Rate Region of the Quadratic Gaussian CEO Problem under Unbiased Estimation Constraint with Application in Distributed Learning

Naifu Zhang, Meixia Tao and Jia Wang

Abstract

One of the main focus in distributed learning is the communication efficiency since the model update from each edge device at each round of the training can consist of millions to billions of parameters. To obtain the trade-off between the convergence rate and the communication cost, we need to find out what is the minimum communication cost to achieve a target variance bound of the gradient estimation at each training round, which falls into the classic rate distortion theory for the CEO problem. However, all the existing rate region results of CEO problem cannot be directly applied to distributed learning due to the fact that the biased gradients cannot guarantee the convergence of distributed learning. In this paper, we study the rate region of quadratic Gaussian CEO problem under unbiased estimation constraint and extend the rate region results to the distributed learning. Finally, we analyse the communication efficiency of convex mini-batch stochastic gradient descent (mini-batch SGD) algorithm and non-convex mini-batch SGD algorithm based on the sum-rate-distortion function, respectively.

Index Terms

Distributed learning, rate distortion theory, quadratic Gaussian CEO problem, communication efficiency.

I. Introduction

In a wide range of artificial intelligence (AI) applications such as image recognition and natural language processing, the size of training datasets has grown significantly over the years due to the growing computation and sensing capabilities of mobile devices. It is becoming crucial to train big machine learning models in a distributed fashion, in which large-scale datasets are distributed

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.).

over multiple worker machines for parallel processing. Compared with traditional learning at a centralized data center, distributed learning offers several distinct advantages, such as preserving privacy, reducing network congestion, and leveraging distributed on-device computation. In a typical distributed learning framework such as federated learning [1]–[4], each edge device downloads a shared model from the edge server, computes an update to the current model by learning from its local dataset using, for instance, the SGD algorithm. Then these updates are sent to the edge server and averaged to improve the shared model. The process is repeated until a good solution is found.

The main bottleneck in distributed learning is the communication cost for model aggregation [5]–[9] since the model update from each edge device at each round of the training can consist of millions to billions of parameters. In general, each edge device needs to compress its local update, e.g. stochastic gradient, then transmit the compressed update to the edge server for aggregation over a shared channel with limited bandwidth. The convergence rate depends on the variance bound of the gradient received at the edge server [10]–[13], and the latter depends on the communication bandwidth. As such, a fundamental question arises: what is the tradeoff between the communication cost and the convergence rate in distributed learning. To this end, several model compression methods have been proposed to improve the communication efficiency such as quantization [6], [14]–[18], sparsification [19]–[22] and methods [14], [16] combining quantization and sparsification. Yet it remains unknown what is the theoretical achievability bound of the communication cost to achieve a target model accuracy. While this question may not be answered directly since the model accuracy not only depends on the number of iterations but also depends on the variance bound of the gradient received at the edge server, we resort to an alternative question: what is the minimum communication cost to achieve a target variance bound of the gradient estimation at each training round.

To characterize the minimum communication cost at a given variance bound falls into the classic rate distortion theory for multiterminal remote source coding problem, or the so-called CEO problem [23]. The primary goal of this paper is to have an insight of communication efficiency for distributed learning from the rate distortion perspective and to provide a achievability bound of communication cost at a particular variance upper bound. To the best of our knowledge, this is the first work aiming at analysing the theoretical limits on the communication efficiency for distributed learning from the rate distortion perspective.

In the CEO problem, the main objective of the CEO is a good estimate of a data sequence $\{X(t)\}_{t=1}^{\infty}$, which cannot be observed directly. The CEO employs a team of K agents who can observe the independently corrupted version of the data sequence $\{Y_k(t)\}_{t=1}^{\infty}, k = 1, 2, ..., K$. The observations $\{Y_k(t)\}_{t=1}^{\infty}, k = 1, 2, ..., K$ are separately encoded and forwarded to the CEO over rate-constrained channels. The model aggregation in distributed learning with SGD system is inherently a CEO problem. Specifically, the edge server works like the chief executive officer, and tries to reconstruct the target gradient sequence from the K local gradients while keeping the distortion lower than an acceptable level. The target gradient should be calculated by gradient descent (GD) on global dataset, while the local gradients are calculated SGD on local datesets, which are noisy versions of the target gradient.

The CEO problem was first studied by Berger, Zhang, and Viswanathan [23], which is considered to give a new direction in multi-terminal information theory. The quadratic Gaussian CEO problem, first studied by Viswanathan and Berger [24], has received particular attention. In [24], the decay rate of the minimum mean squared error (MMSE) with respect to the rate expenditure of the agents is considered and shown to be inversely proportional with the rate expenditure of the agents, when the number of agents goes to infinity. Major progress on the quadratic Gaussian CEO problem was made by Oohama in [25]. This problem is further studied in [26] and [27], where the entire rate-distortion region for the quadratic Gaussian problem is established. The achievability is shown by using the Berger-Tung inner bound [28]. The converse developed in [25] and later refined in [26], [27] is the application of Shannons entropy power inequality to relate various information quantities. Recently, the work [28] provides an alternative proof for the sum-rate of the quadratic Gaussian CEO problem without invoking the entropy-power inequality.

Note, however, that none of the existing rate region results for CEO problem can be directly applied to distributed learning due to the following reason. In the CEO problem, the only target is to minimize the distortion measured by MSE without any biased or unbiased estimation constraint. In distributed learning, however, it explicitly requires that the gradient estimation should be unbiased, i.e. $\mathbb{E}[\hat{X}(t)|X(t)] = X(t)$. Otherwise, the convergence cannot be guaranteed. Intuitively, the mean of the gradient estimator can be viewed as an ”effective information” for model convergence while the variance of the gradient estimator can be viewed as a ”noise” for model convergence. By adjusting the learning rate, the convergence rate only depends on the ratio between the ”effective information” and the ”noise”. This ratio of a biased gradient estimator with bounded mean squared error (MSE) may approach to zero while an unbiased gradient estimator with bounded MSE always has a bounded ratio.

Motivated by the above issue, in this paper, we first derive the rate region of the quadratic Gaussian CEO problem under unbiased estimation constraint. The rate region is derived from the standpoint of multiterminal rate-distortion theory. We formulate the multiterminal source coding problem in distributed learning as the quadratic Gaussian CEO problem based on the observation of the gradient distribution. Then we extend the rate region results of quadratic Gaussian CEO problem to the distributed learning. Our goal is to obtain the minimum achievable rate at a particular upper bound of gradient variance. Our results can guide the design of the practical model compression schemes. In other applications, such as federated learning (FL), our results can also be used as transmission models of the communication-efficient FL works such as aggregation frequency control works and user scheduling works. The main contributions of this work are outlined below:

- *Rate region results:* We derive the rate region results for the quadratic Gaussian CEO problem under unbiased estimation constraint. Our achievable region corresponds to a specialization of the general Berger-Tung achievable region [29] to the case of unbiased estimation constraint. We show that the Berger-Tung achievable region is also tight in this case. Our converse proof is inspired by the converse in work [25]. But unlike there, our problem is under unbiased estimation constraint. We reveal that the multiterminal source coding problem in distributed learning is the quadratic vector Gaussian CEO problem based on a thorough understanding of gradient distributions. We extend the rate region results of quadratic Gaussian CEO problem to the distributed learning.
- *Rate region boundary and sum-rate-distortion function:* To facilitate the analysis of the communication efficiency of distributed learning, we provide an explicit formula of the rate region boundary by solving an optimization problem. Specifically, the rate region are characterized by the tangent hyperplanes to it, since the rate region is convex. We also derive a closed-form sum-rate-distortion function in the special case where gradients are identically distributed over edge device and dimension. The derived function has the form of a sum of two nonnegative functions. One is a Shannon classical capacity of the additive white Gaussian noise channel 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 analyse communication efficiency of convex mini-batch SGD and non-convex mini-batch 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. In Section II, we give problem formulation and results of the scalar CEO problem. In Section III, we prove the direct and converse of the rate region results. In Section IV, we extend our rate region results to the distributed learning. Section V derives the explicit formula of rate region. Section VI provides the communication efficiency for distributed learning. Finally, we conclude the paper in Section VII.

II. Problem formulation and results

In this section, we first formulate a quadratic Gaussian CEO problem under unbiased estimation constraint, and then derive the rate region for this problem.

A. Problem Formulation

Let $\{X(t)\}_{t=1}^{\infty}$ be an independent and identical distributed Gaussian data sequence with mean 0 and variance σ_X^2. Each $X(t), t = 1, 2, ...$ takes value in real space $\mathbb{R}$. For $k = 1, ..., K$, let $\{Y_k(t)\}_{t=1}^{\infty}$ be noisy version of $\{X(t)\}_{t=1}^{\infty}$, each taking value in real space $\mathbb{R}$ and corrupted by an independent additive white Gaussian noise, i.e.,

$$Y_k(t) = X(t) + N_k(t), k = 1, 2, ..., K, \tag{1}$$

where $N_k(t)$ are Gaussian random variables independent over device k and iteration t. For $k = 1, 2, ..., K$ and $t = 1, 2, ...$, we assume that $N_k(t)$ is a Gaussian variable with mean 0 and variance $\sigma_{N_k}^2$. The edge server is interested in the sequence $\{X(t)\}_{t=1}^{\infty}$ that cannot be observed directly. The edge server employs a team of K edge devices who observe independently corrupted versions $\{Y_k(t)\}_{t=1}^{\infty}, k = 1, 2, ..., K$ of $\{X(t)\}_{t=1}^{\infty}$. We write n independent copies of $X(t)$ and $Y_k(t)$ as

$$X^n = (X(1), X(2), ..., X(n)), \tag{2}$$

and

$$Y_k^n = (Y_k(1), Y_k(2), ..., Y_k(n)), k = 1, 2, ..., K, \tag{3}$$

respectively. For $k = 1, 2, ..., K$, each data sequence Y_k^n observed by edge device k is separately encoded to $\phi_k(Y_k^n)$, and those are sent to the edge server, where the edge server observes $\phi_k(Y_k^n), k = 1, 2, ..., K$ and outputs the estimation $\hat{X}^n$ of X^n by using the decoder function ψ_K. The encoder function $\phi_k, k = 1, 2, ..., K$ are defined by

$$\phi_k : \mathbb{R}^n \to \mathcal{C}_k = \{1, 2, ..., |\mathcal{C}_k|\}, \tag{4}$$

and satisfy the total rate constraint

$$\frac{1}{n} \log |\mathcal{C}_k| \leq R_k, k = 1, 2, ..., K. \tag{5}$$

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

$$\phi^K = (\phi_1, \phi_2, ..., \phi_K). \tag{6}$$

Similarly, we write

$$\phi^K(Y^{nK}) = (\phi_1(Y_1^n), \phi_2(Y_2^n), ..., \phi_K(Y_K^n)). \tag{7}$$

The decoder function ψ_K is defined by

$$\psi_K : \mathcal{C}_1 \times \mathcal{C}_2 \times ... \times \mathcal{C}_K \to \mathbb{R}^n. \tag{8}$$

For $\hat{X}^n = \psi_K(\phi^K(Y^{nK}))$, define the average distortion by

$$D^n(X^n, \hat{X}^n) = \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}(X(t) - \hat{X}(t))^2. \tag{9}$$

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 (5) and decoder ψ_K such that $\hat{X}^n$ is unbiased estimator of X^n, i.e., $\mathbb{E}[\hat{X}^n | X^n = x^n] = x^n$ and $D^n(X^n, \hat{X}^n) \leq D$ 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.

B. Rate Region Results

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

$$\mathcal{R}_N(D) = \bigcup_{(r_1,r_2,...,r_K)\in\mathcal{F}(D)} \mathcal{R}(r_1,r_2,...,r_K), \tag{10}$$

where

$$\mathcal{R}(r_1,r_2,...,r_K) = \Bigg\{(R_1,R_2,...,R_K): \\ \sum_{k\in\mathcal{A}} R_k \geq \sum_{k\in\mathcal{A}} r_k + \frac{1}{2}\log\big(\frac{1}{\sigma_X^2}+\frac{1}{D}\big) - \frac{1}{2}\log\left(\frac{1}{\sigma_X^2}+\sum_{k\in\mathcal{A}^c}\frac{1-\exp(-2r_k)}{\sigma_{N_k}^2}\right), \forall\mathcal{A}\subseteq\mathcal{K}\Bigg\}, \tag{11}$$

and

$$\mathcal{F}(D) = \left\{(r_1,r_2,...,r_K)\in\mathbb{R}_+^K : \sum_{k=1}^{K}\frac{1-\exp(-2r_k)}{\sigma_{N_k}^2} = \frac{1}{D}\right\}. \tag{12}$$

Our main result is

Theorem 1:

$$\mathcal{R}_\star(D) = \mathcal{R}_N(D). \tag{13}$$

Remark 1: The rate region $\mathcal{R}_\star(D)$ under unbiased estimation constraint is a subset of the rate region results in work [27]. As will become clear in the next section, parameter r_k can be interpreted as the rate of the k-th edge device for quantizing its observation noise. It can be observed that our rate region with unbiased estimation constraint is a subset of the classic rate region without such constraint. In the special case when the distortion D is larger than the gradient variance σ_X^2, the classic rate region can achieve zero. This indicates that we can simply set $\hat{X} = 0$ at the receiver without any transmission. However, $\hat{X} = 0$ is not an option in our unbiased setting since the corresponding estimator is biased. Thus in this large distortion case, the rate region under unbiased estimator is still bounded away from zero. The extra rate introduced by our rate region results is necessary for model training. For a distortion D larger than the gradient variance σ_X^2, the model will never converge by applying the classic rate region while can converge by applying the rate region under unbiased estimator.

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

III. Proof

A. Achievability of Theorem 1

For the achievability proof, we adopt the classic Berger-Tung scheme [29] but design an unbiased estimator at receiver.

Lemma 1 (Berger-Tung inner bound): If we can find auxiliary random variables $U_1, U_2, ..., U_K$ such that

$$U_1, U_2, ..., U_K | Y_1, Y_2, ..., Y_K \sim p(u_1|y_1)p(u_2|y_2)...p(u_K|y_K) \tag{14}$$

and decoding function $\hat{X}(U_1, U_2, ..., U_K)$ such that

$$\mathbb{E}(X - \hat{X}(U^K))^2 \leq D, \tag{15}$$

then the following rate region is achievable

$$\sum_{k \in \mathcal{A}} R_k \geq I(\boldsymbol{Y}_{\mathcal{A}}; \boldsymbol{U}_{\mathcal{A}} | \boldsymbol{U}_{\mathcal{A}^c}), \forall \mathcal{A} \subseteq \{1, 2, ..., K\}. \tag{16}$$

Now, let us consider the Berger-Tung scheme in our problem. For each edge device, we define the auxiliary random variable $U_k = Y_k + V_k$, where $V_k \sim \mathcal{N}(0, \sigma^2_{V_k})$ are independently distributed and independent of Y_k and X. Parameters $\{\sigma^2_{V_k}\}$ are determined in terms of the target distortion D. After recovering U^{nK}, the decoder reconstructs $\hat{X}^n$ by applying the following weighted averaging function component-wise

$$\hat{x} = \psi_K(U^K) \triangleq \sum_{k=1}^{K} \alpha_k u_k, \tag{17}$$

where $\alpha_k = \frac{(\sigma^2_{N_k} + \sigma^2_{V_k})^{-1}}{\sum_{k=1}^{K} (\sigma^2_{N_k} + \sigma^2_{V_k})^{-1}}$. Note that $\hat{X}$ is unbiased estimator of X and we set the variance of $\hat{X}$ equal to the target distortion D, which is given by

$$D = \left(\sum_{k=1}^{K} \frac{1}{\sigma^2_{N_k} + \sigma^2_{V_k}} \right)^{-1}. \tag{18}$$

Let us define

$$r_k \triangleq I(Y_k; U_k | X) = \frac{1}{2} \log (1 + \frac{\sigma^2_{N_k}}{\sigma^2_{V_k}}), k = 1, 2, ..., K. \tag{19}$$

We can interpret parameter r_k as the rate of the k-th device for quantizing its observation noise. We will use r_k's as the parameters instead of $\sigma^2_{V_k}$. Note that for any choice of $(r_k \geq 0, k =$

$1,2,...,K)$, we can find a corresponding $(\sigma_{V_k}^2 \geq 0, k = 1,2,...,K)$ and therefore, a set of auxiliary random variables. Then we can rewrite (18) in terms of r_k's as

$$\frac{1}{D} = \sum_{k=1}^{K} \frac{1-\exp(-2r_k)}{\sigma_{N_k}^2}, \tag{20}$$

as desired.

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

$$\begin{aligned}
\sum_{k\in\mathcal{A}} R_k &\geq I(\boldsymbol{Y}_{\mathcal{A}}; \boldsymbol{U}_{\mathcal{A}}|\boldsymbol{U}_{\mathcal{A}^c}) && (21)\\
&= I(\boldsymbol{Y}_{\mathcal{A}}, X; \boldsymbol{U}_{\mathcal{A}}|\boldsymbol{U}_{\mathcal{A}^c}) && (22)\\
&= I(X; \boldsymbol{U}_{\mathcal{A}}|\boldsymbol{U}_{\mathcal{A}^c}) + I(\boldsymbol{Y}_{\mathcal{A}}; \boldsymbol{U}_{\mathcal{A}}|X) && (23)\\
&= I(X; \boldsymbol{U}_{\mathcal{A}}|\boldsymbol{U}_{\mathcal{A}^c}) + \sum_{k\in\mathcal{A}} r_k && (24)\\
&= h(X|\boldsymbol{U}_{\mathcal{A}^c}) - h(X|\boldsymbol{U}) + \sum_{k\in\mathcal{A}} r_k && (25)\\
&= -\frac{1}{2}\log\left(\frac{1}{\sigma_X^2} + \sum_{k\in\mathcal{A}^c}(\frac{1}{\sigma_{N_k}^2+\sigma_{V_k}^2})^{-1}\right) + \frac{1}{2}\log(\frac{1}{\sigma_X^2} + \sum_{k\in\mathcal{K}}(\frac{1}{\sigma_{N_k}^2+\sigma_{V_k}^2})^{-1}) + \sum_{k\in\mathcal{A}} r_k && (26)\\
&= -\frac{1}{2}\log\left(\frac{1}{\sigma_X^2} + \sum_{k\in\mathcal{A}^c}\frac{1-\exp(-2r_k)}{\sigma_{N_k}^2}\right) + \frac{1}{2}\log(\frac{1}{\sigma_X^2} + \frac{1}{D}) + \sum_{k\in\mathcal{A}} r_k. && (27)
\end{aligned}$$

Equations (20) and (27) together concludes the achievable proof.

B. Converse of Theorem 1

Our converse proof is inspired by Oohama's converse in [25] but is not straightforward and cannot be derived directly from traditional (biased) CEO. Suppose we achieve an unbiased estimator $\hat{X}^n$ with distortion $D = D^n(X^n, \hat{X}^n)$. 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

$$r_k \triangleq \frac{1}{n} I(Y_k^n; C_k|X^n). \tag{28}$$

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

$$\begin{aligned}
\sum_{k\in\mathcal{A}} R_k &\geq \sum_{k\in\mathcal{A}} \frac{1}{n} H(C_k) && (29)\\
&\geq \frac{1}{n} H(\boldsymbol{C}_{\mathcal{A}}) && (30)\\
&\geq \frac{1}{n} H(\boldsymbol{C}_{\mathcal{A}}|\boldsymbol{C}_{\mathcal{A}^c}) && (31)\\
&\geq \frac{1}{n} I(\boldsymbol{Y}^n_{\mathcal{A}}; \boldsymbol{C}_{\mathcal{A}}|\boldsymbol{C}_{\mathcal{A}^c}) && (32)\\
&= \frac{1}{n} I(X^n, \boldsymbol{Y}^n_{\mathcal{A}}; \boldsymbol{C}_{\mathcal{A}}|\boldsymbol{C}_{\mathcal{A}^c}) && (33)\\
&= \frac{1}{n} I(X^n; \boldsymbol{C}_{\mathcal{A}}|\boldsymbol{C}_{\mathcal{A}^c}) + \frac{1}{n} I(\boldsymbol{Y}^n_{\mathcal{A}}; \boldsymbol{C}_{\mathcal{A}}|\boldsymbol{C}_{\mathcal{A}^c}, X^n) && (34)\\
&= \frac{1}{n} I(X^n; \boldsymbol{C}_{\mathcal{A}}|\boldsymbol{C}_{\mathcal{A}^c}) + \frac{1}{n} I(X^n; \boldsymbol{C}_{\mathcal{A}^c}) - \frac{1}{n} I(X^n; \boldsymbol{C}_{\mathcal{A}^c}) + \frac{1}{n} I(\boldsymbol{Y}^n_{\mathcal{A}}; \boldsymbol{C}_{\mathcal{A}}|X^n) && (35)\\
&= \frac{1}{n} I(X^n; \boldsymbol{C}_{\mathcal{K}}) - \frac{1}{n} I(X^n; \boldsymbol{C}_{\mathcal{A}^c}) + \sum_{k\in\mathcal{A}} \frac{1}{n} I(Y^n_k; C_k|X^n) && (36)\\
&= \frac{1}{n} I(X^n; \boldsymbol{C}_{\mathcal{K}}) - \frac{1}{n} I(X^n; \boldsymbol{C}_{\mathcal{A}^c}) + \sum_{k\in\mathcal{A}} r_k, && (37)
\end{aligned}$$

where (31) follows from the fact that conditioning reduces entropy, (33) follows from the fact that $X^n - \boldsymbol{Y}^n_{\mathcal{A}} - \boldsymbol{C}_{\mathcal{A}}$ are Markov chain, (34) follows from the chain rule for mutual information, (35) follows from the fact that $(\boldsymbol{Y}^n_{\mathcal{A}}, \boldsymbol{C}_{\mathcal{A}}) - X^n - \boldsymbol{C}_{\mathcal{A}^c}$ are Markov chain, (36) follows from the chain rule for mutual information and the fact that $Y^n_k - X^n - \boldsymbol{C}_{\mathcal{K}\setminus\{k\}}$ are Markov chain.

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

$$\frac{1}{n} I(X^n; \hat{X}^n) \geq \frac{1}{2}\log(1 + \frac{\sigma^2_X}{D^n(X^n, \hat{X}^n)}). \tag{38}$$

Proof:

$$\begin{aligned}
\frac{1}{n} I(X^n; \hat{X}^n) &= \frac{1}{n} h(X^n) - \frac{1}{n} h(X^n|\hat{X}^n) && (39)\\
&= \frac{1}{n} h(X^n) - \frac{1}{n}\sum_{t=1}^n h(X(t)|X^{t-1}\hat{X}^n) && (40)\\
&\geq \frac{1}{n} h(X^n) - \frac{1}{n}\sum_{t=1}^n h(X(t)|\hat{X}(t)) && (41)\\
&\geq \frac{1}{2}\log 2\pi e\sigma^2_X - \frac{1}{n}\sum_{t=1}^n \frac{1}{2}\log 2\pi e\sigma^2_{X(t)|\hat{X}(t)}. && (42)
\end{aligned}$$

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

$$\begin{aligned}
\sigma^2_{X|\hat{X}} &= \mathbb{E}\left[X - \mathbb{E}(X|\hat{X})\right]^2 \tag{43}\\
&\leq \min_a \mathbb{E}(X - a\hat{X})^2 \tag{44}\\
&= \min_a \mathbb{E}\left[\mathbb{E}\left((X - a\hat{X})^2|X\right)\right] \tag{45}\\
&= \min_a \mathbb{E}\left[\mathbb{E}^2(X - a\hat{X}|X) + \text{Var}(X - a\hat{X}|X)\right] \tag{46}\\
&= \min_a \mathbb{E}\left\{(X - aX)^2 + \mathbb{E}\left[(X - a\hat{X}) - \mathbb{E}(X - a\hat{X}|X)\right]\right\} \tag{47}\\
&= \min_a \left[(1-a)^2\sigma_X^2 + a^2\mathbb{E}(X - \hat{X})^2\right] \tag{48}\\
&= (\frac{1}{\sigma_X^2} + \frac{1}{\mathbb{E}(X - \hat{X})^2})^{-1}. \tag{49}
\end{aligned}$$

where (47) and (48) follows from the fact that $\mathbb{E}(\hat{X}|X) = X$. Substituting (49) to (42), we have

$$\begin{aligned}
\frac{1}{n}I(X^n;\hat{X}^n) &\geq \frac{1}{2}\log 2\pi e\sigma_X^2 + \frac{1}{n}\sum_{t=1}^{n}\frac{1}{2}\log\frac{1}{2\pi e}(\frac{1}{\sigma_X^2} + \frac{1}{\mathbb{E}(X(t) - \hat{X}(t))^2}) \tag{50}\\
&\geq \frac{1}{2}\log 2\pi e\sigma_X^2 + \frac{1}{2}\log\frac{1}{2\pi e}(\frac{1}{\sigma_X^2} + \frac{1}{D^n(X^n,\hat{X}^n)}) \tag{51}\\
&= \frac{1}{2}\log(1 + \frac{\sigma_X^2}{D^n(X^n,\hat{X}^n)}), \tag{52}
\end{aligned}$$

where (51) follows from Jensens inequality. The proof of Lemma 2 is completed. ■

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

$$\begin{aligned}
\frac{1}{n}I(X^n;\boldsymbol{C}_\mathcal{K}) &\geq \frac{1}{n}I(X^n;\hat{X}^n) \tag{53}\\
&\geq \frac{1}{2}\log(1 + \frac{\sigma_X^2}{D^n(X^n,\hat{X}^n)}) \tag{54}\\
&= \frac{1}{2}\log(1 + \frac{\sigma_X^2}{D}), \tag{55}
\end{aligned}$$

where (53) follows from the data-processing inequality and (54) 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

$$\frac{1}{\sigma_X^2}\exp\left(\frac{2}{n}I(X^n;\boldsymbol{C}_\mathcal{A})\right) \leq \frac{1}{\sigma_X^2} + \sum_{k\in\mathcal{A}}\frac{1 - \exp(-2r_k)}{\sigma_{N_k}^2}. \tag{56}$$

Proof: see Lemma 3.1 in [27]. ■

Using the bounds from (55) and (56) in (37), for all $\mathcal{A} \subseteq \{1, 2, ..., K\}$

$$\sum_{k\in\mathcal{A}} R_k \geq \frac{1}{2}\log(1+\frac{\sigma_X^2}{D}) - \frac{1}{2}\log\left(1+\sigma_X^2 \sum_{k\in\mathcal{A}^c} \frac{1-\exp(-2r_k)}{\sigma_{N_k}^2}\right) + \sum_{k\in\mathcal{A}} r_k \tag{57}$$

$$= -\frac{1}{2}\log\left(\frac{1}{\sigma_X^2} + \sum_{k\in\mathcal{A}^c} \frac{1-\exp(-2r_k)}{\sigma_{N_k}^2}\right) + \frac{1}{2}\log(\frac{1}{\sigma_X^2}+\frac{1}{D}) + \sum_{k\in\mathcal{A}} r_k. \tag{58}$$

Substituting (55) to (56) with $\mathcal{A} = \mathcal{K}$, we have the condition

$$\sum_{k=1}^{K} \frac{1-\exp(-2r_k)}{\sigma_{N_k}^2} \geq \frac{1}{D}. \tag{59}$$

Along with the non-negativity constraints on r_k's, (58) and (59) define an outer bound for $\mathcal{R}_\star(D)$. It is easy to show that replacing the inequality in (59) with an equality will not change this outer bound. Thus we have $\mathcal{R}_\star(D) \subseteq \mathcal{R}_N(D)$.

IV. Application in Distributed Learning

In this section we extend our rate region results to the distributed learning. There are many variants in distributed learning, which have different preconditions and convergence guarantees. Our proposed rate region results are very adaptable. To facilitate the presentation, we only focus on a basic distributed learning setup. However, the results can be extended to other distributed learning cases such as federated learning and distributed reinforcement learning.

A. System Model

We consider a distributed learning framework as illustrated in Fig. 1, 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 the model error is defined as

$$F(\boldsymbol{w}) = \sum_{k\in\mathcal{K}} \frac{|\mathcal{S}_k|}{|\mathcal{S}|} F_k(\boldsymbol{w}), \tag{60}$$

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 $\boldsymbol{G}(t) \triangleq \nabla F(\boldsymbol{w}(t)) \in \mathbb{R}^P$

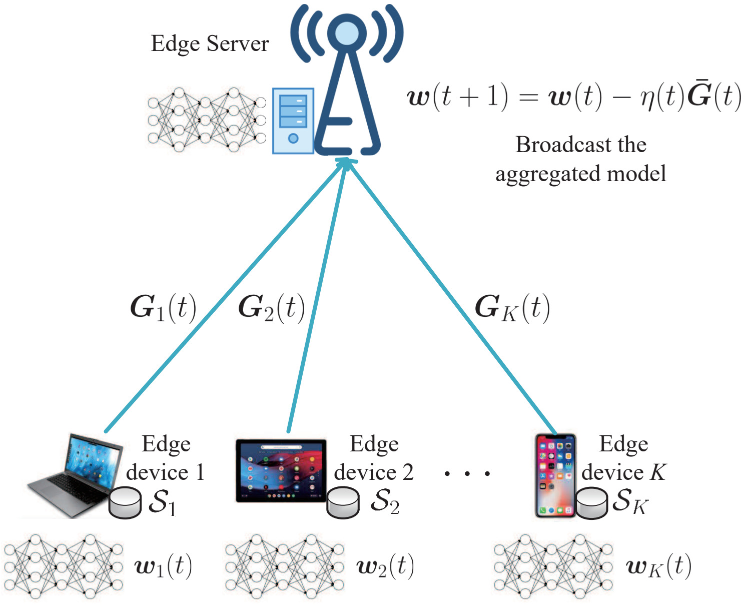


Fig. 1. Illustration of distributed learning system in wireless networks.

denote the gradient vector calculated through the GD algorithm at iteration t. The minimization of $F(\boldsymbol{w})$ is typically carried out through the mini-batch stochastic gradient descent (mini-batch SGD) algorithm, where device k's local dataset $\mathcal{S}_k$ is split into mini-batches of size B_k and at each iteration $t = 1, 2, ..., T$, we draw one mini-batch $\mathcal{B}_k(t)$ randomly and calculate the local gradient vector as

$$\boldsymbol{G}_k(t) = \nabla \frac{1}{B_k} \sum_{i \in \mathcal{B}_k(t)} f_i(\boldsymbol{w}). \tag{61}$$

When the mini-batch size $B_k = 1$, the mini-batch 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 updates the model parameter as

$$\boldsymbol{w}(t+1) = \boldsymbol{w}(t) - \eta(t)\bar{\boldsymbol{G}}(t), \tag{62}$$

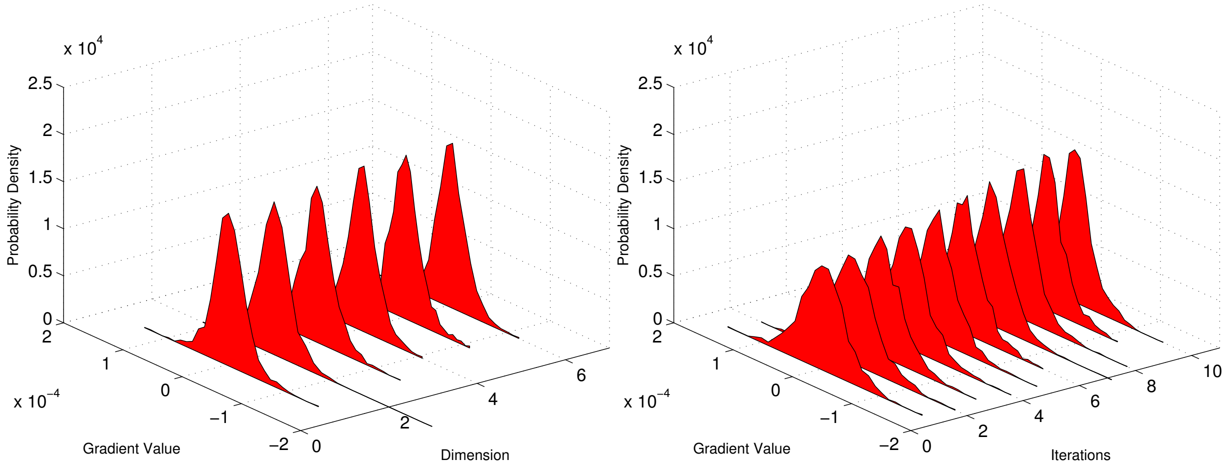
with $\eta(t)$ being the learning rate at iteration t.

B. 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 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 an 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 [30]–[33]. The assumption of independence is valid when the learning rate is large enough.
- The gradient elements $\{G_p(t)\}, \forall p \in \{1, 2, ..., P\}$ 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 [34], [35] and parameter sharing/tying [36], [37].

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. 2 illustrates the experimental results of the distribution of global gradient of dataset MNIST. Fig. 2(a) shows the distribution of the global gradient $\boldsymbol{G}(t)$ over 6 dimensions at iteration 10. Fig. 2(b) shows the distribution of the

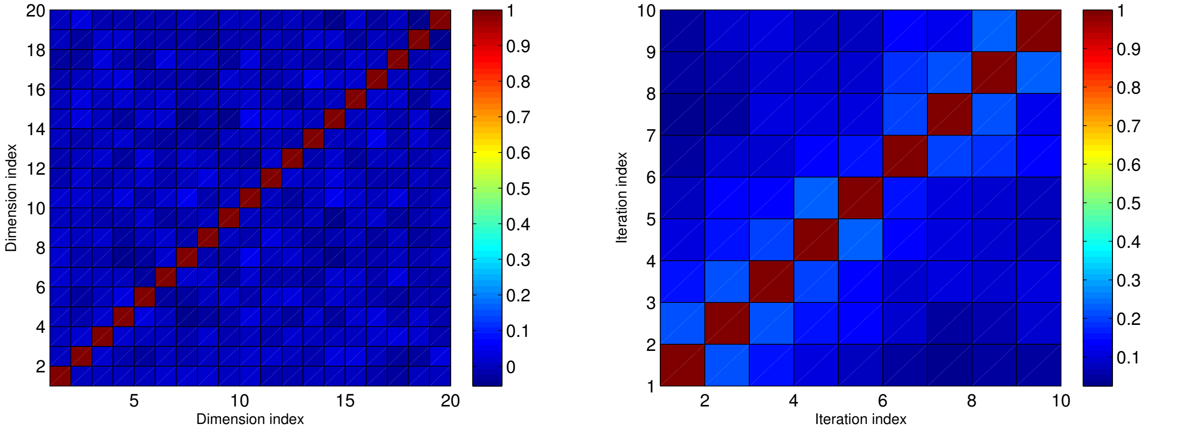


(a) The distribution of global gradient at different dimension.

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

Fig. 2. The distribution of global gradient.

global gradient $\boldsymbol{G}(t)$ dimension of global gradient $\boldsymbol{G}(t)$ over iteration [1, 10] at one dimension. It is observed that the global gradient $\boldsymbol{G}(t)$ follows a Gaussian distribution with mean zero. It is also observed that the variance of the global gradient gradually decreases with the model converges.



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

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

Fig. 3. Correlation coefficient of global gradient.

The correlation coefficient of Gaussian variables can indicate the independence of them. Fig. 3 shows the correlation coefficient of global gradient over iteration t's and dimension p's. Fig. 3(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, which justifies that the global gradient elements $\{G_p(t)\}, \forall p \in \{1, 2, ..., P\}$ are independent over dimension p's. Fig. 3(b) shows the correlation coefficient of global gradient over iteartion t's. It is observed that the correlation coefficient of gradients on different iterations are almost zero, which justifies that the global gradient $\{\boldsymbol{G}(t)\}_{t=1}^{\infty}$ is an independent distributed sequence.

2) Local Gradients: For $k = 1, 2, ..., K$, edge device k carries out mini-batch SGD in distributed learning. Recall that $\{\boldsymbol{G}_k(t)\}_{t=1}^{\infty}$ is the local gradient vector sequence calculated through the mini-batch 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 $\{\boldsymbol{N}_k(t)\}_{t=1}^{\infty}$ is an independent distributed vector Gaussian sequence with zero mean, which is independent of the $\boldsymbol{G}(t)$ process. Note that the gradient noise $\boldsymbol{N}_k(t)$ depends on the selection of local batch at edge device k. The assumption of independence is valid since the selection of local batch is independent of the global gradient $\boldsymbol{G}(t)$.
- The gradient noises $\{\boldsymbol{N}_k(t)\}, \forall k \in \mathcal{K}$ 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 elements $\{N_{k,p}(t)\}, \forall p \in \{1, 2, ..., P\}$ 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.

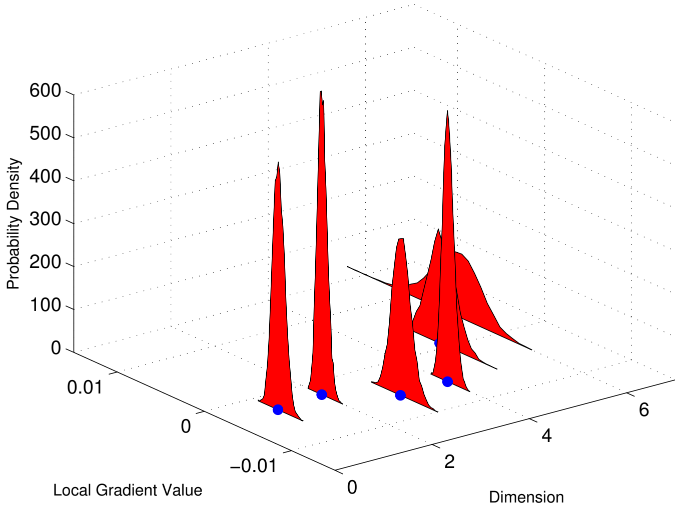
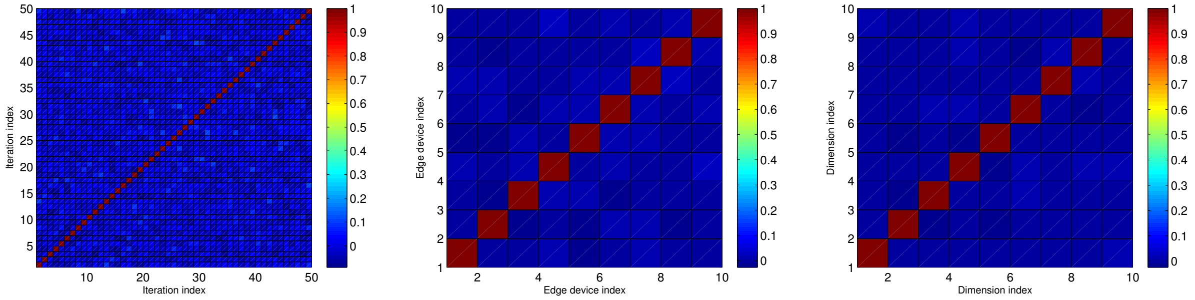


Fig. 4. The conditional distribution of local gradient $\boldsymbol{G}_k(t)$ given global gradient $\boldsymbol{G}(t)$, where the blue dot represents the given global gradient $\boldsymbol{G}(t)$.

Fig. 4 illustrates the experimental results of the conditional distribution of local gradient $\boldsymbol{G}_k(t)$ given global gradient $\boldsymbol{G}(t)$. We sampled 10000 local gradients within one iteration. From Fig. 4, it is observed that the conditional local gradient $\boldsymbol{G}_k(t)$ given $\boldsymbol{G}(t)$ follows a Gaussian distribution, which justifies that the gradient noise $\boldsymbol{N}_k(t)$ follows a Gaussian distribution. It is also observed that the conditional expectation of local gradient is equal to global gradient, i.e.,

$$\mathbb{E}[\boldsymbol{G}_k(t)|\boldsymbol{G}(t) = \boldsymbol{g}(t)] = \boldsymbol{g}(t). \tag{63}$$

Hence, we have justified that the mean of noise $\boldsymbol{N}_k(t)$ is zero.



(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. 5. Correlation coefficient of gradient noise over iteration, device and dimension.

Fig. 5 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 $t = 1, ..., 50$ and $k = 1, ..., 10$. We randomly select 10 dimensions in the gradient noise $\boldsymbol{N}_k(t)$ to show the correlation coefficient of gradient noise over dimension p. It is observed that the correlation coefficient of gradient noise are nearly zero over dimension p's, device k's and iteration t's, which justifies that the noise sequence $\{\boldsymbol{N}_k(t)\}_{t=1}^{\infty}$ is independent over dimension p's, device k's and iteration t's.

C. Rate Region in Distributed Learning

Fig. 6 shows the CEO Problem in distributed learning. The edge server is interested in the sequence $\{\boldsymbol{G}(t)\}_{t=1}^{\infty}$ that cannot be observed directly. The edge server employs K edge devices who observe independently corrupted versions $\{\boldsymbol{G}_k(t)\}_{t=1}^{\infty}, k = 1, ..., K$ of $\{\boldsymbol{G}(t)\}_{t=1}^{\infty}$. We assume that the global gradient $\{\boldsymbol{G}(t)\}_{t=1}^{\infty}$ is an independent Gaussian vector sequence with mean 0 and variance diag$(\sigma_{X_1}^2, \sigma_{X_2}^2, ..., \sigma_{X_P}^2)$. Each $\boldsymbol{G}(t), t = 1, ..., T$ takes value in real space $\mathbb{R}^P$. We assume that the local gradient $\boldsymbol{G}_k(t)$ is a 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.,

$$\boldsymbol{G}_k(t) = \boldsymbol{G}(t) + \boldsymbol{N}_k(t), \tag{64}$$

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 Gaussian

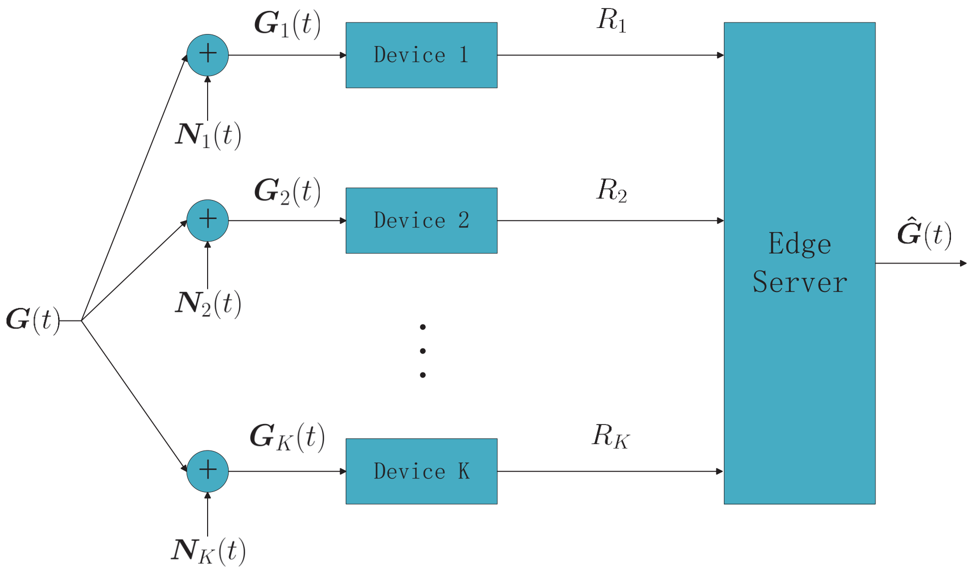


Fig. 6. The Gaussian vector CEO problem in distributed learning.

variable with mean 0 and variance $\text{diag}(\sigma^2_{N_{k,1}}(t), \sigma^2_{N_{k,2}}(t), ..., \sigma^2_{N_{k,P}}(t))$. We write n independent copies of $\boldsymbol{G}(t)$ and $\boldsymbol{G}_k(t)$ as $\boldsymbol{G}^n$ and $\boldsymbol{G}_k^n$, respectively.

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

$$D^n(\boldsymbol{G}^n, \hat{\boldsymbol{G}}^n) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\|\boldsymbol{G}[i] - \hat{\boldsymbol{G}}[i]\|^2. \tag{65}$$

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 (5) and decoder ψ_K such that $\hat{\boldsymbol{G}}^n$ is unbiased estimator of $\boldsymbol{G}^n$, i.e., $\mathbb{E}[\hat{\boldsymbol{G}}^n|\boldsymbol{G}^n] = \boldsymbol{G}^n$ 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.

The rate region results can be easily extended from Theorem 1:

Corollary 1:

$$\begin{aligned}\mathcal{R}_\star(D) = \Bigg\{(R_1, ..., R_K) \in \mathbb{R}_+^K :& R_k \geq \sum_{p=1}^{P} R_{k,p}, k = 1, 2, ..., K \\ & \forall (R_{1,p}, R_{2,p}, ..., R_{K,p}) \in \mathcal{R}_p(r_{1,p}, r_{2,p}, ..., r_{K,p}), p = 1, 2, ..., P \\ & \forall (r_{1,p}, r_{2,p}, ..., r_{K,p}) \in \mathcal{F}_p(D_p), p = 1, 2, ..., P \\ & \forall (D_1, D_2, ..., D_P) \in \mathbb{R}_+^P, \sum_{p=1}^{P} D_p = D\Bigg\}\end{aligned} \tag{66}$$

where

$$\mathcal{R}_p(r_{1,p}, r_{2,p}, ..., r_{K,p}) = \Bigg\{(R_{1,p}, R_{2,p}, ..., R_{K,p}) :$$
$$\sum_{k\in\mathcal{A}} R_{k,p} \geq \sum_{k\in\mathcal{A}} r_{k,p} + \frac{1}{2}\log\big(\frac{1}{\sigma^2_{X_p}} + \frac{1}{D_p}\big) - \frac{1}{2}\log\left(\frac{1}{\sigma^2_{X_p}} + \sum_{k\in\mathcal{A}^c} \frac{1-\exp(-2r_{k,p})}{\sigma^2_{N_{k,p}}}\right), \forall\mathcal{A}\subseteq\mathcal{K}\Bigg\}, \tag{67}$$

and

$$\mathcal{F}_p(D_p) = \left\{(r_{1,p}, r_{2,p}, ..., r_{K,p}) \in \mathbb{R}^K_+ : \sum_{k=1}^{K} \frac{1-\exp(-2r_{k,p})}{\sigma^2_{N_{k,p}}} = \frac{1}{D_p}\right\}. \tag{68}$$

Proof: First, we prove the direct parts of this corollary. Since the local gradients $\boldsymbol{G}_k$ and global gradient $\boldsymbol{G}$ are independent over dimension p's, we can prove the achievability of corollary by encoding each dimension of the local gradient $\boldsymbol{G}_{k,p}(t)$ separately. For all $(D_1, D_2, ..., D_P) \in \mathbb{R}^P_+, \sum_{p=1}^P D_p = D$, from the Theorem 1, $\forall(r_{1,p}, r_{2,p}, ..., r_{K,p}) \in \mathcal{F}_p(D_p), p = 1, 2, ..., P$, $\forall(R_{1,p}, ..., R_{K,p}) \in \mathcal{R}_p(r_{1,p}, r_{2,p}, ..., r_{K,p}), p = 1, 2, ..., P$ are achievable such that $D^n(G^n_p, \hat{G}^n_p) \leq D_p, p = 1, 2, ..., P$. Then $R_k \geq \sum_{p=1}^P R_{k,p}, k = 1, 2, ..., K$ are achievable such that

$$D^n(\boldsymbol{G}^n, \hat{\boldsymbol{G}}^n) = \sum_{p=1}^{P} D^n(G^n_p, \hat{G}^n_p) \leq \sum_{p=1}^{P} D_p = D, \tag{69}$$

We have completed the achievable proof of this corollary.

Second, we prove the converse parts of this corollary. Suppose we achieve an unbiased gradient estimator $\hat{\boldsymbol{G}}^n$ with distortion $D = \sum_{p=1}^P D_p$, where $D_p = D^n(G^n_p, \hat{G}^n_p)$. 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

$$r_{k,p} \triangleq \frac{1}{n} I(G^n_{k,p}; C_k | G^n_p). \tag{70}$$

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

$$\sum_{k\in\mathcal{A}} R_k \geq \frac{1}{n} I(\boldsymbol{G}^n; \boldsymbol{C}_\mathcal{K}) - \frac{1}{n} I(\boldsymbol{G}^n; \boldsymbol{C}_{\mathcal{A}^c}) + \sum_{k\in\mathcal{A}} \frac{1}{n} I(\boldsymbol{G}^n_k; C_k | \boldsymbol{G}^n) \tag{71}$$
$$= \sum_{p=1}^{P}\left(\frac{1}{n} I(G^n_p; \boldsymbol{C}_\mathcal{K}) - \frac{1}{n} I(G^n_p; \boldsymbol{C}_{\mathcal{A}^c}) + \sum_{k\in\mathcal{A}} r_{k,p}\right) \tag{72}$$
$$\geq \sum_{p=1}^{P}\left(-\frac{1}{2}\log\left(\frac{1}{\sigma^2_{X_p}} + \sum_{k\in\mathcal{A}^c}\frac{1-\exp(-2r_{k,p})}{\sigma^2_{N_{k,p}}}\right) + \frac{1}{2}\log\big(\frac{1}{\sigma^2_{X_p}} + \frac{1}{D_p}\big) + \sum_{k\in\mathcal{A}} r_{k,p}\right), \tag{73}$$

where (71) is similar to the derivation of the inequality (36), and (73) follows from Lemma 2 and Lemma 3. For each $p = 1, 2, ..., P$, similar to the derivation of (59), we have the condition

$$\sum_{k=1}^{K} \frac{1 - \exp(-2r_{k,p})}{\sigma^2_{N_{k,p}}} \geq \frac{1}{D_p}, p = 1, 2, ..., P. \tag{74}$$

It is easy to show that replacing the inequality in (74) with an equality will not change this outer bound. Thus we have completed the proof of the converse part. ■

Remark 2: Parameter D_p can be interpreted as the distortion of gradient estimator on the p-th dimension, $r_{k,p}$ can be interpreted as the rate of the k-th edge device for quantizing its observation noise on the p-th dimension, and $R_{k,p}$ can be interpreted as the rate contributed by edge device k on the p-th dimension.

V. Explicit Formula of Rate Region

The expression of rate region in section IV is implicit. To facilitate the analysis of the communication efficiency of distributed learning, we will derive the rate region boundary and a closed-form sum-rate-distortion function in this section.

A. Rate region boundary

We can explicitly compute the the boundary of the rate region $\mathcal{R}_\star(D)$ by solving the following optimization problem,

$$\min_{(R_1, R_2, ..., R_K) \in \mathcal{R}_\star(D)} \sum_{k \in \mathcal{K}} \alpha_k R_k = \min_{D_p} \min_{r_{k,p}} \min_{R_{k,p}} \sum_{p=1}^{P} \sum_{k \in \mathcal{K}} \alpha_k R_{k,p}, \tag{75}$$

for all choices of $(\alpha_1, \alpha_2, ..., \alpha_K) \in \mathbb{R}^K_+$. Without loss of generality, we 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

$$R_{k,p} = r_{k,p} + \frac{1}{2} \log \left(\frac{1}{\sigma^2_{X_p}} + \sum_{i=k}^{K} \frac{1 - \exp(-2r_{i,p})}{\sigma^2_{N_{i,p}}} \right) - \frac{1}{2} \log \left(\frac{1}{\sigma^2_{X_p}} + \sum_{i=k+1}^{K} \frac{1 - \exp(-2r_{i,p})}{\sigma^2_{N_{i,p}}} \right). \tag{76}$$

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 (76).

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

$$\sum_{k\in\mathcal{A}} R_{k,p} = \sum_{k\in\mathcal{A}} I(G_{k,p}; U_{k,p}|U_{k+1,p}, ..., U_{K,p}) \tag{77}$$

$$\geq \sum_{k\in\mathcal{A}} I(G_{k,p}; U_{k,p}|U_{\mathcal{A}^c\cup\{k+1,...,K\},p}) \tag{78}$$

$$= I(G_{\mathcal{A},p}; U_{\mathcal{A},p}|U_{\mathcal{A}^c,p}). \tag{79}$$

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

$$\sum_{i=1}^{k} R^{'}_{i,p} \geq I(G_{1,p}, ..., G_{k,p}; U_{1,p}, ..., U_{k,p}|U_{k+1,p}, ..., U_{K,p}) = \sum_{i=1}^{k} R_{i,p}, k = 1, 2, ..., K. \tag{80}$$

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

$$\sum_{k\in\mathcal{K}} \alpha_k R^{'}_{k,p} - \sum_{k\in\mathcal{K}} \alpha_k R_{k,p} \tag{81}$$

$$= \sum_{k\in\mathcal{K}} \alpha_k (R^{'}_{k,p} - R_{k,p}) \tag{82}$$

$$= \sum_{k\in\mathcal{K}} \left[(\alpha_k - \alpha_{k+1}) + (\alpha_{k+1} - \alpha_{k+2}) + ... + (\alpha_{K-1} - \alpha_K) + (\alpha_K)\right] (R^{'}_{k,p} - R_{k,p}) \tag{83}$$

$$= \sum_{k=1}^{K-1} (\alpha_k - \alpha_{k+1})(\sum_{i=1}^{k} R_{i,p} - \sum_{i=1}^{k} R^{'}_{i,p}) + \alpha_K(\sum_{i\in\mathcal{K}} R_{i,p} - \sum_{i\in\mathcal{K}} R^{'}_{i,p}) \tag{84}$$

$$\geq 0. \tag{85}$$

The inequality in (85) 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

$$R_{k,p} = I(G_{k,p}; U_{k,p}|U_{k+1,p}, ..., U_{K,p}) \tag{86}$$

$$= I(G_{1,p}...G_{k,p}; U_{1,p}...U_{k,p}|U_{k+1,p}, ..., U_{K,p}) - I(G_{1,p}...G_{k-1,p}; U_{1,p}...U_{k-1,p}|U_{k,p}, ..., U_{K,p}) \tag{87}$$

$$= \sum_{i\in\{1,...,k\}} r_{i,p} + \frac{1}{2}\log\left(\frac{1}{\sigma^2_{X_p}} + \frac{1}{D_p}\right) - \frac{1}{2}\log\left(\frac{1}{\sigma^2_{X_p}} + \sum_{i\in\{1,...,k\}^c} \frac{1-\exp(-2r_{i,p})}{\sigma^2_{N_{i,p}}}\right) \tag{88}$$

$$- \left[\sum_{i\in\{1,...,k-1\}} r_{i,p} + \frac{1}{2}\log\left(\frac{1}{\sigma^2_{X_p}} + \frac{1}{D_p}\right) - \frac{1}{2}\log\left(\frac{1}{\sigma^2_{X_p}} + \sum_{i\in\{1,...,k-1\}^c} \frac{1-\exp(-2r_{i,p})}{\sigma^2_{N_{i,p}}}\right)\right] \tag{89}$$

$$= r_{k,p} + \frac{1}{2}\log\left(\frac{1}{\sigma^2_{X_p}} + \sum_{i=k}^{K} \frac{1-\exp(-2r_{i,p})}{\sigma^2_{N_{i,p}}}\right) - \frac{1}{2}\log\left(\frac{1}{\sigma^2_{X_p}} + \sum_{i=k+1}^{K} \frac{1-\exp(-2r_{i,p})}{\sigma^2_{N_{i,p}}}\right). \tag{90}$$

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

$$\mathcal{P}_1: \quad \min \quad \sum_{p=1}^{P}\sum_{k=1}^{K}\alpha_k\left[-\frac{1}{2}\log\left(1-\sigma^2_{N_{k,p}}z_{k,p}\right) + \frac{1}{2}\log\left(\frac{1}{\sigma^2_{X_p}} + \sum_{i=k}^{K} z_{i,p}\right) - \frac{1}{2}\log\left(\frac{1}{\sigma^2_{X_p}} + \sum_{i=k+1}^{K} z_{i,p}\right)\right] \tag{91a}$$

$$s.t. \qquad \sum_{k\in\mathcal{K}} z_{k,p} \geq \frac{1}{D_p}, \quad p = 1, 2, ..., P \tag{91b}$$

$$\sum_{p=1}^{P} D_p \leq D. \tag{91c}$$

It is obvious that it is a convex optimization problem and the inequality constraints (91b) and (91c) are active, which can be solved using Lagrange minimization. We calculate the partial derivative of $L(\boldsymbol{z}, \boldsymbol{D}, \boldsymbol{\lambda})$,

$$\frac{\partial L}{\partial z_{k,p}} = \frac{\alpha_k\sigma^2_{N_{k,p}}}{2(1-\sigma^2_{N_{k,p}}z_{k,p})} + \sum_{i=1}^{k}\frac{\alpha_i - \alpha_{i-1}}{2}\left(\frac{1}{\sigma^2_{X_p}} + \sum_{j=i}^{K} z_{j,p}\right)^{-1} + \lambda_p, k = 1, 2, ..., K, p = 1, 2, ..., P \tag{92}$$

$$\frac{\partial L}{\partial D_p} = \frac{\lambda_p}{D_p^2} + \lambda_{P+1}, p = 1, 2, ..., P \tag{93}$$

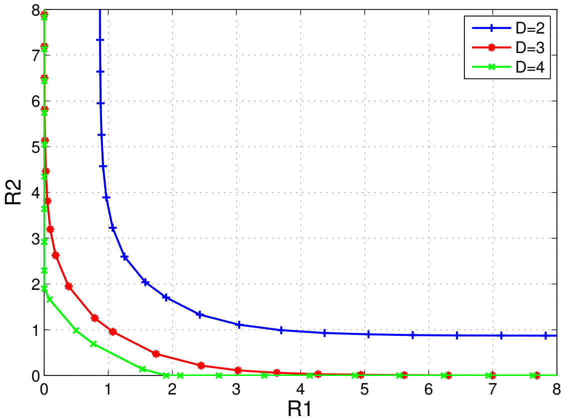


Fig. 7. An example of the rate region $\mathcal{R}_{\star}(D)$ for different distortion.

$$\frac{\partial L}{\partial \lambda_p} = \sum_{k \in \mathcal{K}} z_{k,p} - \frac{1}{D_p}, p = 1, 2, ..., P \tag{94}$$

$$\frac{\partial L}{\partial \lambda_{P+1}} = \sum_{p=1}^{P} D_p - D. \tag{95}$$

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 rate region boundary of a distributed learning 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. 7 shows the rate region $\mathcal{R}_{\star}(D)$ for different distortion. It is observed that given distortion D, the rate R_1 decreases 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.

B. *Sum-rate-distortion function*

In most cases of analysing communication efficiency of distributed learning, such as model compress and aggregation frequency control, we only care about the sum-rate-distortion function. Sum-rate-distortion function are defined by

$$R_{sum}(D) \triangleq \min_{(R_1, R_2, ..., R_K) \in \mathcal{R}_{\star}(D)} \sum_{k \in \mathcal{K}} R_k. \tag{96}$$

Recall that the rate region $\mathcal{R}_\star(D)$ in (66), we have

$$\min_{(R_1,R_2,...,R_K)\in\mathcal{R}_\star(D)}\sum_{k\in\mathcal{K}}R_k=\min_{D_p}\min_{r_{k,p}}\min_{R_{k,p}}\sum_{k\in\mathcal{K}}\sum_{p=1}^{P}R_{k,p}. \tag{97}$$

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

$$\begin{aligned}\mathcal{P}_2:\quad &\min_{D_p,r_{k,p}}\quad \sum_{p=1}^{P}\left[\sum_{k\in\mathcal{K}}r_{k,p}+\frac{1}{2}\log(1+\frac{\sigma^2_{X_p}}{D_p})\right] &\text{(98a)}\\ &s.t.\qquad \sum_{k\in\mathcal{K}}\frac{1-\exp(-2r_{k,p})}{\sigma^2_{N_{k,p}}}\geq\frac{1}{D_p},\quad p=1,2,...,P &\text{(98b)}\\ &\qquad\qquad \sum_{p=1}^{P}D_p\leq D. &\text{(98c)}\end{aligned}$$

Obviously, it is a convex optimization problem, and we can easily know that the inequality constraints (98b) and (98c) must be active. Therefore it can be solved using Lagrange minimization. For the method of Lagrange multipliers,

$$L(\boldsymbol{r},\boldsymbol{D},\boldsymbol{\lambda})=\sum_{p=1}^{P}\left[\sum_{k\in\mathcal{K}}r_{k,p}+\frac{1}{2}\log(1+\frac{\sigma^2_{X_p}}{D_p})\right]+\sum_{p=1}^{P}\lambda_p(\sum_{k\in\mathcal{K}}\frac{1-\exp(-2r_{k,p})}{\sigma^2_{N_{k,p}}}-\frac{1}{D_p})+\lambda_{P+1}(\sum_{p=1}^{P}D_p-D). \tag{99}$$

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 $\lambda_p,p=1,2,...,P$,

$$\frac{\partial L}{\partial r_{k,p}}=1+\lambda_p\frac{2\exp(-2r_{k,p})}{\sigma^2_{N_{k,p}}},k=1,2,...,K,p=1,2,...,P \tag{100}$$

$$\frac{\partial L}{\partial\lambda_p}=\sum_{k\in\mathcal{K}}\frac{1-\exp(-2r_{k,p})}{\sigma^2_{N_{k,p}}}-\frac{1}{D_p},p=1,2,...,P. \tag{101}$$

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,

$$\lambda^*_p=\frac{K}{2}(\frac{1}{D_p}-\sum_{k\in\mathcal{K}}\frac{1}{\sigma^2_{N_{k,p}}})^{-1},p=1,2,...,P \tag{102}$$

$$r^*_{k,p}=-\frac{1}{2}\log\left[\frac{\sigma^2_{N_{k,p}}}{K}(\sum_{k\in\mathcal{K}}\frac{1}{\sigma^2_{N_{k,p}}}-\frac{1}{D_p})\right],k=1,2,...,K,p=1,2,...,P. \tag{103}$$

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}

$$\frac{\partial L}{\partial D_p} = \frac{\frac{\sigma_{X_p}^2}{D_p^2}}{2(1+\frac{\sigma_{X_p}^2}{D_p})} + \frac{\lambda_p}{D_p^2} + \lambda_{P+1}, p=1,2,...,P \tag{104}$$

$$\frac{\partial L}{\partial \lambda_{P+1}} = \sum_{p=1}^{P} D_p - D. \tag{105}$$

Substituting λ_p^* in (102) back to (104), 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

$$\lambda_{P+1}^* = \frac{K}{2D_p^{*2}}(\sum_{k\in\mathcal{K}}\frac{1}{\sigma_{N_{k,p}}^2} - \frac{1}{D_p^*})^{-1} + \frac{1}{2D_p^{*2}}(\frac{1}{\sigma_{X_p}^2}+\frac{1}{D_p^*})^{-1}, p=1,2,...,P \tag{106}$$

$$\sum_{p=1}^{P} D_p^* = D. \tag{107}$$

We find that λ_{P+1}^* decreases monotonically with D_p^* for all $p=1,2,...,P$. The problem can be solved by finding λ_{P+1}^* such that $\sum_{p=1}^{P} D_p^*$ equals the given value of D.

C. 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 P2. Then, we can obtain a closed-form formula of the sum-rate-distortion function shown in the following corollary,

Corollary 2: For every $D>0$, when the global gradient and local gradients are identical over dimension p's,

$$R_{sum}(D) = P\left[-\frac{1}{2}\sum_{k\in\mathcal{K}}\log\left(\frac{\sigma_{N_k}^2}{K}(\sum_{k\in\mathcal{K}}\frac{1}{\sigma_{N_k}^2}-\frac{1}{D})\right)+\frac{1}{2}\log(1+\frac{\sigma_X^2}{D})\right]. \tag{108}$$

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 (106) we have $D_p^* = \frac{D}{P}, p=1,2,...,P$. Substituting $D_p^* = \frac{D}{P}, p=1,2,...,P$ back to

(103), we have

$$r^*_{k,p} = -\frac{1}{2}\log\left[\frac{\sigma^2_{N_k}}{K}(\sum_{k\in\mathcal{K}}\frac{1}{\sigma^2_{N_k}} - \frac{1}{D})\right], \tag{109}$$

Substituting $D^*_p, p = 1, 2, ..., P$ and $r^*_{k,p}, k = 1, 2, ..., K, p = 1, 2, ..., P$ back to problem P2, we conclude the corollary 1. ■

Corollary 3: For every $D > 0$, when the gradients are identical over dimension p's and local gradients are identical over device k's,

$$R_{sum}(D) = P\left[-\frac{K}{2}\log\left(1 - \frac{\sigma^2_N}{KD}\right) + \frac{1}{2}\log(1 + \frac{\sigma^2_X}{D})\right]. \tag{110}$$

Proof: It can be easily reduced from Corollary 2 by letting $\sigma^2_{N_1} = ... = \sigma^2_{N_K} = \sigma^2_N$. ■

Remark 3: The derived function has the form of a sum of two nonnegative functions. The first term decreases with the number of edge devices and is dominant for relatively small D. The second term is a classical channel capacity for a Gaussian channel with power constraint σ^2_X and noise variance D.

VI. Communication Efficiency of Distributed Learning

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

A. Convex Mini-batch SGD

Combining the sum-rate-distortion function given in Corollary 3 and the convergence guarantees [10] for mini-batch SGD algorithms on smooth, convex functions yields the following results:

Theorem 2 (Smooth Convex Optimization): Let $F, L, \boldsymbol{w}(0)$ and A be as in Theorem 1. Fix $\epsilon > 0$. Let gradient variance be diag$(\frac{\sigma^2_X}{P}, ..., \frac{\sigma^2_X}{P})$ and noise variance be diag$(\frac{\sigma^2_N}{P}, ..., \frac{\sigma^2_N}{P})$. 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 satisfy

$$T \geq A^2\left(\sqrt{\frac{D}{2\epsilon^2} + \frac{L}{\epsilon}} + \sqrt{\frac{D}{2\epsilon^2}}\right)^2 = O\left(A^2 \max(\frac{2D}{\epsilon^2}, \frac{L}{\epsilon})\right) \tag{111}$$

Moreover, the required communication bits at iteration t is given by

$$P\left[-\frac{K}{2}\log\left(1 - \frac{\sigma_N^2}{KD}\right) + \frac{1}{2}\log(1 + \frac{\sigma_X^2}{D})\right]. \tag{112}$$

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 [10], Theorem 6.3 we should have

$$A\sqrt{\frac{2D}{T}} + \frac{LA^2}{T} \leq \epsilon \Leftrightarrow T \geq A^2\left(\sqrt{\frac{D}{2\epsilon^2} + \frac{L}{\epsilon}} + \sqrt{\frac{D}{2\epsilon^2}}\right)^2. \tag{113}$$

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

Remark 4: In the most reasonable regimes, the number of iterations is dominated by the first term of the maximum in (111). Specifically, the number of iterations will depend linearly on the variance bound of the gradient estimator 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 (112), 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 Mini-batch 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 mini-batch SGD at least always provably converages to a local minima, when F is smooth. For instance, by applying Theorem 2.1 in [11], we immediately obtain the following communication efficiency results for distributed learning.

Theorem 3 (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

$$\frac{1}{L}\mathbb{E}[\|\nabla F(\boldsymbol{w})\|^2] \leq O\left(\frac{F(\boldsymbol{w}(0)) - F^*}{N} + \frac{D}{L}\right). \tag{114}$$

Moreover, the lower bound of communication bits at iteration t is the same as in Theorem 2.

VII. Conclusion

This paper studied the quadratic Gaussian CEO problem under unbiased estimation constraint. We have characterized the rate region by showing that the Berger-Tung achievable region is also tight in the case of unbiased estimation constraint. We also derived the explicit rate region characterized by the tangent hyperplanes and the closed-form sum-rate-distortion functions in special cases. Finally, we apply the sum-rate-distortion function to analyse the communication efficiency of distributed learning. 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 adaptive practical quantization and coding schemes based on the estimated gradient distribution to achieve the rate region results in distributed learning. We can also study the communication-efficient user scheduling schemes and aggregation frequency control schemes based on the rate region results derived in this paper.

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] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, Q. V. Le, M. Z. Mao, M. Ranzato, A. W. Senior, P. A. Tucker *et al.*, “Large scale distributed deep networks,” in *NIPS*, 2012.

[6] 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.

[7] N. Strom, “Scalable distributed dnn training using commodity gpu cloud computing,” in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[8] H. Qi, E. R. Sparks, and A. Talwalkar, “Paleo: A performance model for deep neural networks.” in *ICLR (Poster)*, 2017.

[9] D. Grubic, L. K. Tam, D. Alistarh, and C. Zhang, “Synchronous multi-gpu deep learning with low-precision communication: An experimental study,” in *Proceedings of the 21st International Conference on Extending Database Technology*. OpenProceedings, 2018, pp. 145–156.

[10] S. Bubeck *et al.*, “Convex optimization: Algorithms and complexity,” *Foundations and Trends® in Machine Learning*, vol. 8, no. 3-4, pp. 231–357, 2015.

[11] 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.

[12] J. Wang and G. Joshi, “Cooperative sgd: A unified framework for the design and analysis of communication-efficient SGD algorithms,” *arXiv preprint arXiv:1808.07576*, 2018.

[13] S. Zhang, A. Choromanska, and Y. LeCun, “Deep learning with elastic averaging SGD,” in *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*, 2015, pp. 685–693.

[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. Berger, Z. Zhang, and H. Viswanathan, “The CEO problem [multiterminal source coding],” *IEEE Transactions on Information Theory*, vol. 42, no. 3, pp. 887–902, 1996.

[24] H. Viswanathan and T. Berger, “The quadratic gaussian CEO problem,” *IEEE Transactions on Information Theory*, vol. 43, no. 5, pp. 1549–1559, 1997.

[25] 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.

[26] Y. Oohama, “Rate-distortion theory for gaussian multiterminal source coding systems with several side informations at the decoder,” *IEEE Transactions on Information Theory*, vol. 51, no. 7, pp. 2577–2593, 2005.

[27] V. Prabhakaran, D. Tse, and K. Ramachandran, “Rate region of the quadratic gaussian CEO problem,” in *International Symposium on Information Theory, 2004. ISIT 2004. Proceedings.*, 2004, pp. 119–.

[28] 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.

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

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

[31] 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.

[32] 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.

[33] 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.

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

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

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

[37] 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.