : \quad \min_{D_p, r_{k,p}} \quad \sum_{p=1}^{P}\left[\sum_{k\in\mathcal{K}} r_{k,p} + \frac{1}{2}\log(1+\frac{\sigma^2_{X_p}}{D_p})\right] \tag{69a}$$

$$s.t. \qquad \sum_{k\in\mathcal{K}} \frac{1-\exp(-2r_{k,p})}{\sigma^2_{N_{k,p}}} \geq \frac{1}{D_p}, \quad p = 1, 2, ..., P \tag{69b}$$

$$\sum_{p=1}^{P} D_p \leq D. \tag{69c}$$

Obviously, it is a convex optimization problem, and we can easily know that the inequality constraints (69b) and (69c) must be active. Therefore it can be solved using Lagrange minimization. For the method of Lagrange multipliers,

$$L(\boldsymbol{r}, \boldsymbol{D}, \boldsymbol{\lambda}) = \sum_{p=1}^{P} \left[\sum_{k \in \mathcal{K}} r_{k,p} + \frac{1}{2} \log(1 + \frac{\sigma_{X_p}^2}{D_p}) \right] + \sum_{p=1}^{P} \lambda_p (\sum_{k \in \mathcal{K}} \frac{1 - \exp(-2r_{k,p})}{\sigma_{N_{k,p}}^2} - \frac{1}{D_p}) + \lambda_{P+1} (\sum_{p=1}^{P} D_p - D). \tag{70}$$

First, we calculate the partial derivative of $L(\boldsymbol{r}, \boldsymbol{D}, \boldsymbol{\lambda})$ with respect to $r_{k,p}, k \in \mathcal{K}, p = 1, 2, ..., P$ and $\lambda_p, p = 1, 2, ..., P$,

$$\frac{\partial L}{\partial r_{k,p}} = 1 + \lambda_p \frac{2 \exp(-2r_{k,p})}{\sigma_{N_{k,p}}^2}, k = 1, 2, ..., K, p = 1, 2, ..., P \tag{71}$$

$$\frac{\partial L}{\partial \lambda_p} = \sum_{k \in \mathcal{K}} \frac{1 - \exp(-2r_{k,p})}{\sigma_{N_{k,p}}^2} - \frac{1}{D_p}, p = 1, 2, ..., P. \tag{72}$$

The Lagrange cost is minimized when $\partial L / \partial r_{k,p} = 0$ for all $k = 1, 2, ..., K, p = 1, 2, ..., P$ and $\partial L / \partial \lambda_p = 0$ for all $p = 1, 2, ..., P$, then we have the optimal λ_p^* and $r_{k,p}^*$ at any given D_p as,

$$\lambda_p^* = \frac{K}{2} (\frac{1}{D_p} - \sum_{k \in \mathcal{K}} \frac{1}{\sigma_{N_{k,p}}^2})^{-1}, p = 1, 2, ..., P \tag{73}$$

$$r_{k,p}^* = -\frac{1}{2} \log \left[\frac{\sigma_{N_{k,p}}^2}{K} (\sum_{k \in \mathcal{K}} \frac{1}{\sigma_{N_{k,p}}^2} - \frac{1}{D_p}) \right], k = 1, 2, ..., K, p = 1, 2, ..., P. \tag{74}$$

Second, we calculate the partial derivative of $L(\boldsymbol{r}, \boldsymbol{D}, \boldsymbol{\lambda})$ with respect to $D_p, p = 1, 2, ..., P$ and λ_{P+1}

$$\frac{\partial L}{\partial D_p} = \frac{\frac{\sigma_{X_p}^2}{D_p^2}}{2(1 + \frac{\sigma_{X_p}^2}{D_p})} + \frac{\lambda_p}{D_p^2} + \lambda_{P+1}, p = 1, 2, ..., P \tag{75}$$

$$\frac{\partial L}{\partial \lambda_{P+1}} = \sum_{p=1}^{P} D_p - D. \tag{76}$$

Substituting λ_p^* in (73) back to (75), the Lagrange cost is minimized when $\frac{\partial L}{\partial D_p} = 0$ for all $p = 1, 2, ..., P$ and $\frac{\partial L}{\partial \lambda_{P+1}} = 0$, then the optimal D_p^* and λ_{P+1}^* satisfy

$$\lambda_{P+1}^* = \frac{K}{2D_p^{*2}} (\sum_{k \in \mathcal{K}} \frac{1}{\sigma_{N_{k,p}}^2} - \frac{1}{D_p^*})^{-1} + \frac{1}{2D_p^{*2}} (\frac{1}{\sigma_{X_p}^2} + \frac{1}{D_p^*})^{-1}, p = 1, 2, ..., P \tag{77}$$

$$\sum_{p=1}^{P} D_p^* = D. \tag{78}$$

We find that λ_{P+1}^* decreases monotonically with D_p^* for all $p = 1, 2, ..., P$. The goal is to find the optimal λ_{P+1}^* such that the total distortion equals the given value of D in the problem.

B. Sum-rate distortion function in special cases

In this subsection, we will provide closed-form sum-rate distortion functions in special cases. In particular, when the global gradient $\boldsymbol{G}(t)$ and the local gradients $\boldsymbol{G}_k(t)$ are identical over dimension p's, i.e., $\sigma_{X_1}^2 = ... = \sigma_{X_P}^2 = \frac{\sigma_X^2}{P}$ and $\sigma_{N_{k,1}}^2 = ... = \sigma_{N_{k,P}}^2 = \frac{\sigma_{N_k}^2}{P}$ for all $k = 1, 2, ..., K$. The optimal solution $\boldsymbol{r}^*, \boldsymbol{D}^*, \boldsymbol{\lambda}^*$ in this case can be reduced directly from the solution of problem P1. Then, we can obtain a closed-form formula of the sum-rate-distortion function shown in the following corollary,

Corollary 1: For every $D > 0$, when the gradients elements are identical over dimension p's,

$$R_{sum}(D) = P\left[-\frac{1}{2}\sum_{k\in\mathcal{K}} \log\left(\frac{\sigma_{N_k}^2}{K}(\sum_{k\in\mathcal{K}} \frac{1}{\sigma_{N_k}^2} - \frac{1}{D})\right) + \frac{1}{2}\log(1 + \frac{\sigma_X^2}{D})\right]. \tag{79}$$

Proof: Since $\sigma_{N_{k,1}}^2 = ... = \sigma_{N_{k,P}}^2 = \frac{\sigma_{N_k}^2}{P}$ for all $k = 1, 2, ..., K$ and $\sigma_{X_1}^2 = ... = \sigma_{X_P}^2 = \frac{\sigma_X^2}{P}$, based on (77) we have $D_p^* = \frac{D}{P}, p = 1, 2, ..., P$. Substituting $D_p^* = \frac{D}{P}, p = 1, 2, ..., P$ back to (74), we have

$$r_{k,p}^* = -\frac{1}{2}\log\left[\frac{\sigma_{N_k}^2}{K}(\sum_{k\in\mathcal{K}} \frac{1}{\sigma_{N_k}^2} - \frac{1}{D})\right], \tag{80}$$

Substituting $D_p^*, p = 1, 2, ..., P$ and $r_{k,p}^*, k = 1, 2, ..., K, p = 1, 2, ..., P$ back to problem P1, we conclude the corollary 1. ■

Corollary 2: For every $D > 0$, when the gradient elements are identical over dimension p's and local gradients are identical over device k's, i.e., $\sigma_{N_1}^2 = ... = \sigma_{N_K}^2 = \sigma_N^2$.

$$R_{sum}(D) = P\left[-\frac{K}{2}\log\left(1 - \frac{\sigma_N^2}{KD}\right) + \frac{1}{2}\log(1 + \frac{\sigma_X^2}{D})\right]. \tag{81}$$

Corollary 3: For every $D > 0$, gradient elements are identical over dimension p's and local gradients are identical over device k's, and $K \to \infty$.

$$R_{sum}(D) = P\left[\frac{\sigma_N^2}{2D} + \frac{1}{2}\log(1 + \frac{\sigma_X^2}{D})\right]. \tag{82}$$

C. Explicit rate region

To study user scheduling for communication-efficient FL, we can also explicitly compute the rate region $\mathcal{R}_\star(D)$ by solving the following optimization problem

$$\min_{(R_1,R_2,...,R_K)\in\mathcal{R}_\star(D)}\sum_{k\in\mathcal{K}}\alpha_k R_k=\min_{D_p}\min_{r_{k,p}}\min_{R_{k,p}}\sum_{p=1}^{P}\sum_{k\in\mathcal{K}}\alpha_k R_{k,p}. \tag{83}$$

for all choices of $(\alpha_1,\alpha_2,...,\alpha_K)\in\mathbb{R}_+^K$. Without loss of generality assume that $(\alpha_1\geq\alpha_2\geq ...\geq\alpha_K)$.

Lemma 4: Given D_p, $r_{k,p}$ and $(\alpha_1\geq\alpha_2\geq...\geq\alpha_K)$ the optimal choice of $(R_{1,p},R_{2,p},...,R_{K,p})\in\mathcal{R}_p(r_{1,p},r_{2,p},...,r_{K,p})$ for minimizing $\sum_{k\in\mathcal{K}}\alpha_k R_{k,p}$ is

$$R_{k,p}=r_{k,p}+\frac{1}{2}\log\left(\frac{1}{\sigma^2_{X_p}}+\sum_{i=k}^{K}\frac{1-\exp(-2r_{i,p})}{\sigma^2_{N_{i,p}}}\right)-\frac{1}{2}\log\left(\frac{1}{\sigma^2_{X_p}}+\sum_{i=k+1}^{K}\frac{1-\exp(-2r_{i,p})}{\sigma^2_{N_{i,p}}}\right). \tag{84}$$

Proof: Let $R_{k,p}=I(G_{k,p};U_{k,p}|U_{k+1,p},...,U_{K,p})$, $k=1,2,...,K$. We first prove that $(R_{1,p},R_{2,p},...,R_{K,p})$ lies in $\mathcal{R}_p(r_{1,p},r_{2,p},...,r_{K,p})$. Second, we prove that $(R_{1,p},R_{2,p},...,R_{K,p})$ is the optimal choice in $\mathcal{R}_p(r_{1,p},r_{2,p},...,r_{K,p})$ for minimizing $\sum_{k\in\mathcal{K}}\alpha_k R_{k,p}$. Finally, we show that $I(G_{k,p};U_{k,p}|U_{k+1,p},...,U_{K,p})$ is equal to the equation (84).

First, for all $\mathcal{A}\subseteq\mathcal{K}$, we have

$$\begin{aligned}\sum_{k\in\mathcal{A}}R_{k,p}&=\sum_{k\in\mathcal{A}}I(G_{k,p};U_{k,p}|U_{k+1,p},...,U_{K,p}) &(85)\\ &\geq\sum_{k\in\mathcal{A}}I(G_{k,p};U_{k,p}|U_{\mathcal{A}^c\cup\{k+1,...,K\},p}) &(86)\\ &=I(G_{\mathcal{A},p};U_{\mathcal{A},p}|U_{\mathcal{A}^c,p}). &(87)\end{aligned}$$

Based on the definition of $\mathcal{R}_p(r_{1,p},r_{2,p},...,r_{K,p})$, we can easily know that $(R_{1,p},R_{2,p},...,R_{K,p})$ is in $\mathcal{R}_p(r_{1,p},r_{2,p},...,r_{K,p})$.

Then, to prove that $(R_{1,p},R_{2,p},...,R_{K,p})$ is the optimal choice in $\mathcal{R}_p(r_{1,p},r_{2,p},...,r_{K,p})$ is equal to prove that for all $(R^{'}_{1,p},R^{'}_{2,p},...,R^{'}_{K,p})\in\mathcal{R}_p(r_{1,p},r_{2,p},...,r_{K,p})$ we have $\sum_{k\in\mathcal{K}}\alpha_k R_{k,p}\leq\sum_{k\in\mathcal{K}}\alpha_k R^{'}_{k,p}$. For all $(R^{'}_{1,p},R^{'}_{2,p},...,R^{'}_{K,p})\in\mathcal{R}_p(r_{1,p},r_{2,p},...,r_{K,p})$, based on the definition of $\mathcal{R}_p(r_{1,p},r_{2,p},...,r_{K,p})$, we have

$$\sum_{i=1}^{k}R^{'}_{i,p}\geq I(G_{1,p},...,G_{k,p};U_{1,p},...,U_{k,p}|U_{k+1,p},...,U_{K,p})=\sum_{i=1}^{k}R_{i,p}, k=1,2,...,K. \tag{88}$$

To compare $\sum_{k\in\mathcal{K}}\alpha_k R_{k,p}$ and $\sum_{k\in\mathcal{K}}\alpha_k R'_{k,p}$, we have

$$\sum_{k\in\mathcal{K}}\alpha_k R'_{k,p}-\sum_{k\in\mathcal{K}}\alpha_k R_{k,p} \tag{89}$$

$$=\sum_{k\in\mathcal{K}}\alpha_k(R'_{k,p}-R_{k,p}) \tag{90}$$

$$=\sum_{k\in\mathcal{K}}\left[(\alpha_k-\alpha_{k+1})+(\alpha_{k+1}-\alpha_{k+2})+...+(\alpha_{K-1}-\alpha_K)+(\alpha_K)\right](R'_{k,p}-R_{k,p}) \tag{91}$$

$$=\sum_{k=1}^{K-1}(\alpha_k-\alpha_{k+1})(\sum_{i=1}^{k}R_{i,p}-\sum_{i=1}^{k}R'_{i,p})+\alpha_K(\sum_{i\in\mathcal{K}}R_{i,p}-\sum_{i\in\mathcal{K}}R'_{i,p}) \tag{92}$$

$$\geq 0. \tag{93}$$

The inequality in (93) holds as $\alpha_1\geq\alpha_2\geq...\geq\alpha_K$ and $\sum_{i=1}^{k}R'_{i,p}\geq\sum_{i=1}^{k}R_{i,p}$ for all $k=1,2,...,K$. This indicates that for all $(R'_{1,p},R'_{2,p},...,R'_{K,p})\in\mathcal{R}_p(r_{1,p},r_{2,p},...,r_{K,p})$, $(R_{1,p},R_{2,p},...,R_{K,p})$ is a better solution than $(R'_{1,p},R'_{2,p},...,R'_{K,p})$.

Therefore, the optimal choice of $(R_{1,p},R_{2,p},...,R_{K,p})$ is

$$R_{k,p}=I(G_{k,p};U_{k,p}|U_{k+1,p},...,U_{K,p}) \tag{94}$$

$$=I(G_{1,p}...G_{k,p};U_{1,p}...U_{k,p}|U_{k+1,p},...,U_{K,p})-I(G_{1,p}...G_{k-1,p};U_{1,p}...U_{k-1,p}|U_{k,p},...,U_{K,p}) \tag{95}$$

$$=\sum_{i\in\{1,...,k\}}r_{i,p}+\frac{1}{2}\log\left(\frac{1}{\sigma^2_{X_p}}+\frac{1}{D_p}\right)-\frac{1}{2}\log\left(\frac{1}{\sigma^2_{X_p}}+\sum_{i\in\{1,...,k\}^c}\frac{1-\exp(-2r_{i,p})}{\sigma^2_{N_{i,p}}}\right) \tag{96}$$

$$-\left[\sum_{i\in\{1,...,k-1\}}r_{i,p}+\frac{1}{2}\log\left(\frac{1}{\sigma^2_{X_p}}+\frac{1}{D_p}\right)-\frac{1}{2}\log\left(\frac{1}{\sigma^2_{X_p}}+\sum_{i\in\{1,...,k-1\}^c}\frac{1-\exp(-2r_{i,p})}{\sigma^2_{N_{i,p}}}\right)\right] \tag{97}$$

$$=r_{k,p}+\frac{1}{2}\log\left(\frac{1}{\sigma^2_{X_p}}+\sum_{i=k}^{K}\frac{1-\exp(-2r_{i,p})}{\sigma^2_{N_{i,p}}}\right)-\frac{1}{2}\log\left(\frac{1}{\sigma^2_{X_p}}+\sum_{i=k+1}^{K}\frac{1-\exp(-2r_{i,p})}{\sigma^2_{N_{i,p}}}\right). \tag{98}$$

The proof of Lemma 4 is completed. ■

Based on Lemma 4, we can rewrite the optimization in terms of parameters $z_{k,p} = \frac{1-\exp(-2r_{k,p})}{\sigma^2_{N_{k,p}}}$ as

$$\mathcal{P}_2: \quad \min \quad \sum_{p=1}^{P}\sum_{k=1}^{K}\alpha_k\left[-\frac{1}{2}\log\left(1-\sigma^2_{N_{k,p}}z_{k,p}\right)+\frac{1}{2}\log\left(\frac{1}{\sigma^2_{X_p}}+\sum_{i=k}^{K}z_{i,p}\right)-\frac{1}{2}\log\left(\frac{1}{\sigma^2_{X_p}}+\sum_{i=k+1}^{K}z_{i,p}\right)\right] \tag{99a}$$

$$s.t. \qquad \sum_{k\in\mathcal{K}} z_{k,p} \geq \frac{1}{D_p}, \quad p=1,2,...,P \tag{99b}$$

$$\sum_{p=1}^{P} D_p \leq D. \tag{99c}$$

It is obvious that it is a convex optimization problem and the inequality constraints (99b) and (99c) are active, which can be solved using Lagrange minimization. We calculate the partial derivative of $L(\boldsymbol{z},\boldsymbol{D},\boldsymbol{\lambda})$,

$$\frac{\partial L}{\partial z_{k,p}} = \frac{\alpha_k\sigma^2_{N_{k,p}}}{2(1-\sigma^2_{N_{k,p}}z_{k,p})}+\sum_{i=1}^{k}\frac{\alpha_i-\alpha_{i-1}}{2}\left(\frac{1}{\sigma^2_{X_p}}+\sum_{j=i}^{K}z_{j,p}\right)^{-1}+\lambda_p, k=1,2,...,K, p=1,2,...,P \tag{100}$$

$$\frac{\partial L}{\partial D_p} = \frac{\lambda_p}{D_p^2}+\lambda_{P+1}, p=1,2,...,P \tag{101}$$

$$\frac{\partial L}{\partial \lambda_p} = \sum_{k\in\mathcal{K}} z_{k,p}-\frac{1}{D_p}, p=1,2,...,P \tag{102}$$

$$\frac{\partial L}{\partial \lambda_{P+1}} = \sum_{p=1}^{P} D_p - D. \tag{103}$$

where we set $\alpha_0 = 0$. The Lagrange cost is minimized when $\nabla_{\boldsymbol{z},\boldsymbol{D},\boldsymbol{\lambda}}L(\boldsymbol{z},\boldsymbol{D},\boldsymbol{\lambda}) = 0$.

We show the explicit rate region of a FL system with $K = 2$ devices and $P = 2$ dimensions for illustration. The results are based on the global gradient distribution given by $\sigma^2_{X_1} = 1, \sigma^2_{X_2} = 2$ and the local gradient distribution given by $\sigma^2_{N_{1,1}} = 1, \sigma^2_{N_{2,1}} = 1, \sigma^2_{N_{1,2}} = 2, \sigma^2_{N_{2,2}} = 2$. The result can be easily extended to general cases.

Fig. 8 shows the rate region $\mathcal{R}_\star(D)$ for different distortion. It is observed that given distortion D, the rate R_1 decrease with the rate R_2. In the case of distortion $D = 3$, the rate R_1 approaches infinity when the rate R_2 approaches zero, where $D_1 = 1$ and $D_2 = 2$. In the case of distortion $D = 4$, the rate R_2 is zero when the rate R_1 is larger than 1.9.

Fig. 9 depicts the distortion at 1-st dimension, D_1, as a function of parameter α_1/α_2. It is observed that the distortion D_1 is almost a constant for all the α in the case of $D = 2$. In the

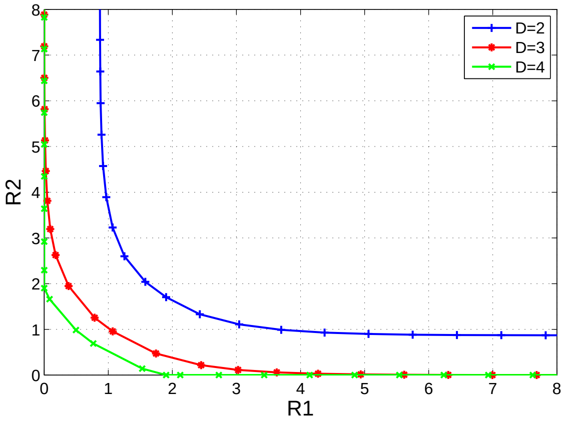


Fig. 8. Rate region $\mathcal{R}_\star(D)$ for different distortion.

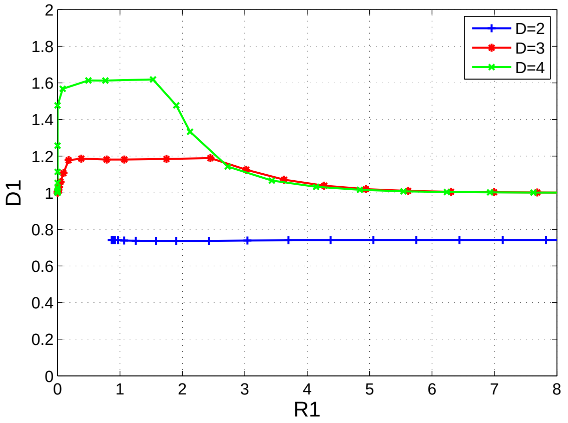


Fig. 9. The distortion at 1-st dimension, D_1, as a function of parameter α_1/α_2.

case of $D = 3$ and $D = 4$, the distortion D_1 has a platform stage when the rate R_1 and R_2 larger than zero, which is a typical case in practical.

VII. Communication Efficiency of Federated Learning

In this section, we apply the sum-rate-distortion function to analyze the communication efficiency of FL. We provide an inherent trade-off between communication cost and convergence guarantees based on the sum-rate distortion function. Our rate distortion function is quite portable, and can be applied to almost any stochastic gradient method. For illustration, we analysis communication efficiency of convex Minibatched SGD and non-convex Minibatched SGD, respectively.

A. Convex Minibatched SGD

Combining the sum-rate-distortion function given in Corollary 2 and the guarantees for Minibatched SGD algorithms on smooth, convex functions yield the following results:

Theorem 3 (Smooth Convex Optimization): Let $F, L, \boldsymbol{w}(0)$ and A be as in Theorem 1. Fix $\epsilon > 0$. Let gradient variance be $\text{diag}(\frac{\sigma_X^2(t)}{P}, ..., \frac{\sigma_X^2(t)}{P})$ and noise variance be $\text{diag}(\frac{\sigma_N^2(t)}{P}, ..., \frac{\sigma_N^2(t)}{P})$. Suppose the edge server outputs the gradient estimate $\{\hat{\boldsymbol{G}}(t)\}_{t=1}^{T}$ from K identical edge devices accessing independent stochastic gradients with variance bound D, i.e., $\mathbb{E}\|\boldsymbol{G}(t) - \hat{\boldsymbol{G}}(t)\|^2 \leq D$ for all $t = 1, 2, ..., T$, and with step size $\eta(t) = \frac{\gamma}{L+1}$, where γ is as in Theorem 1.

To guarantee the convergence rate $\mathbb{E}\left[L\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{w}(t)\right)\right] - \min_{\boldsymbol{w}\in\mathbb{R}^P} L(\boldsymbol{w}) \leq \epsilon$, the number of iterations should be satisfy

$$T \geq A^2\left(\sqrt{\frac{D}{2\epsilon^2} + \frac{L}{\epsilon}} + \sqrt{\frac{D}{2\epsilon^2}}\right)^2 = O\left(A^2 \max(\frac{2D}{\epsilon^2}, \frac{L}{\epsilon})\right) \tag{104}$$

Moreover, the lower bound of communication bits at iteration t is given by

$$P\left[-\frac{K}{2}\log\left(1 - \frac{\sigma_N^2(t)}{KD}\right) + \frac{1}{2}\log(1 + \frac{\sigma_X^2(t)}{D})\right]. \tag{105}$$

Proof: To guarantee the convergence rate $\mathbb{E}\left[L\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{w}(t)\right)\right] - \min_{\boldsymbol{w}\in\mathbb{R}^P} L(\boldsymbol{w}) \leq \epsilon$, from Theorem 1 we should have

$$A\sqrt{\frac{2D}{T}} + \frac{LA^2}{T} \leq \epsilon \Leftrightarrow T \geq A^2\left(\sqrt{\frac{D}{2\epsilon^2} + \frac{L}{\epsilon}} + \sqrt{\frac{D}{2\epsilon^2}}\right)^2. \tag{106}$$

By the definition of rate region, there are encoders $\phi^K(t)$ and decoder $\psi_K(t)$ satisfying communication bits per round less than $R_{sum}(D)(t)$ such that $\hat{\boldsymbol{G}}^n(t)$ is unbiased estimator of $\boldsymbol{G}^n(t)$, i.e., $\mathbb{E}[\hat{\boldsymbol{G}}^n(t)] = \boldsymbol{G}^n(t)$ and $D^n(\boldsymbol{G}^n(t), \hat{\boldsymbol{G}}^n(t)) \leq D$ for some n. In FL, the sequence length n is always 1 due to the fact that edge devices only upload the local gradients once per iteration. Therefore, $R_{sum}(D)(t) = P\left[-\frac{K}{2}\log\left(1 - \frac{\sigma_N^2(t)}{KD}\right) + \frac{1}{2}\log(1 + \frac{\sigma_X^2(t)}{D})\right]$ is the lower bound of communication bits at iteration t. ■

Remark 4: In the most reasonable regimes, the first term of the max in (104) will dominate the number of iterations necessary. Specifically, the number of iterations will depend linearly on the estimate gradient variance bound D. In SGD-based learning, the gradient variance $\sigma_X^2(t)$ is large at the begining, then gradually approaches to zero when the model converges. The noise variance $\sigma_N^2(t)$, on the other hand, remains approximately unchanged due to the randomness of local datasets throughout the training process. As a result, according to (105), the lower bound of communication bits per iteration decreases during the training and converges to $-P\frac{K}{2}\log\left(1-\frac{\sigma_N^2(t)}{KD}\right)$.

B. Non-convex Minibatched SGD

In many interesting applications such as neural network training, however, the objective is non-convex, where much less is known. However, there has been an interesting line of recent work which shows that Minibatched SGD at least always provably converages to a local minima, when F is smooth. For instance, by applying Theorem 2.1 in [], we immediately obtain the following communication efficiency results for FL.

Theorem 4 (Smooth Non-convex Optimization): Let $F : \mathbb{R}^P \to \mathbb{R}$ be a L-smooth (possibly non-convex) function, and Let $\boldsymbol{w}(0)$ be given. Let the iteration limit $T > 0$ be fixed and the random stoping iteration with probability mass function supported on $\{1, ..., T\}$. Suppose the edge server outputs the gradient estimate $\{\hat{\boldsymbol{G}}(t)\}_{t=1}^{T}$ from K identical edge devices accessing independent stochastic gradients with variance bound D, i.e., $\mathbb{E}\|\boldsymbol{G}(t) - \hat{\boldsymbol{G}}(t)\|^2 \leq D$ for all $t = 1, 2, ..., T$, and with step size $\eta = O(\frac{1}{L})$, then we have

$$\frac{1}{L}\mathbb{E}[\|\nabla F(\boldsymbol{w})\|^2] \leq O\left(\frac{F(\boldsymbol{w}(0)) - F^*}{N} + \frac{D}{L}\right). \tag{107}$$

Moreover, the lower bound of communication bits at iteration t is the same as in Theorem 3.

VIII. Conclusion

This paper studied the indirect multiterminal source coding problem in FL. We have characterized the rate region by showing that the Berger-Tung achievable region is tight. We also derived the explicit rate region and closed-form sum-rate distortion functions in special cases. Finally, we apply the sum-rate-distortion function to analyze the communication efficiency of FL. We provide an inherent trade-off between communication cost and convergence guarantees based on the sum-rate distortion function. In the future work, we can study the practical quantization and coding schemes based on the estimated gradient distribution to achieve the rate distortion function in FL. We can also study the communication-efficient user scheduling schemes based on the explicit rate region in FL.

REFERENCES

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*, 2017, pp. 1273–1282.

[2] J. Konen, H. B. McMahan, F. X. Yu, P. Richtrik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," 2016.

[3] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konecny, S. Mazzocchi, H. B. McMahan *et al.*, "Towards federated learning at scale: System design," *arXiv preprint arXiv:1902.01046*, 2019.

[4] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, p. 12, 2019.

[5] F. Ang, L. Chen, N. Zhao, Y. Chen, W. Wang, and F. R. Yu, "Robust federated learning with noisy communication," *IEEE Trans. Commun.*, vol. 68, no. 6, pp. 3452–3464, 2020.

[6] W. Liu, L. Chen, Y. Chen, and W. Zhang, "Accelerating federated learning via momentum gradient descent," *IEEE Trans. Parallel Distrib. Syst.*, vol. 31, no. 8, pp. 1754–1766, 2020.

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

[8] M. Tao and K. Huang, "Editorial: Special topic on machine learning at network edges," *ZTE communications*, vol. 18, no. 2, 2020.

[9] S. Bubeck, "Convex optimization: Algorithms and complexity," *Foundations and Trends in Machine Learning*, vol. 8, no. 3-4, pp. 231–357, 2015.

[10] S. Ghadimi and G. Lan, "Stochastic first- and zeroth-order methods for nonconvex stochastic programming," *SIAM Journal on Optimization*, vol. 23, no. 4, pp. 2341–2368, 2013.

[11] K. Hsieh, A. Harlap, N. Vijaykumar, D. Konomis, G. R. Ganger, P. B. Gibbons, and O. Mutlu, "Gaia: Geo-distributed machine learning approaching LAN speeds," in *14th USENIX Symposium on Networked Systems Design and Implementation (NSDI 17)*. Boston, MA: USENIX Association, Mar. 2017, pp. 629–647.

[12] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, "Adaptive federated learning in resource constrained edge computing systems," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 6, pp. 1205–1221, 2019.

[13] F. Seide, H. Fu, J. Droppo, G. Li, and D. Yu, "1-bit stochastic gradient descent and its application to data-parallel distributed training of speech DNNs," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[14] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "QSGD: Communication-efficient SGD via randomized quantization and encoding," *Advances in Neural Information Processing Systems 30*, vol. 3, pp. 1710–1721, 2018.

[15] S. Zhou, Y. Wu, Z. Ni, X. Zhou, H. Wen, and Y. Zou, "Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients," 2018.

[16] W. Wen, C. Xu, F. Yan, C. Wu, Y. Wang, Y. Chen, and H. Li, "Terngrad: Ternary gradients to reduce communication in distributed deep learning," vol. 30. Curran Associates, Inc., 2017, pp. 1509–1519.

[17] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "Xnor-net: Imagenet classification using binary convolutional neural networks," in *Computer Vision – ECCV 2016*. Springer International Publishing, 2016, pp. 525–542.

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

[19] R. Leblond, F. Pedregosa, and S. Lacoste-Julien, "ASAGA: Asynchronous Parallel SAGA," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, vol. 54. PMLR, 20–22 Apr 2017, pp. 46–54.

[20] A. F. Aji and K. Heafield, "Sparse communication for distributed gradient descent," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 440–445.

[21] Y. Lin, S. Han, H. Mao, Y. Wang, and B. Dally, "Deep gradient compression: Reducing the communication bandwidth for distributed training," in *International Conference on Learning Representations*, 2018.

[22] Y. Tsuzuku, H. Imachi, and T. Akiba, "Variance-based gradient compression for efficient distributed deep learning," *arXiv preprint arXiv:1802.06058*, 2018.

[23] T. Nishio and R. Yonetani, "Client selection for federated learning with heterogeneous resources in mobile edge," in *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, 2019, pp. 1–7.

[24] D. Liu, G. Zhu, J. Zhang, and K. Huang, "Data-importance aware user scheduling for communication-efficient edge machine learning," *IEEE Transactions on Cognitive Communications and Networking*, pp. 1–1, 2020.

[25] M. M. Amiri, D. Gunduz, S. R. Kulkarni, and H. V. Poor, "Convergence of update aware device scheduling for federated learning at the wireless edge," 2020.

[26] H. H. Yang, A. Arafa, T. Q. S. Quek, and H. Vincent Poor, "Age-based scheduling policy for federated learning in mobile edge networks," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 8743–8747.

[27] Y. Oohama, "The rate-distortion function for the quadratic gaussian CEO problem," *IEEE Transactions on Information Theory*, vol. 44, no. 3, pp. 1057–1070, 1998.

[28] S. Sra, S. Nowozin, and S. J. Wright, *Optimization for machine learning*. Mit Press, 2012.

[29] Y. Ida, T. Nakamura, and T. Matsumoto, "Domain-dependent/independent topic switching model for online reviews with numerical ratings," in *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, 2013, pp. 229–238.

[30] Y. Fukuda, Y. Ida, T. Matsumoto, N. Takemura, and K. Sakatani, "A bayesian algorithm for anxiety index prediction based on cerebral blood oxygenation in the prefrontal cortex measured by near infrared spectroscopy," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 2, pp. 1–10, 2014.

[31] H. Miyashita, T. Nakamura, Y. Ida, T. Matsumoto, and T. Kaburagi, "Nonparametric bayes-based heterogeneous drosophila melanogaster gene regulatory network inference: T-process regression," in *international conference on artificial intelligence and applications,(Innsbruck, Austria, 11–13 Feb. 2013)*, 2013, pp. 51–58.

[32] M. D. Collins and P. Kohli, "Memory bounded deep convolutional networks," *arXiv preprint arXiv:1412.1442*, 2014.

[33] S. Scardapane, D. Comminiello, A. Hussain, and A. Uncini, "Group sparse regularization for deep neural networks," *Neurocomputing*, vol. 241, pp. 81–89, 2017.

[34] S. Dieleman, J. De Fauw, and K. Kavukcuoglu, "Exploiting cyclic symmetry in convolutional neural networks," *arXiv preprint arXiv:1602.02660*, 2016.

[35] W. Chen, J. Wilson, S. Tyree, K. Weinberger, and Y. Chen, "Compressing neural networks with the hashing trick," in *International conference on machine learning*, 2015, pp. 2285–2294.

[36] V. Prabhakaran, D. Tse, and K. Ramachandran, "Rate region of the quadratic gaussian ceo problem," in *International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings.*, 2004, pp. 119–.

[37] J. Wang, J. Chen, and X. Wu, "On the sum rate of gaussian multiterminal source coding: New proofs and results," *IEEE Transactions on Information Theory*, vol. 56, no. 8, pp. 3946–3960, 2010.

[38] T. Berger, “Multiterminal source coding,” *The Information Theory Approach to Communications*, vol. 229 of CISM Courses and Lectures, pp. 171–231, 1978.