2021 IEEE International Symposium on Information Theory (ISIT) | 978-1-5386-8209-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISIT45174.2021.9517797

Analog Joint Source-Channel Coding for Distributed Functional Compression using Deep Neural Networks

Yashas Malur Saidutta, Afshin Abdi, and Faramarz Fekri
Dept. of Electrical and Computer Engineering,
Georgia Institute of Technology,
Atlanta, Georgia, USA
Email: {yashas.saidutta, abdi}@gatech.edu, fekri@ece.gatech.edu

***Abstract*—In this paper, we study Joint Source-Channel Coding (JSCC) for distributed analog functional compression over both Gaussian Multiple Access Channel (MAC) and AWGN channels. Notably, we propose a deep neural network based solution for learning encoders and decoders. We propose three methods of increasing performance. The first one frames the problem as an autoencoder; the second one incorporates the power constraint in the objective by using a Lagrange multiplier; the third method derives the objective from the information bottleneck principle. We show that all proposed methods are variational approximations to upper bounds on the indirect rate-distortion problem's minimization objective. Further, we show that the third method is the variational approximation of a tighter upper bound compared to the other two. Finally, we show empirical performance results for image classification. We compare with existing work and showcase the performance improvement yielded by the proposed methods.**

I. Introduction

With the number of IoT devices set to exceed 75 billion by 2025 [1] and promising new applications like cooperative autonomous driving [2], it is important to design systems where distributed sensors communicate efficiently in a target-aware manner. Analog communication based JSCC has seen a recent resurgence due to its attractive robustness properties to channel conditions [3], [4]. Along these lines, in this paper, we study the problem of analog distributed functional joint source-channel coding (JSCC) using deep neural networks (DNN).

Fig. 1 shows the problem setup. Consider an example where separate wireless cameras observe different parts of a scene that has to be classified. We know that a system that uses all the parts in predicting the class label will perform better than any system using only a single part. Instead of transmitting all the observed raw data to a centralized location, it is more efficient in terms of communication to send the features relevant for the classification task. Further, in delay critical applications, employing asymptotic coding schemes may not be viable. Thus, in this paper, we study the problem where separate sensors independently encode a single sample for transmission across a noisy channel. The central receiver uses this received data for classification.

We categorize the prior works as below:

Deep neural networks for JSCC: Many works perform JSCC for data recovery using DNNs over orthogonal channels [3]–[9], and MAC [10]. Recent works have looked at the centralized analog functional JSCC problem [11]–[13]. Very recently, [13] proposed a system for analog JSCC for face recognition from multiple cameras. However, their method is similar to our first baseline method, and subsequent methods proposed in this paper show better performance.

Distributed Functional Compression: Unlike the problem here, the CEO problem assumes that given the random variable to be reconstructed, the observations at the sensors are independent [14]–[19]. Further, in contrast to our one sample encoding, others have studied asymptotic methods [20]–[22]. Additionally, many works study the problem for linear target functions [23]–[25]. Also, none of these works perform JSCC.

Distributed Functional JSCC over MAC: Beginning with [26], many works have looked into this problem when target functions have known simple forms [27]–[29]. Interestingly [30] showed the existence of discontinuous universal encoding functions when the sensors observe scalar variables.

Variational Information Bottleneck: Information Bottleneck was proposed as a generalization of Rate-Distortion Theory [31] and used in centralized applications [32], [33]. Recently, [34]–[36] introduced distributed variational information bottleneck. However, their objective function simplifications make use of independence assumptions similar to the CEO problem and also do not look at JSCC.

Distributed Learning: Model parallelism in distributed learning assumes some communication between nodes [37], [38].

Miscellaneous: The work [39] on distributed quantization for classification is the closest to our work. However, not accounting for the communication channel leads to a different training criterion.

The contributions of the paper are

1) We propose three deep learning based methods to perform analog JSCC for distributed functional compres-

This work is supported by National Science Foundation under award ID MLWiNS-2003002 and a Gift from Intel Co.

978-1-5386-8209-8/21/$31.00 ©2021 IEEE

sion for use over Gaussian MAC and AWGN channels.

2) We showcase the theoretical connection of proposed methods to the indirect rate-distortion problem. We also theoretically contrast the proposed methods.
3) Finally, we show empirical results on the CIFAR-10 dataset to validate these methods and insights.

Notation: Bold uppercase letters denote random variables. Bold lowercase letters denote the samples.

II. Problem Definition

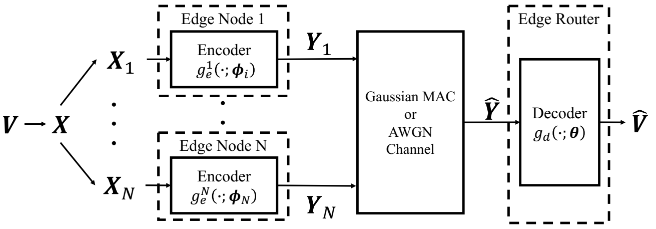

Fig. 1: JSCC for functional compression.

Fig. 1 shows the problem under consideration where $\boldsymbol{V}$ is an information source and $\boldsymbol{X}$ is a correlated random variable such that $I(\boldsymbol{V};\boldsymbol{X}) > 0$. The sensor nodes observe parts of the correlated random variable $\boldsymbol{X}$. We represent the random variable observed by the edge node-i as $\boldsymbol{X}_i$ where $i \in \{1, ..., N\}$ indexes the nodes. We assume that knowing the value of $\boldsymbol{x}$ is equivalent to knowing all the values of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ and vice-versa. Each Node-i uses some deterministic encoding function denoted as a mapping $g_e^i(\cdot;\boldsymbol{\phi}_i)$ where $\boldsymbol{\phi}_i$ are the parameters. $\boldsymbol{Y}_i \in \mathbb{R}^K$ is used to denote the signal transmitted by node-i. The transmitters are subject to an average power constraint of the form $\frac{1}{KN}\sum_{i=1}^N \mathbb{E}\left[\|\boldsymbol{Y}_i\|_2^2\right] \leq P_T$. $\hat{\boldsymbol{Y}}$ represents the noisy received signal at the receiver. We represent the decoder by a function $g_d(\cdot;\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is its parameters. The output of the decoder is the recovered value of $\boldsymbol{v}$ denoted by $\hat{\boldsymbol{v}}$. The objective of this problem is to minimize the distortion between $\boldsymbol{V}$ and $\hat{\boldsymbol{V}}$ while satisfying a rate constraint $I(\boldsymbol{X};\hat{\boldsymbol{V}}) \leq C$ where C is the channel capacity. Revisiting the earlier example, $\boldsymbol{x}_i$ corresponds to different parts of the scene observed by the cameras (edge nodes), and $\boldsymbol{v}$ is the unknown true class label of the observed scene.

In the case of the orthogonal AWGN channel, the received noisy signal $\hat{\boldsymbol{Y}}$ is given by

$$\hat{\boldsymbol{Y}} = \begin{bmatrix}\hat{\boldsymbol{Y}}_1\\ \vdots \\ \hat{\boldsymbol{Y}}_N\end{bmatrix} = \begin{bmatrix}\boldsymbol{Y}_1\\ \vdots \\ \boldsymbol{Y}_N\end{bmatrix} + \begin{bmatrix}\boldsymbol{Z}_1\\ \vdots \\ \boldsymbol{Z}_N\end{bmatrix}. \tag{1}$$

Here, $\boldsymbol{Z}_i \sim \mathcal{N}\left(\boldsymbol{0}, \sigma_z^2 \boldsymbol{I}_K\right)$ are iid random variables representing noise. $\boldsymbol{I}_K$ represents the identity matrix of dimension K and σ_z^2 is the noise power. In the Gaussian MAC scenario the received noisy signal $\hat{\boldsymbol{Y}}$ is given by

$$\hat{\boldsymbol{Y}} = \sum_{i=1}^N \boldsymbol{Y}_i + \boldsymbol{Z}, \tag{2}$$

where $\boldsymbol{Z} \sim \mathcal{N}\left(\boldsymbol{0}, \sigma_z^2 \boldsymbol{I}_K\right)$ is the noise. The Channel Signal to Noise Ratio (CSNR) in dB is defined as $10\log_{10}\left(1 + \frac{P_T}{\sigma_z^2}\right)$.

III. Methodology

In this section, we propose three methods. The first one frames the problem as an autoencoder; the second one incorporates the power constraint in the objective by using a Lagrange multiplier; the third method derives the objective from the information bottleneck principle [31].

A. DiFJ-AU: Distributed Functional JSCC via Autoencoders

In this methodology, we train the system to minimize the empirical distortion loss as

$$\mathcal{L}_{AU} = \mathbb{E}_{\boldsymbol{V},\boldsymbol{X},\boldsymbol{Z}}\left[\mathcal{D}_{\boldsymbol{V}}\left(\boldsymbol{v}, g_d\left(\sum_{i=1}^N g_e^i(\boldsymbol{x}_i;\boldsymbol{\theta}_i) + \boldsymbol{z};\boldsymbol{\phi}\right)\right)\right]. \tag{3}$$

Here $\mathcal{D}_{\boldsymbol{V}}(\cdot,\cdot)$ is some differentiable distortion measure used to compute the discrepancy between the true value $\boldsymbol{v}$ and its recovered value $\hat{\boldsymbol{v}}$. To ensure the power constraint is satisfied, the output of the encoding function $\boldsymbol{y}_i$ is normalized as $\tilde{\boldsymbol{y}}_i = \sqrt{KP_T}\frac{\boldsymbol{y}_i}{\|\boldsymbol{y}_i\|_2}$ prior transmission. In this method, the power constraint implicitly enforces the rate constraint. Prior works have popularized this approach in joint source-channel coding [3], [12]. In the absence of prior works, this method will serve as a strong performing baseline.

B. DiFJ-SL: Distributed Functional JSCC via Standard Lagrangian

The optimization problem is written as

$$\underset{\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_N,\boldsymbol{\theta}}{\arg\min}\ \mathbb{E}_{\boldsymbol{V},\boldsymbol{X},\boldsymbol{Z}}\left[\mathcal{D}_{\boldsymbol{V}}\left(\boldsymbol{v}, g_d\left(\sum_{i=1}^N g_e^i(\boldsymbol{x}_i;\boldsymbol{\theta}_i) + \boldsymbol{z};\boldsymbol{\phi}\right)\right)\right]$$
$$\text{such that } \frac{1}{KN}\sum_{i=1}^N \mathbb{E}\left[\|\boldsymbol{Y}_i\|_2^2\right] \leq P_T. \tag{4}$$

The constrained optimization problem in (4) is converted into an unconstrained optimization problem by using Lagrange multipliers. The loss function for minimization is

$$\mathcal{L}_{SL} = \mathbb{E}_{\boldsymbol{V},\boldsymbol{X},\boldsymbol{Z}}\left[\mathcal{D}_{\boldsymbol{V}}\left(\boldsymbol{v}, g_d\left(\sum_{i=1}^N g_e^i(\boldsymbol{x}_i;\boldsymbol{\theta}_i) + \boldsymbol{z};\boldsymbol{\phi}\right)\right)\right.$$
$$\left. + \lambda\frac{1}{KN}\sum_{i=1}^N \|\boldsymbol{Y}_i\|_2^2\right]. \tag{5}$$

Here λ is the Lagrange multiplier. Similar to the earlier method, the power constraint implicitly enforces the rate constraint.

C. DiFJ-VIB: Distributed Functional JSCC via Variational Information Bottleneck

The information bottleneck principle proposed by [31] is the generalization of the rate-distortion theory proposed by Shannon [40]. The rate-distortion theory is a principled way of finding a compressed representation of $\boldsymbol{X}$, denoted by

$\hat{\boldsymbol{Y}}$, that minimizes $\mathcal{D}_{\boldsymbol{X}}(\boldsymbol{X}, \hat{\boldsymbol{Y}})$ while simultaneously ensuring $I(\boldsymbol{X}; \hat{\boldsymbol{Y}}) \leq R$, where R is the rate constraint. The choice of the distortion measure governs which features are relevant and preserved in the compressed representation. However, [31] proposed the use of another random variable $\boldsymbol{V}$ to determine the features relevant for preservation. Thus, the information bottleneck theory is a principled way of finding a compressed representation of $\boldsymbol{X}$, denoted by $\hat{\boldsymbol{Y}}$, that maximizes $I(\boldsymbol{V}; \hat{\boldsymbol{Y}})$ while ensuring $I(\boldsymbol{X}; \hat{\boldsymbol{Y}}) \leq R$.

The resulting minimization objective is written as

$$\underset{p(\hat{\boldsymbol{y}}|\boldsymbol{x})}{\arg\min} -I\left(\boldsymbol{V}; \hat{\boldsymbol{Y}}\right) + \lambda I\left(\boldsymbol{X}; \hat{\boldsymbol{Y}}\right). \tag{6}$$

However, it is not possible to compute these quantities in closed form. Hence, we use variational upper bounds [32] like

$$\begin{aligned} &- \mathbb{E}_{\boldsymbol{V},\hat{\boldsymbol{Y}}}\left[\log\left(q(\boldsymbol{v} \mid \hat{\boldsymbol{y}})\right)\right] - H(\boldsymbol{V}) \\ &\quad + \lambda \mathbb{E}_{\boldsymbol{X},\hat{\boldsymbol{Y}}}\left[\log p(\hat{\boldsymbol{y}} \mid \boldsymbol{x})\right] - \lambda \mathbb{E}_{\boldsymbol{X},\hat{\boldsymbol{Y}}}\left[\log r(\hat{\boldsymbol{y}})\right]. \end{aligned} \tag{7}$$

Here, $q(\boldsymbol{v}|\hat{\boldsymbol{y}})$ is the variational approximation to $p(\boldsymbol{v}|\hat{\boldsymbol{y}})$, $r(\hat{\boldsymbol{y}})$ is the variational approximation to $p(\hat{\boldsymbol{y}})$.

Based on our problem setup in Fig. 1, we make further simplification to (7). $-H(\boldsymbol{V})$ is a constant w.r.t. the encoder and decoder parameters. Since the encoder is deterministic and the noise is independent of the encoding values, computing expectations w.r.t $\boldsymbol{X}, \hat{\boldsymbol{Y}}$ is equivalent to computing expectations w.r.t. $\boldsymbol{X}, \boldsymbol{Z}$. Since the goal is to reduce the distortion measure $\mathcal{D}_{\boldsymbol{V}}(\cdot, \cdot)$ between the true and the predicted value, we model the variational distribution $q(\boldsymbol{v}|\hat{\boldsymbol{y}})$ as $\frac{1}{Z}\exp^{-\mathcal{D}_{\boldsymbol{V}}(\boldsymbol{v}, g_d(\hat{\boldsymbol{y}}))}$ where Z is the normalization constant. We use a separate external parametric model to represent the variational distribution $r(\hat{\boldsymbol{y}})$. Fig. 1 shows that $\boldsymbol{V} \rightarrow \boldsymbol{X} \rightarrow (\boldsymbol{X}_1, \ldots \boldsymbol{X}_N) \rightarrow (\boldsymbol{Y}_1, \ldots \boldsymbol{Y}_N) \rightarrow \hat{\boldsymbol{Y}} \rightarrow \hat{\boldsymbol{V}}$ forms a Markov Chain. Also, $\boldsymbol{Y}_i \perp\!\!\!\perp \boldsymbol{X}_j | \boldsymbol{X}_i \; \forall i \neq j$. Hence the joint distribution decomposes as

$$\begin{aligned} p(&\boldsymbol{x}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_N, \hat{\boldsymbol{y}}) \\ &= p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N | \boldsymbol{x}) \left(\prod_{i=1}^{N} p(\boldsymbol{y}_i | \boldsymbol{x}_i)\right) p(\hat{\boldsymbol{y}} | \boldsymbol{y}_1, \ldots, \boldsymbol{y}_N). \end{aligned} \tag{8}$$

From our problem definition, knowledge of $\boldsymbol{x}$ provides complete knowledge about the samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$. Hence, $p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N | \boldsymbol{x})$ is a delta-dirac function at the appropriate value of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ determined by $\boldsymbol{x}$. Since the encoders are deterministic functions, $p(\boldsymbol{y}_i | \boldsymbol{x}_i) = \delta\left(\boldsymbol{y}_i - g_e^i(\boldsymbol{x}_i)\right)$. If the communication is over a Gaussian MAC $p(\hat{\boldsymbol{y}} | \boldsymbol{y}_1, \ldots, \boldsymbol{y}_N) = p_{\boldsymbol{Z}}\left(\hat{\boldsymbol{y}} - \sum_i \boldsymbol{y}_i\right)$. Hence, $\mathbb{E}_{\boldsymbol{X},\hat{\boldsymbol{Y}}}\left[\log p(\hat{\boldsymbol{y}} \mid \boldsymbol{x})\right] = H(\boldsymbol{Z})$, where $H(\cdot)$ signifies the entropy. If the channel is orthogonal AWGN then, $p(\hat{\boldsymbol{y}} | \boldsymbol{y}_1, \ldots, \boldsymbol{y}_N) = \prod_{i=1} p_{\boldsymbol{Z}_i}\left(\hat{\boldsymbol{y}}_i - \boldsymbol{y}_i\right)$. Hence, $\mathbb{E}_{\boldsymbol{X},\hat{\boldsymbol{Y}}}\left[\log p(\hat{\boldsymbol{y}} \mid \boldsymbol{x})\right] = \sum_i H(\boldsymbol{Z}_i)$. Both $H(\boldsymbol{Z})$ and $\sum_i H(\boldsymbol{Z}_i)$ are constant w.r.t. the encoder and decoder parameters. Thus, the objective becomes

$$\begin{aligned} \mathcal{L}_{IB} = \mathbb{E}_{\boldsymbol{X},\boldsymbol{V},\boldsymbol{Z}} &\left[\mathcal{D}_{\boldsymbol{V}}\left(\boldsymbol{v}, g_d\left(\sum_{i=1}^{N} g_e^i(\boldsymbol{x}_i) + \boldsymbol{z}\right)\right)\right. \\ &\left. - \lambda \log\left(r\left(\sum_{i=1}^{N} g_e^i(\boldsymbol{x}_i) + \boldsymbol{z}\right)\right)\right]. \end{aligned} \tag{9}$$

Unlike the previously proposed methods, the second term in the loss function explicitly imposes the rate constraint.

D. Theoretical connections to the Indirect Rate-Distortion Problem

This subsection shows that all objectives are variational approximations to upper bounds on the indirect rate-distortion problem's minimization objective [41], [42]. In the presentation of both the theorems here, we assume a Gaussian MAC; however, they apply to any channel with an independent additive noise component.

The work closest in deriving the rate-distortion function for distributed functional computation is that of [22, Theorem 43]. Even though our power constraint is similar to the sum rate constraint in (14) of [22], they require computation of graph entropies that are neither easy to compare or compute. Hence, we compare it with the single letter indirect rate-distortion problem for the source $\boldsymbol{X} = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N]$. In the indirect rate-distortion problem [41], the encoder observes the source $\boldsymbol{V}$ through a noisy channel whose output is $\boldsymbol{X}$. The receiver uses the encoded $\boldsymbol{x}$ got through another noisy communication channel to recover the value $\boldsymbol{v}$ [42] . The rate-distortion objective for this problem is [43]

$$\min_{p(\hat{\boldsymbol{v}}|\boldsymbol{x}): I(\boldsymbol{X}; \hat{\boldsymbol{V}}) \leq R} \mathbb{E}_{\boldsymbol{V},\hat{\boldsymbol{V}}}\left[\mathcal{D}_{\boldsymbol{V}}(\boldsymbol{v}, \hat{\boldsymbol{v}})\right]. \tag{10}$$

For, JSCC $R \leq C$ where C is the capacity of the communication channel. Hence, we can write (10) as an unconstrained Lagrangian optimization of the form

$$\min_{p(\hat{\boldsymbol{v}}|\boldsymbol{x})} \mathbb{E}_{\boldsymbol{V},\hat{\boldsymbol{V}}}\left[\mathcal{D}(\boldsymbol{v}, \hat{\boldsymbol{v}})\right] + \lambda\left(I(\boldsymbol{X}; \hat{\boldsymbol{V}}) - R\right). \tag{11}$$

Here λ is the Lagrange multiplier.

Theorem 1. *For a fixed λ along with deterministic encoders and decoder, the DiFJ-VIB objective function for the system in Fig. 1*

$$\begin{aligned} \mathbb{E}_{\boldsymbol{X},\boldsymbol{V},\boldsymbol{Z}} &\left[\mathcal{D}_{\boldsymbol{V}}\left(\boldsymbol{v}, g_d\left(\sum_{i=1}^{N} g_e^i(\boldsymbol{x}_i) + \boldsymbol{z}\right)\right)\right. \\ &\left. - \lambda \log\left(r\left(\sum_{i=1}^{N} g_e^i(\boldsymbol{x}_i) + \boldsymbol{z}\right)\right)\right] - H(\boldsymbol{Z}) \end{aligned} \tag{12}$$

is an upper bound on the indirect rate-distortion problem objective

$$\mathbb{E}_{\boldsymbol{V},\hat{\boldsymbol{V}}}\left[\mathcal{D}_{\boldsymbol{V}}(\boldsymbol{v}, \hat{\boldsymbol{v}})\right] + \lambda I(\boldsymbol{X}; \hat{\boldsymbol{V}}) \tag{13}$$

Proof. From Fig. 1, $\boldsymbol{X} \rightarrow \hat{\boldsymbol{Y}} \rightarrow \hat{\boldsymbol{V}}$ is a Markov Chain. Hence, $I\left(\boldsymbol{X}; \hat{\boldsymbol{V}}\right) \leq I\left(\boldsymbol{X}; \hat{\boldsymbol{Y}}\right)$. Further, $I\left(\boldsymbol{X}; \hat{\boldsymbol{Y}}\right) = H(\hat{\boldsymbol{Y}}) - H(\boldsymbol{Z})$ (refer simplification made in Sec. III-C). Thus, we can write an upper bound on (13) as

$$\mathbb{E}_{\boldsymbol{V},\hat{\boldsymbol{V}}}\left[\mathcal{D}_{\boldsymbol{V}}\left(\boldsymbol{v}, \hat{\boldsymbol{v}}\right)\right] + \lambda H(\hat{\boldsymbol{Y}}) - H(\boldsymbol{Z}). \tag{14}$$

Since the encoders and the decoder are deterministic, we can simplify (14) as

$$\mathbb{E}_{\boldsymbol{X},\boldsymbol{V},\boldsymbol{Z}}\left[\mathcal{D}_{\boldsymbol{V}}\left(\boldsymbol{v}, g_d\left(\sum_{i=1}^{N} g_e^i\left(\boldsymbol{x}_i\right)+\boldsymbol{z}\right)\right)\right.$$
$$\left.-\lambda\log\left(p_{\hat{\boldsymbol{Y}}}\left(\sum_{i=1}^{N} g_e^i\left(\boldsymbol{x}_i\right)+\boldsymbol{z}\right)\right)\right]-H(\boldsymbol{Z}). \quad (15)$$

For any random variable $\hat{\boldsymbol{Y}}$ with true distribution $p_{\hat{\boldsymbol{Y}}}(\cdot)$, any arbitrary (variational) approximation $r_{\hat{\boldsymbol{Y}}}(\cdot)$ satisfies $H(\hat{\boldsymbol{Y}}) \leq -\mathbb{E}_{\hat{\boldsymbol{y}}\sim p_{\hat{\boldsymbol{Y}}}(\cdot)}\left[\log r_{\hat{\boldsymbol{Y}}}(\hat{\boldsymbol{y}})\right]$. Thus, we get an upper bound on (15) as

$$\mathbb{E}_{\boldsymbol{X},\boldsymbol{V},\boldsymbol{Z}}\left[\mathcal{D}_{\boldsymbol{V}}\left(\boldsymbol{v}, g_d\left(\sum_{i=1}^{N} g_e^i\left(\boldsymbol{x}_i\right)+\boldsymbol{z}\right)\right)\right.$$
$$\left.-\lambda\log\left(r_{\hat{\boldsymbol{Y}}}\left(\sum_{i=1}^{N} g_e^i\left(\boldsymbol{x}_i\right)+\boldsymbol{z}\right)\right)\right]-H(\boldsymbol{Z}). \quad (16)$$

□

Theorem 2. *For a fixed λ along with deterministic encoders and decoder, the objective functions*

$$\mathbb{E}_{\boldsymbol{V},\boldsymbol{X},\boldsymbol{Z}}\left[\mathcal{D}_{\boldsymbol{V}}\left(\boldsymbol{v}, g_d\left(\sum_{i=1}^{N} g_e^i\left(\boldsymbol{x}_i\right)+\boldsymbol{z}\right)\right)\right]+B_1, \quad (17a)$$

$$\mathbb{E}_{\boldsymbol{V},\boldsymbol{X},\boldsymbol{Z}}\left[\mathcal{D}_{\boldsymbol{V}}\left(\boldsymbol{v}, g_d\left(\sum_{i=1}^{N} g_e^i\left(\boldsymbol{x}_i\right)+\boldsymbol{z}\right)\right)\right]$$
$$+\beta\frac{1}{KN}\sum_{i=1}^{N}\mathbb{E}_{\boldsymbol{X}}\left[\left\|g_e^i\left(\boldsymbol{x}_i\right)\right\|_2^2\right]+B_2 \quad (17b)$$

where B_1 and B_2 are constants, are the variational approximations to an upper bound on the indirect rate-distortion problem objective

$$\mathbb{E}_{\boldsymbol{V},\hat{\boldsymbol{V}}}\left[\mathcal{D}_{\boldsymbol{V}}(\boldsymbol{v},\hat{\boldsymbol{v}})\right]+\lambda I(\boldsymbol{X};\hat{\boldsymbol{V}}). \quad (18)$$

Proof. From Fig. 1, $\boldsymbol{X} \rightarrow (\boldsymbol{Y}_1,\dots\boldsymbol{Y}_N) \rightarrow \hat{\boldsymbol{Y}} \rightarrow \hat{\boldsymbol{V}}$ is a Markov Chain. Hence $I\left(\boldsymbol{X};\hat{\boldsymbol{V}}\right) \leq I\left(\boldsymbol{X};(\boldsymbol{Y}_1,\dots\boldsymbol{Y}_N)\right)$. We also note that $H\left((\boldsymbol{Y}_1,\dots\boldsymbol{Y}_N)\,|\boldsymbol{X}\right)=0$ because the encoders are deterministic, and the mapping between $\boldsymbol{X}$ to $(\boldsymbol{X}_1,\dots\boldsymbol{X}_N)$ is one-to-one. Thus (18) is upper bounded by

$$\mathbb{E}_{\boldsymbol{V},\hat{\boldsymbol{V}}}\left[\mathcal{D}_{\boldsymbol{V}}(\boldsymbol{v},\hat{\boldsymbol{v}})\right]+\lambda H\left(\boldsymbol{Y}_1,\dots\boldsymbol{Y}_N\right). \quad (19)$$

Since for any random variable $(\boldsymbol{Y}_1,\dots\boldsymbol{Y}_N)$ with true distribution $p_{\boldsymbol{Y}_1,\dots\boldsymbol{Y}_N}(\cdot)$, any arbitrary (variational) approximation $\prod_{i=1}^{N} q_i(\boldsymbol{y}_i)$ satisfies $H(\boldsymbol{Y}) \leq -\mathbb{E}_{\boldsymbol{Y}_1,\dots\boldsymbol{Y}_N}\left[\sum_{i=1}^{N}\log q_i(\boldsymbol{y}_i)\right]$. Thus, we get an upper bound on (15). Further, using the deterministic behavior of the encoders and the decoder, we can write the upper bound on (15) as

$$\mathbb{E}_{\boldsymbol{V},\boldsymbol{X},\boldsymbol{Z}}\left[\mathcal{D}_{\boldsymbol{V}}\left(\boldsymbol{v}, g_d\left(\sum_{i=1}^{N} g_e^i\left(\boldsymbol{x}_i\right)+\boldsymbol{z}\right)\right)\right]$$
$$-\lambda\sum_{i=1}^{N}\mathbb{E}_{\boldsymbol{X}}\left[\log(q_i(g_e^i\left(\boldsymbol{x}_i\right)))\right]+A+\log(Z). \quad (20)$$

In the DiFJ-AU system, the encoders ensure that the encoded samples $\boldsymbol{y}_i$ always have the norm $\sqrt{KP_T}$. Thus, the support of $\boldsymbol{Y}_i$ is an improper subset of the surface of a hypersphere in K dimensions with radius $\sqrt{KP_T}$. In this system, if we assume that the variational approximation $q_i(\boldsymbol{y}_i)$ is a uniform distribution over the support of $\boldsymbol{Y}_i$ and to be of the form $q_i(\boldsymbol{y}_i)=\frac{1}{D_i}$. Then, we can see that (20) is the same as (17a) where $B_1=\lambda\sum_i\log(D_i)+A+\log(Z)$.

If we assume $q_i(\boldsymbol{y}_i)=\mathcal{N}\left(\boldsymbol{y}_i;\boldsymbol{0},P_T\boldsymbol{I}_K\right)$, then (20) takes the form of (17b). Here, β corresponds to $\frac{KN\lambda}{P_T}$ and $B_2 = A+log(Z)+\frac{K\lambda}{2}\log(2\pi eP_T)$. □

Remark 1. *All three loss functions $\mathcal{L}_{AU}$, $\mathcal{L}_{SL}$, and $\mathcal{L}_{IB}$ are variational approximations to upper bounds on the indirect rate-distortion problem objective. Since $\boldsymbol{X} \rightarrow (\boldsymbol{X}_1,\dots\boldsymbol{X}_N) \rightarrow (\boldsymbol{Y}_1,\dots\boldsymbol{Y}_N) \rightarrow \hat{\boldsymbol{Y}} \rightarrow \hat{\boldsymbol{V}}$ is a Markov Chain, then $I(\boldsymbol{X};\boldsymbol{Y}_1,\dots,\boldsymbol{Y}_N) \geq I(\boldsymbol{X};\hat{\boldsymbol{Y}}) \geq I(\boldsymbol{X};\hat{\boldsymbol{V}})$. Thus, $\mathcal{L}_{IB}$ is the variational approximation to a **tighter upper bound** on the indirect rate-distortion objective.*

IV. Experimental Results and Discussion

A. Implementation Details

We use the CIFAR-10 dataset for showing empirical results on the proposed methods [44]. Here, we use 45000 images for training, 5000 for validation, and 10000 for testing. We report results over ten transmissions for all the images in the test set. We divide each image into equal-sized square patches made available to the encoders; these patches represent the $\boldsymbol{x}_i$s. Let us consider the case of $N=4$ encoders. We divide the original CIFAR-10 image into four disjoint quadrants. We show the first quadrant to the encoder in edge node-1, the second quadrant to edge node-2, and so on. The random variable $\boldsymbol{v}$ is the class label. Our objective is to recover this label at the edge router.

We model the distribution over the received $\hat{\boldsymbol{Y}}$, $r_{\hat{\boldsymbol{Y}}}(\cdot)$ as a product of independent Generalized Gaussian Distributions (GGD), one for each dimension of $\hat{\boldsymbol{Y}}$.

$$r_{\hat{\boldsymbol{Y}}}(\hat{\boldsymbol{y}})=\prod_{k=1}^{K}\frac{\beta_k}{2\alpha_k\Gamma\left(\frac{1}{\beta_k}\right)}\exp^{-\left(\frac{|\hat{\boldsymbol{y}}[k]-\mu_k|}{\alpha_k}\right)^{\beta_k}} \quad (21)$$

where $\hat{\boldsymbol{y}}[k]$ is the k^{th} element of the vector $\hat{\boldsymbol{y}}$; $\alpha_k>0$ and $\beta_k>0$ are the scale and shape parameters, respectively, and μ_k is the mean. We learn the parameters by performing gradient descent on the negative log-likelihood of the observed samples of $\hat{\boldsymbol{y}}$. For low values of CSNR, we found that restricting $\beta_1=\cdots=\beta_k=2$ and $\alpha_1=\cdots=\alpha_k$ helped prevent overfitting.

TABLE I: Architecture of the DNNs used for the system in Fig. 1.

Name	Architecture Details
VGG_Block(F)	[Conv(F,3 $\times$ 3),Conv(F,3 $\times$ 3),MaxPool(2 $\times$ 2)]
Gaussian MAC - Encoder	[VGG_Block(32),VGG_Block(64),VGG_Block(128),FCN(1024),FCN(512),FCN(K)]
Gaussian MAC - Decoder	[FCN(512),FCN(1024),FCN(2048),FCN(128),FCN(10)]
Orthogonal AWGN - Encoder	[VGG_Block(64),VGG_Block(128),VGG_Block(512),VGG_Block(512),FCN(1024),FCN(512),FCN(K)]
Orthogonal AWGN - Decoder	[FCN(512),FCN(1024),FCN(2048),FCN(512),FCN(10)]

Table I gives the details of the architectures used. We denote the convolutional layer as Conv(F), where F represents the number of filters. We do the max-pooling operation on non-overlapping patches. FC(H) represents a fully connected layer with H hidden neurons. We train the systems for 400 epochs using an Adam Optimizer [45] with an initial learning rate of 10^{-3}, subject to a decay of 0.5 when the validation loss stagnates.

B. Simulation Results

TABLE II: Classification Accuracy for $N = 4$ over GMAC

Method	K	CSNR=0dB	CSNR=10dB	CSNR=20dB
DiFJ-AU	20	80.86%	81.41%	81.48%
DiFJ-SL	20	82.55%	83.58%	83.72%
DiFJ-VIB	20	**82.78%**	**84.12%**	**84.43%**
DiFJ-AU	5	76.29%	78.09%	78.09%
DiFJ-SL	5	**79.29%**	82.25%	82.58%
DiFJ-VIB	5	79.02%	**82.71%**	**83.07%**

Gaussian MAC: Table II presents the classification accuracies for the simulations over a Gaussian MAC. Here, the DiFJ-AU method serves as a baseline. We observe that both DiFJ-SL and DiFJ-VIB outperform the results of DiFJ-AU. We also observe that DiFJ-VIB and DiFJ-SL are close to each other at low CSNR, but DiFJ-VIB performs better at higher CSNR. The increasing gap is probably because of the MI gap $I(\boldsymbol{X}; \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N) - I(\boldsymbol{X}; \hat{\boldsymbol{Y}})$. A more in-depth study is required to verify this hypothesis.

To verify that all encoders transmit meaningful information, we performed two analyses. First, across all experiments over GMAC, we observed that the average power of each individual encoder is in the range $[0.8P_T, 1.2P_T]$ where P_T is the global power constraint. Second, dropping any one of the encoders for $K = 5$ and CSNR=0dB (trained with DiFJ-VIB) results in a 5% drop in, accuracy.

TABLE III: Classification Accuracy for $N = 4$ over orthogonal AWGN Channels

Method	C=4	C=8	C=12	C=16	C=20
NN-REG [39]	48.63%	63.32%	68.07%	73.43%	78.12%
NN-GBI [39]	48.33%	60.88%	65.16%	71.57%	81.18%
DiFJ-AU	46.66%	61.89%	69.55%	71.4%3	73.74%
DiFJ-SL	**61.76%**	**74.52%**	**79.62%**	**80.97%**	81.79%
DiFJ-VIB	61.19%	**74.46%**	**79.66%**	**81.03%**	**82.23%**

Orthogonal AWGN Channels: Table III showcases the results for the experiments performed on orthogonal AWGN Channels. The total channel capacity in bits is computed as

$$C = \frac{NK}{2} \log_2^+ \left(1 + \frac{P_T}{\sigma_z^2}\right). \tag{22}$$

We vary the values of P_T and vary K to get the different capacities. We choose the values of capacity and $N = 4$ to match that of [39]. Since [39] does not model the communication channel, their performance serves as an upper bound on the performance of a digital communication system. We note that DiFJ-AU is the only system whose performance is below this upper bound. Both DiFJ-SL and DiFJ-VIB perform better than [39]. Similar to our observations in the low CSNR regime of Gaussian MAC, the performance gap between the DiFJ-SL and DiFJ-VIB remains small.

C. Robustness

TABLE IV: Robustness of DiFJ-VIB over GMAC with systems trained at a particular CSNR (CSNR_{Tr}) and tested over a range of CSNRs (CSNR_{Te})

K	CSNR_{Tr}	CSNR_{Te}=0dB	CSNR_{Te}=10dB	CSNR_{Te}=20dB
20	0dB	82.78%	83.62%	83.76%
20	10dB	82.78%	84.12%	84.17%
20	20dB	82.88%	84.12%	84.43%
5	0dB	79.02%	81.66%	81.84%
5	10dB	78.67%	82.71%	82.53%
5	20dB	77.96%	82.48%	83.07%

Due to space constraints, we only present the robustness results for DiFJ-VIB systems operating over the Gaussian MAC in Table IV. All learned systems are robust to deviations in noise power. We observe that even when the training CSNR varies from the CSNR at the testing time by 20dB, the accuracy drops by only around 1% w.r.t. the system whose CSNR_{Te}=CSNR_{Tr}. Further, the robustness increases with K.

V. Conclusion

In this paper, we studied analog joint source-channel coding for distributed functional computation using deep neural networks. We proposed and studied three methods, DiFJ-AU, DiFJ-SL, and DiFJ-VIB. We first framed the problem like an autoencoder by minimizing only the distortion and enforcing the power constraint by scaling the individual encodings to have a norm equal to the power constraint. We incorporated the power constraint into the objective by using a Lagrange multiplier in the second method. Finally, we used the Variational Information Bottleneck to derive the objective for training. We showed that all objectives are variational approximations to upper bounds on the minimization objective from the indirect rate-distortion problem. Further, the objective of DiFJ-VIB is a variational approximation of a tighter upper bound when compared to DiFJ-AU and DiFJ-SL. Finally, we showed empirical performance results over the CIFAR-10 dataset. For both Gaussian MAC and AWGN channels, we found that the system trained using the objective from DiFJ-VIB performed better than or close to the performance of DiFJ-SL which, in turn was better than the system DiFJ-AU. Further, our systems performed better than the upper bounds obtained from other existing methods for the case of orthogonal AWGN channels.

REFERENCES

[1] S. Lucero *et al.*, "Iot platforms: enabling the internet of things," *White paper*, 2016.

[2] J. Nie, J. Zhang, W. Ding, X. Wan, X. Chen, and B. Ran, "Decentralized cooperative lane-changing decision-making for connected autonomous vehicles," *IEEE Access*, vol. 4, pp. 9413–9420, 2016.

[3] E. Bourtsoulatze, D. B. Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," *IEEE Transactions on Cognitive Communications and Networking*, vol. 5, no. 3, pp. 567–579, 2019.

[4] Y. M. Saidutta, A. Abdi, and F. Fekri, "Joint source-channel coding of gaussian sources over awgn channels via manifold variational autoencoders," in *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2019, pp. 514–520.

[5] ——, "M to 1 joint source-channel coding of gaussian sources via dichotomy of the input space based on deep learning," in *2019 Data Compression Conference (DCC)*, 2019, pp. 488–497.

[6] M. Rao, N. Farsad, and A. Goldsmith, "Variable length joint source-channel coding of text using deep neural networks," in *2018 IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, 2018, pp. 1–5.

[7] Y. M. Saidutta, A. Abdi, and F. Fekri, "Joint source-channel coding for gaussian sources over awgn channels using variational autoencoders," in *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2019, pp. 1327–1331.

[8] K. Choi, K. Tatwawadi, A. Grover, T. Weissman, and S. Ermon, "Neural joint source-channel coding," in *International Conference on Machine Learning*. PMLR, 2019, pp. 1182–1192.

[9] Y. M. Saidutta, A. Abdi, and F. Fekri, "Joint source-channel coding over additive noise analog channels using mixture of variational autoencoders," *IEEE Journal on Selected Areas in Communications*, p. to appear in July, 2021.

[10] ——, "Vae for joint source-channel coding of distributed gaussian sources over awgn mac," in *2020 IEEE 21st International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, 2020, pp. 1–5.

[11] M. Jankowski, D. Gündüz, and K. Mikolajczyk, "Deep joint source-channel coding for wireless image retrieval," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 5070–5074.

[12] ——, "Wireless image retrieval at the edge," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 1, pp. 89–100, 2020.

[13] ——, "Joint device-edge inference over wireless links with pruning," in *2020 IEEE 21st International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, 2020, pp. 1–5.

[14] 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.

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

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

[17] 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.

[18] X. He, X. Zhou, P. Komulainen, M. Juntti, and T. Matsumoto, "A lower bound analysis of hamming distortion for a binary ceo problem with joint source-channel coding," *IEEE Transactions on Communications*, vol. 64, no. 1, pp. 343–353, 2015.

[19] Y. Uğur, I. E. Aguerri, and A. Zaidi, "Vector gaussian ceo problem under logarithmic loss and applications," *IEEE Transactions on Information Theory*, vol. 66, no. 7, pp. 4183–4202, 2020.

[20] V. Doshi, D. Shah, M. Medard, and S. Jaggi, "Distributed functional compression through graph coloring," in *2007 Data Compression Conference (DCC'07)*, 2007, pp. 93–102.

[21] V. Doshi, D. Shah, M. Médard, and M. Effros, "Functional compression through graph coloring," *IEEE Transactions on Information Theory*, vol. 56, no. 8, pp. 3901–3917, 2010.

[22] S. Feizi and M. Médard, "On network functional compression," *IEEE transactions on information theory*, vol. 60, no. 9, pp. 5387–5401, 2014.

[23] D. Krithivasan and S. S. Pradhan, "Lattices for distributed source coding: Jointly gaussian sources and reconstruction of a linear function," *IEEE Transactions on Information Theory*, vol. 55, no. 12, pp. 5628–5651, 2009.

[24] V. Lalitha, N. Prakash, K. Vinodh, P. V. Kumar, and S. S. Pradhan, "Linear coding schemes for the distributed computation of subspaces," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 4, pp. 678–690, 2013.

[25] A. B. Wagner, "On distributed compression of linear functions," *IEEE Transactions on Information Theory*, vol. 57, no. 1, pp. 79–94, 2010.

[26] B. Nazer and M. Gastpar, "Computation over multiple-access channels," *IEEE Transactions on information theory*, vol. 53, no. 10, pp. 3498–3516, 2007.

[27] A. Kortke, M. Goldenbaum, and S. Stańczak, "Analog computation over the wireless channel: A proof of concept," in *SENSORS, 2014 IEEE*. IEEE, 2014, pp. 1224–1227.

[28] M. Goldenbaum, H. Boche, and S. Stańczak, "Harnessing interference for analog function computation in wireless sensor networks," *IEEE Transactions on Signal Processing*, vol. 61, no. 20, pp. 4893–4906, 2013.

[29] M. Goldenbaum and S. Stanczak, "Robust analog function computation via wireless multiple-access channels," *IEEE Transactions on Communications*, vol. 61, no. 9, pp. 3863–3877, 2013.

[30] M. Goldenbaum, H. Boche, and S. Stańczak, "Analog computation via wireless multiple-access channels: Universality and robustness," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 2921–2924.

[31] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *arXiv preprint physics/0004057*, 2000.

[32] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," *arXiv preprint arXiv:1612.00410*, 2016.

[33] A. Kolchinsky, B. D. Tracey, and D. H. Wolpert, "Nonlinear information bottleneck," *Entropy*, vol. 21, no. 12, p. 1181, 2019.

[34] I. E. Aguerri and A. Zaidi, "Distributed information bottleneck method for discrete and gaussian sources," *arXiv preprint arXiv:1709.09082*, 2017.

[35] A. Zaidi and I. E. Aguerri, "Distributed deep variational information bottleneck," in *2020 IEEE 21st International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, 2020, pp. 1–5.

[36] I. E. Aguerri and A. Zaidi, "Distributed variational representation learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 1, pp. 120–138, 2019.

[37] H. Zhang, Z. Hu, J. Wei, P. Xie, G. Kim, Q. Ho, and E. Xing, "Poseidon: A system architecture for efficient gpu-based deep learning on multiple machines," *arXiv preprint arXiv:1512.06216*, 2015.

[38] T. Ben-Nun and T. Hoefler, "Demystifying parallel and distributed deep learning: An in-depth concurrency analysis," *ACM Computing Surveys (CSUR)*, vol. 52, no. 4, pp. 1–43, 2019.

[39] O. A. Hanna, Y. H. Ezzeldin, T. Sadjadpour, C. Fragouli, and S. Diggavi, "On distributed quantization for classification," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 237–249, 2020.

[40] T. M. Cover, *Elements of information theory*. John Wiley & Sons, 1999.

[41] R. Dobrushin and B. Tsybakov, "Information transmission with additional noise," *IRE Transactions on Information Theory*, vol. 8, no. 5, pp. 293–304, 1962.

[42] H. Witsenhausen, "Indirect rate distortion problems," *IEEE Transactions on Information Theory*, vol. 26, no. 5, pp. 518–521, 1980.

[43] K. Eswaran and M. Gastpar, "Remote source coding under gaussian noise: Dueling roles of power and entropy power," *IEEE Transactions on Information Theory*, vol. 65, no. 7, pp. 4486–4498, 2019.

[44] A. Krizhevsky, "Learning multiple layers of features from tiny images," 2009.

[45] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.