Task-aware Network Coding over Butterfly Network

Jiangnan Cheng*, Sandeep Chinchali†, Ao Tang*
*School of Electrical and Computer Engineering, Cornell University, Ithaca, NY, USA
†Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX, USA
{jc3377, atang}@cornell.edu, sandeepc@utexas.edu

2023 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-7554-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISIT54713.2023.10206825

***Abstract*—Network coding allows distributed information sources such as sensors to efficiently compress and transmit data to distributed receivers across a bandwidth-limited network. Classical network coding is largely task-agnostic – the coding schemes mainly aim to faithfully reconstruct data at the receivers, regardless of what ultimate task the received data is used for. In this paper, we analyze a new *task-driven* network coding problem, where distributed receivers pass transmitted data through machine learning (ML) tasks, which provides an opportunity to improve efficiency by transmitting salient task-relevant data representations. Specifically, we formulate a *task-aware* network coding problem over a butterfly network in real-coordinate space, where lossy analog compression through principal component analysis (PCA) can be applied. A lower bound for the total loss function for the formulated problem is given, and necessary and sufficient conditions for achieving this lower bound are also provided. We introduce ML algorithms to solve the problem in the general case, and our evaluation demonstrates the effectiveness of task-aware network coding on distributed image classification tasks on MNIST, CIFAR10, CIFAR100, and satellite imagery datasets.**

I. Introduction

Distributed sensors measure rich sensory data which potentially are consumed by multiple distributed data receivers. On the other hand, network bandwidths remain limited and expensive, especially for wireless networks. For example, low Earth orbit satellites collect high-resolution Earth imagery, whose size goes up to a few terabytes per day and is sent to geographically distributed ground stations, while in the best case one ground station can only download 80 GB from one satellite in a single pass [1]. Therefore, one is motivated to make efficient use of existing network bandwidth for *distributed* data sources and receivers.

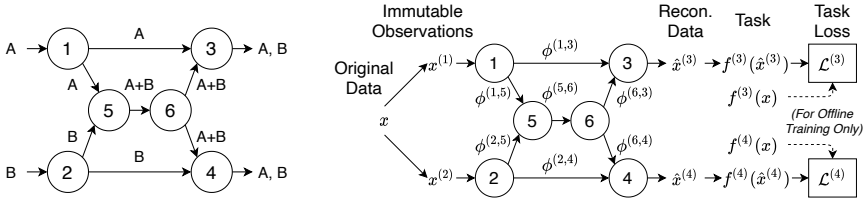
Network coding [2] is an important technology which aims at maximizing the network throughput for multi-source multicasting with limited network bandwidth. Classical network coding literature [3]–[8] considers a pure network information flow problem from the information-theoretic view, where the demands for all the data receivers, either homogeneous or heterogeneous, are specified and the objective is to satisfy each demand with a rate (i.e., mutual information between the demand and the received data) as high as possible. However, in reality, each data receiver may apply the received data to a different task, such as inference, perception and control, where different lossy data representations, even with the same rate, can produce totally different task losses. Hence it is crucial to transmit *task-relevant* data representations to distributed receivers that satisfy the network topology and bandwidth constraints, rather than representations with the highest rate.

We formulate a concrete task-aware network coding problem in this paper – *task-aware* linear network coding over a butterfly network, as shown in Fig. 1(b). The butterfly network is a representative topology in many existing network coding works [9]–[11], and hence a good example to demonstrate the benefit of making network coding task-aware. Moreover, the domain of our problem is the multi-dimensional real-coordinate space $\mathbb{R}^n$ rather than finite field $\mathrm{GF}(\cdot)$ as in classical network coding literature, which enables us to consider lossy analog compression (similar to [12]) through principal component analysis (PCA) [13] rather than information-theoretic discrete compression.

Related work. Our work is broadly related to network coding and task-aware representation learning. First, beyond the classical network coding literatures the two closest works to ours are [14] and [15], where a *data-driven* approach is adopted in the general network coding and distributed source coding settings respectively, to determine a coding scheme that minimizes *task-agnostic* reconstruction loss. In stark contrast, we aim at finding a linear network coding scheme that minimizes an overall *task-aware* loss which incorporates heterogeneous task objectives of different receivers, and we show that in some cases such linear coding schemes can even be determined analytically. Second, our work is also related to network functional compression [16]–[22], where a general function with distributed inputs over finite space is compressed. There's a similar task-aware loss function in our work, yet it corresponds to machine learning tasks over multi-dimensional real-coordinate spaces. Lastly, there have been a variety of works [23]–[27] focusing on task-aware data compression for inference, perception and control tasks under a single-source single-destination setting which is similar to Shannon's rate-distortion theory [28], while in contrast we consider task-aware data compression in a *distributed* setting.

Contributions. In light of prior work, our contributions are three-fold. First, we formulate a task-aware network coding problem over a butterfly network in real-coordinate space where lossy analog compression through PCA can be applied (Sec. III). Second, we give a lower bound for the formulated problem, and provide necessary and sufficient conditions for achieving such a lower bound (Sec. IV). Third, we adopt standard gradient descent algorithms to solve the formulated

This material is based upon work supported by the National Science Foundation under Grant No. 2133481 and 2133403. A full version of this paper is accessible at: https://arxiv.org/abs/2201.11917.pdf.



(a) (b)

Fig. 1. Network coding over butterfly network. Left (classical setting): Task-agnostic network coding in finite field. Node 3 can decode B through A+(A+B) where '+' represents exclusive or logic. Right (our setting): Task-aware network coding in real-coordinate space. Salient task-relevant data representations are transmitted to make efficient use of network bandwidths.

problem in the general case, and validate the effectiveness of task-aware network coding in our evaluation (Sec. V - VI).

II. Preliminaries

A. Network Coding with a Classical Example

Network coding [2] is a technique to increase the network throughput for multi-source multicasting under limited network bandwidths. A classical example over butterfly network in finite field GF(2), as shown in Fig. 1(a), is widely used to illustrate the benefit of network coding. The butterfly network can be represented by a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, 2, \cdots, 6\}$ and $\mathcal{E} = \{(1,3), (1,5), (2,4), (2,5), (5,6), (6,3), (6,4)\}$ are the set of nodes and edges, respectively. Suppose each edge in $\mathcal{E}$ can only carry a single bit, and node 1 and 2 each have a single bit of information, denoted by A and B respectively, which are supposed to be multicast to both node 3 and 4. In this case, network coding makes such multicasting possible by encoding A and B as A+B at node 5, while routing cannot.

B. Task-aware PCA

PCA is a widely-used dimensionality-reduction technique for determining the optimal orthogonal linear transformation that compresses a random vector $x \in \mathbb{R}^n$ to a Z-dimensional representation, where $Z \leq n$. For task-aware data compression under a single-source single-destination setting [24], [27], we assume mean $\mathbb{E}_x[x] = \mathbf{0}$ and covariance matrix $\Psi \triangleq \mathbb{E}_x[xx^\top] \succ \mathbf{0}$ (i.e., $\text{rank}(\Psi) = n$), and consider the following problem

$$\min_{D,E} \quad \mathcal{L} = \mathbb{E}_x[\|f(x) - f(\hat{x})\|_2^2] \tag{1}$$

$$\text{s.t.} \quad \hat{x} = DEx, D \in \mathbb{R}^{n \times Z}, E \in \mathbb{R}^{Z \times n} \tag{2}$$

where $\hat{x}$ is the reconstructed vector through a bottlenecked channel which only transmits a low-dimensional vector in $\mathbb{R}^Z$, and $E \in \mathbb{R}^{Z \times n}$ and $D \in \mathbb{R}^{n \times Z}$ are the corresponding encoding and decoding matrices respectively. Loss function $\mathcal{L}$ is associated with a task function $f(\cdot) \in \mathbb{R}^m$ and captures the mean-squared error between $f(x)$ and $f(\hat{x})$. In this paper we consider linear task function $f(x) = Kx$, where $K \in \mathbb{R}^{m \times n}$ is called task matrix. According to PCA, the optimal task loss $\mathcal{L}^*$ can be determined as follows. Suppose the Cholesky decomposition of Ψ is $\Psi = LL^\top$ where $L \in \mathbb{R}^{n \times n}$ is a lower triangular matrix with positive diagonal entries, and the eigen-values in descending order and the corresponding normalized eigen-vectors of Gram matrix $S = L^\top K^\top K L$ are $\mu_1, \mu_2, \cdots, \mu_n$ and $u_1, u_2, \cdots, u_n$ respectively. Then we have $\mathcal{L}^* = \sum_{i=Z+1}^{n} \mu_i$, and if the eigen-gap $\mu_Z - \mu_{Z+1} > 0$ (define $\mu_{n+1} = 0$), we must have $\text{col}(E^\top) = \text{span}(\{L^{-\top}u_1, L^{-\top}u_2, \cdots, L^{-\top}u_Z\})$ to achieve minimum task loss, where $\text{col}(\cdot)$ denotes the column space of a matrix and $\text{span}(\cdot)$ denotes the linear span of a set of vectors.

III. Problem Formulation

We now formulate a task-aware network coding problem over a butterfly network, as shown in Fig. 1(b). The key differences between our formulation and the classical example in Fig. 1(a) are: 1) our formulation has a heterogeneous task objective for each receiver while the classical example does not; 2) the domain of our code is multi-dimensional real-coordinate space rather than a finite space as in the classical example, and hence PCA can be applied.

Data. The original data is a random vector $x = [x_1, x_2, \cdots, x_n]^\top \in \mathbb{R}^n$, where $x_i \in \mathbb{R}$ is a random variable, $\forall i \in \{1, 2, \cdots, n\}$. Without loss of generality, we assume $\mathbb{E}_x[x] = \mathbf{0}$, or else we replace x by $x - \mathbb{E}_x[x]$. We also let $\Psi = \mathbb{E}_x[xx^\top]$ be the covariance matrix of x.

Data observations. Node 1 and 2 have immutable partial observations of x, denoted by $x^{(1)} \in \mathbb{R}^a$ and $x^{(2)} \in \mathbb{R}^b$, respectively. Here observations $x^{(1)}$ and $x^{(2)}$ are composed of a and b different dimensions of x, respectively; and each x_i exists in *at least* one of the two observations. Therefore, we have $\max\{a, b\} \leq n \leq a + b$. Without loss of generality, we let $x^{(1)} = [x_1, x_2, \cdots, x_a]^\top$ and $x^{(2)} = [x_{n-b+1}, x_{n-b+2}, \cdots, x_n]^\top$. That is, $x_{1:n-b}$ and $x_{a+1:n}$ are node 1's and node 2's exclusive observations respectively, and $x_{n-b+1:a}$ are their mutual observations.

Data transmission. We assume all the edges have the same capacity Z, which represents the number of dimensions in real-coordinate space here. And $\forall (i,j) \in \mathcal{E}$, we use $\phi^{(i,j)} \in \mathbb{R}^Z$ to denote the random vector that transmits over the edge (i,j). Notice that for each edge $(i,j) \in \mathcal{E}' = \{(1,3), (1,5), (2,4), (2,5), (5,6)\}$, the overall number of input dimensions for node i can be larger than Z, so we use linear mappings to transform the input signal to a low-dimensional signal in $\mathbb{R}^Z$:

$$\phi^{(1,3)} = E^{(1,3)}x^{(1)}, \phi^{(1,5)} = E^{(1,5)}x^{(1)}, \tag{3}$$

$$\phi^{(2,4)} = E^{(2,4)}x^{(2)}, \phi^{(2,5)} = E^{(2,5)}x^{(2)}, \tag{4}$$

$$\phi^{(5,6)} = E^{(5,6)} \begin{bmatrix} \phi^{(1,5)} \\ \phi^{(2,5)} \end{bmatrix}, \tag{5}$$

where $E^{(1,3)}, E^{(1,5)} \in \mathbb{R}^{Z\times a}$, $E^{(2,4)}, E^{(2,5)} \in \mathbb{R}^{Z\times b}$, and $E^{(5,6)} \in \mathbb{R}^{Z\times 2Z}$ are encoding matrices. Node 6 simply multicasts the data received from node 5 to node 3 and 4, i.e., $\phi^{(6,3)} = \phi^{(6,4)} = \phi^{(5,6)}$.

Data reconstructions. Node 3 and 4 aim to reconstruct the original data x, through the aggregated inputs they received from their input edges. The decoder functions are:

$$\hat{x}^{(3)} = D^{(3)} \begin{bmatrix} \phi^{(1,3)} \\ \phi^{(6,3)} \end{bmatrix}, \hat{x}^{(4)} = D^{(4)} \begin{bmatrix} \phi^{(2,4)} \\ \phi^{(6,4)} \end{bmatrix}, \tag{6}$$

where $D^{(3)}, D^{(4)} \in \mathbb{R}^{n\times 2Z}$ are decoding matrices for node 3 and node 4 respectively, and $\hat{x}^{(3)}$ and $\hat{x}^{(4)}$ are the reconstructed data at node 3 and node 4 respectively.

Task objectives. Node i ($\forall i \in \{3,4\}$) uses the reconstructed data $\hat{x}^{(i)}$ as the input for a task with loss function

$$\mathcal{L}^{(i)} = \mathbb{E}_x[\|f^{(i)}(x) - f^{(i)}(\hat{x}^{(i)})\|_2^2], \quad \forall i \in \{3,4\} \tag{7}$$

where $f^{(i)}(x) = K^{(i)}x$ with task matrix $K^{(i)} \in \mathbb{R}^{m_i\times n}$.

Task-aware network coding problem. The problem can be written as an optimization problem:

$$\min_{E^{(i,j)}, D^{(i)}} \quad \mathcal{L}_{\text{total}}, \quad \text{s.t.} \quad \text{Eq.}(3) - (6) \tag{8}$$

where we find the optimal encoder and decoder parameters to minimize the overall task loss $\mathcal{L}_{\text{total}} \triangleq \mathcal{L}^{(3)} + \mathcal{L}^{(4)}$.

IV. Analysis

In this section, we summarize the main results towards the task-aware network coding problem. We first provide a lower bound $\mathcal{L}_{\text{total,lb}}$ which may not be always achievable, and then discuss necessary condition and sufficient conditions for $\mathcal{L}^*_{\text{total}} = \mathcal{L}_{\text{total,lb}}$.

A. Lower bound $\mathcal{L}_{total,lb}$

We start the analysis by making the assumption of $\text{rank}(\Psi) = n$, which doesn't make the task-aware network coding problem lose generality. Next, we let the Cholesky decomposition of Ψ be $LL^\top$, where $L \in \mathbb{R}^{n\times n}$. Moreover, notice that $\phi^{(i,j)}$ is a linear transformation from x and hence is also a linear transformation from $L^{-1}x$. Therefore, for the convenience of the following analysis we let $\phi^{(i,j)} = \Phi^{(i,j)\top}L^{-1}x$ where $\Phi^{(i,j)} \in \mathbb{R}^{n\times Z}$ is a transformation matrix. Furthermore, we assume $Z \le n$, or else the network bandwidth is enough to make $\mathcal{L}^*_{\text{total}} = 0$.

For task matrix $K^{(i)}$, $\forall i \in \{3,4\}$, we define Gram matrix $S^{(i)} = L^\top K^{(i)\top} K^{(i)} L \in \mathbb{R}^{n\times n}$. Moreover, let the eigen-values in descending order and the corresponding normalized eigen-vectors of $S^{(i)}$ be $\mu_1^{(i)}, \mu_2^{(i)}, \cdots, \mu_n^{(i)}$ and $u_1^{(i)}, u_2^{(i)}, \cdots, u_n^{(i)}$, respectively. Since node 3 and 4 both receive $2Z$ dimensions, according to PCA, we have $\mathcal{L}^{(3)} \ge \sum_{j=2Z+1}^n \mu_j^{(3)}$ and $\mathcal{L}^{(4)} \ge \sum_{j=2Z+1}^n \mu_j^{(4)}$. Therefore, $\mathcal{L}_{\text{total}} \ge \mathcal{L}_{\text{total,lb}} \triangleq \sum_{i\in\{3,4\}} \sum_{j=2Z+1}^n \mu_j^{(i)}$.

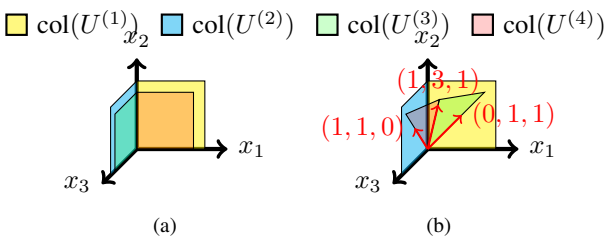


Fig. 2. Two illustrative examples for Theorem 1, where the left one doesn't achieve $\mathcal{L}_{\text{total,lb}}$ while the right one does.

Clearly, the lower bound $\mathcal{L}_{\text{total,lb}}$ may not be always achievable. Hence in the next two subsections, we focus on exploring the necessary conditions and sufficient conditions for $\mathcal{L}^*_{\text{total}} = \mathcal{L}_{\text{total,lb}}$. For further analysis, $\forall i \in \{3,4\}$, we define

$$U^{(i)} = [u_1^{(i)}, u_2^{(i)}, \cdots, u_{\min\{2Z,n\}}^{(i)}] \in \mathbb{R}^{n\times \min\{2Z,n\}}, \tag{9}$$

where the column vectors of $U^{(i)}$ are the top-$\min\{2Z,n\}$ normalized eigen-vectors of $S^{(i)}$. Making $\text{col}(U^{(3)}) \subseteq \text{col}([\Phi^{(1,3)}, \Phi^{(5,6)}])$ and $\text{col}(U^{(4)}) \subseteq \text{col}([\Phi^{(2,4)}, \Phi^{(5,6)}])$ is one way to achieve $\mathcal{L}_{\text{total,lb}}$. Moreover, we let $U^{(1)} \in \mathbb{R}^{n\times a}$ and $U^{(2)} \in \mathbb{R}^{n\times b}$ be matrices whose column vectors are the first a and last b column vectors of matrix L, respectively. The network topology constrains $\text{col}(\Phi^{(1,3)}), \text{col}(\Phi^{(1,5)}) \subseteq \text{col}(U^{(1)})$ and $\text{col}(\Phi^{(2,4)}), \text{col}(\Phi^{(2,5)}) \subseteq \text{col}(U^{(2)})$. Therefore, we say $\Phi^{(1,3)}$ is *valid* if $\text{col}(\Phi^{(1,3)}) \subseteq \text{col}(U^{(1)})$, and $\Phi^{(2,4)}$ is *valid* if $\text{col}(\Phi^{(2,4)}) \subseteq \text{col}(U^{(2)})$. On the other hand, any $\Phi^{(5,6)}$ is valid, since $\forall \Phi^{(5,6)} \in \mathbb{R}^{n\times Z}$, $\exists \Phi^{(1,5)}, \Phi^{(2,5)}$ and $E^{(5,6)}$ s.t. $\Phi^{(5,6)\top} = E^{(5,6)}[\Phi^{(1,5)}, \Phi^{(2,5)}]^\top$, $\text{col}(\Phi^{(1,5)}) \subseteq \text{col}(U^{(1)})$, and $\text{col}(\Phi^{(2,5)}) \subseteq \text{col}(U^{(2)})$.

Furthermore, we also let $r_+^{(i,j)} = \dim(\text{col}([U^{(i)}, U^{(j)}]))$ and $r_-^{(i,j)} = \dim(\text{col}(U^{(i)}) \cap \text{col}(U^{(j)}))$, $\forall i,j \in \{1,2,3,4\}$, where $\dim(\cdot)$ is the dimension of a vector space.

B. Necessary condition

The following theorem provides a necessary condition for achieving $\mathcal{L}_{\text{total,lb}}$ under a mild assumption. It constrains vector spaces' dimensions from the network bandwidth perspective.

Theorem 1. *Assume the eigen-gap* $\mu^{(i)}_{\min\{2Z,n\}} - \mu^{(i)}_{\min\{2Z,n\}+1} > 0$ *(define* $\mu^{(i)}_{n+1} = 0$*),* $\forall i \in \{3,4\}$. *Then* $\mathcal{L}_{total,lb}$ *is achievable only when*

$$r_+^{(3,4)} \le 3Z, \quad \text{and} \tag{10}$$

$$r_-^{(1,3)}, r_-^{(2,4)} \ge \min\{Z, n-Z\}. \tag{11}$$

To show the conditions in Theorem 1 are only necessary but not sufficient, we present two examples in Fig. 2, where the left one doesn't achieve $\mathcal{L}_{\text{total,lb}}$ while the right one does. Here we have $n = 3$, $\Psi = I$, $Z = 1$, $a = b = 2$. And we also assume eigen-gap $\mu_2^{(i)} - \mu_3^{(i)} > 0$, $\forall i \in \{3,4\}$. Therefore, to achieve $\mathcal{L}_{\text{total,lb}}$, we must have $\text{col}([\Phi^{(1,3)}, \Phi^{(5,6)}]) = \text{col}(U^{(3)})$ and $\text{col}([\Phi^{(2,4)}, \Phi^{(5,6)}]) = \text{col}(U^{(4)})$. In Fig. 2(a), we assume $u_1^{(3)} = u_1^{(4)} = [0,1,0]^\top$, $u_2^{(3)} = [0,0,1]^\top$ and $u_2^{(4)} = [1,0,0]^\top$. So we have $r_+^{(3,4)} = 3$ and $r_-^{(1,3)} = r_-^{(1,4)} = 1$. The

conditions in Theorem 1 are satisfied, but we cannot make $\text{col}([\Phi^{(1,3)}, \Phi^{(5,6)}]) = \text{col}(U^{(3)})$ and $\text{col}([\Phi^{(2,4)}, \Phi^{(5,6)}]) = \text{col}(U^{(4)})$ simultaneously, and hence $\mathcal{L}_{\text{total,lb}}$ is not achievable. In Fig. 2(b), we assume $u_1^{(3)} = u_1^{(4)} = \frac{1}{\sqrt{11}}[1,1,3]^\top$, $u_2^{(3)} = \frac{1}{\sqrt{66}}[4,-7,1]^\top$ and $u_2^{(4)} = \frac{1}{\sqrt{66}}[-7,4,1]^\top$. For $\Phi^{(1,3)} = [0,1,1]^\top$, $\Phi^{(2,4)} = [1,0,1]^\top$ and $\Phi^{(5,6)} = [1,3,1]^\top$, $\mathcal{L}_{\text{total,lb}}$ is achievable because $\text{col}([\Phi^{(1,3)}, \Phi^{(5,6)}]) = \text{col}(U^{(3)})$ and $\text{col}([\Phi^{(2,4)}, \Phi^{(5,6)}]) = \text{col}(U^{(4)})$.

C. Sufficient Conditions

We have seen that constrain the dimensions of vector spaces, as in Theorem 1, is not enough to achieve $\mathcal{L}_{\text{total,lb}}$. In the following theorem, we add a requirement of the data dependencies between different $U^{(i)}$'s on top of the necessary conditions, which guarantees the achievability of $\mathcal{L}_{\text{total,lb}}$.

Theorem 2. *If Eq. (10) and (11) hold, and*

$$\text{col}(U^{(3)}) = \text{span}((\text{col}(U^{(1)}) \cap \text{col}(U^{(3)})) \cup (\text{col}(U^{(3)}) \cap \text{col}(U^{(4)}))), \quad (12)$$

$$\text{col}(U^{(4)}) = \text{span}((\text{col}(U^{(2)}) \cap \text{col}(U^{(4)})) \cup (\text{col}(U^{(3)}) \cap \text{col}(U^{(4)}))), \quad (13)$$

then $\mathcal{L}_{total,lb}$ is achievable.

Eq. (12) (and similarly for Eq. (13)) has the following interpretation: we can find vectors in $\text{col}(U^{(1)})$ that extend a basis of $\text{col}(U^{(3)}) \cap \text{col}(U^{(4)})$ to a basis of $\text{col}(U^{(3)})$. This makes it possible for us to assign column vectors of $\Phi^{(1,3)}$ to achieve $\mathcal{L}_{\text{total,lb}}$ (which is not possible for Fig. 2(a)).

We further have the following two corollaries.

Corollary 3. *If Eq. (10), (12) and (13) hold, and*

$$\text{col}(U^{(3)}) \cap \text{col}(U^{(4)}) \subseteq \text{col}(U^{(1)}) \cap \text{col}(U^{(2)}), \quad (14)$$

then $\mathcal{L}_{total,lb}$ is achievable.

Corollary 4. *If Eq. (10), (12) and (13) hold, and*

$$n \leq Z + \min\{a, b\}, \quad (15)$$

then $\mathcal{L}_{total,lb}$ is achievable.

V. Algorithm

In the last section we have discussed the sufficient conditions for achieving $\mathcal{L}_{\text{total,lb}}$, and corresponding optimal encoder and decoder parameters can be determined analytically. In the general case when these sufficient conditions are not satisfied, we resort to standard gradient descent algorithms to determine the encoder and decoder parameters jointly. The encoders and decoders are connected as per network information flow (i.e., Eq. (3)-(6)). We initialize $E^{(i,j)}$ and $D^{(i)}$ randomly and update them for multiple epochs through back-propagation:

$$E^{(i,j)} \leftarrow E^{(i,j)} - \eta \frac{\nabla \mathcal{L}_{\text{total}}}{\nabla E^{(i,j)}}, \quad \forall (i,j) \in \mathcal{E}'; \quad (16)$$

$$D^{(i)} \leftarrow D^{(i)} - \eta \frac{\nabla \mathcal{L}_{\text{total}}}{\nabla D^{(i)}}, \quad \forall i \in \{3,4\} \quad (17)$$

where η is the learning rate.

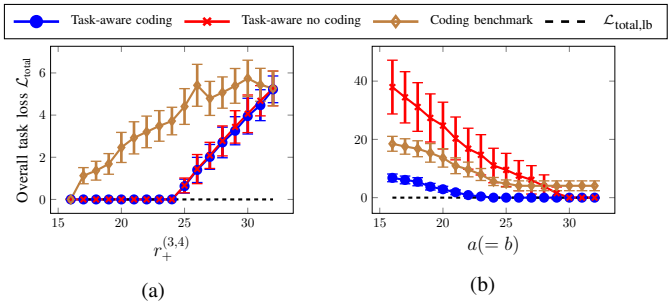


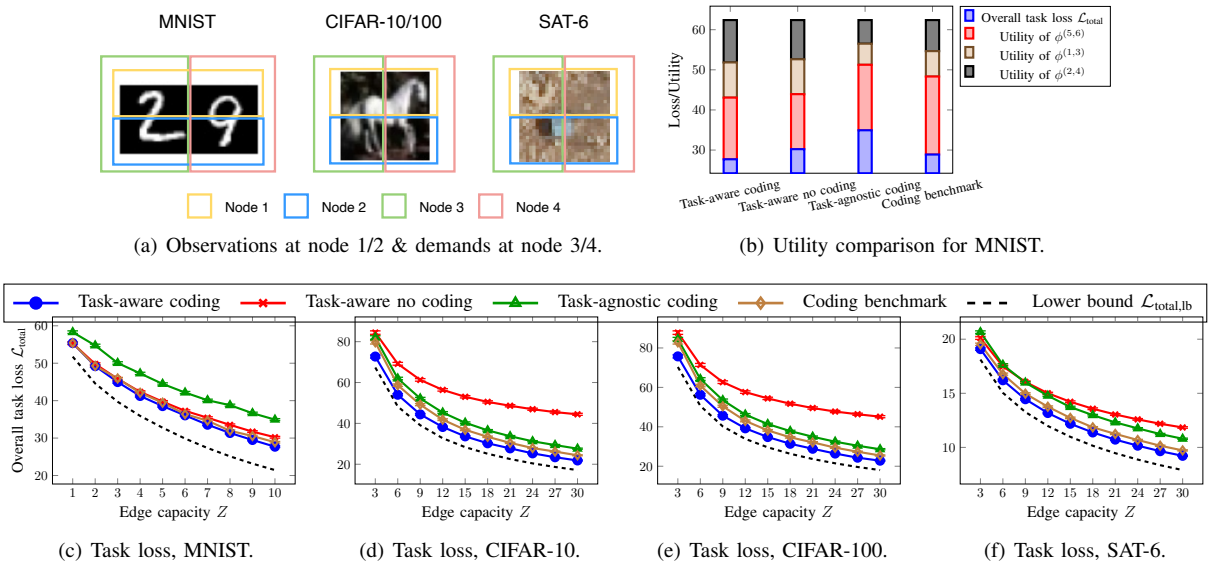
Fig. 3. Simulation result with synthetic data: overall task loss $\mathcal{L}_{\text{total}}$ under different $r_+^{(3,4)}$ (left) and different a (right). The task losses for task-agnostic coding are too large and have to be omitted from this figure.

To show our algorithm converges to a near-optimal solution for low-dimensional data and to verify our conclusions in the last section numerically, we run simulations with synthetic data for our task-aware coding approach and compare against three benchmark approaches. The benchmark approaches are: 1) **Task-aware no coding** approach, where network coding at node 5 is not allowed, i.e., each dimension of $\phi^{(5,6)}$ can only be a dimension of $\phi^{(1,5)}$ or $\phi^{(2,5)}$; 2) **Task-agnostic coding** approach (used in [14]), where the objective is to minimize the reconstruction loss at node 3 and 4, i.e., $K^{(3)} = K^{(4)} = I$; 3) **Coding benchmark** approach, which is also a task-aware coding approach but the encoder parameters associated with edge $(5,6)$ is determined greedily first and then other parameters. Such a greedy approach doesn't ensure global optimality but provides a general analytical solution.

The simulation results are shown in Fig. 3. The parameters are as follows: we fix $n = 32$, $\Psi = I$, $a = b \geq 16$, $Z = 8$. Next we let eigen-values $\mu_1^{(3)}, \cdots, \mu_{2Z}^{(3)}$ and $\mu_1^{(4)}, \cdots, \mu_{2Z}^{(4)}$ be positive, and other eigen-values of $S^{(3)}$ and $S^{(4)}$ be 0. Hence $\mathcal{L}_{\text{total,lb}} = 0$. In Fig. 3(a), we fix $a = b = 24$ and change eigen-vectors $u_1^{(3)}, \cdots, u_{2Z}^{(3)}$ and $u_1^{(4)}, \cdots, u_{2Z}^{(4)}$ to make $r_+^{(3,4)}$ different, while in the meantime keep Eq. (12), (13) and (14). We can observe our task-aware coding approach achieves $\mathcal{L}_{\text{total,lb}}$ when $r_+^{(3,4)} \leq 24$, i.e., Eq. (10) is satisfied, which verifies our conclusion in Corollary 3. We also notice that the task-aware no coding approach achieves $\mathcal{L}_{\text{total,lb}}$ when $r_+^{(3,4)} \leq 24$ as well, since coding is not required to achieve $\mathcal{L}_{\text{total,lb}}$. In Fig. 3(b), we fix $u_1^{(3)}, \cdots, u_{2Z}^{(3)}$ and $u_1^{(4)}, \cdots, u_{2Z}^{(4)}$ such that $r_+^{(3,4)} = 18$, and change a, while in the meantime keep Eq. (12) and (13). We can observe our task-aware coding approach achieves $\mathcal{L}_{\text{total,lb}}$ when $a = b \geq 24$, i.e., Eq. (15) is satisfied, which verifies our conclusion in Corollary 4. Furthermore, in both Fig. 3(a) and 3(b), our task-aware coding approach beats all the other benchmark approaches under varying $r_+^{(3,4)}$'s with respect to overall task loss $\mathcal{L}_{\text{total}}$.

VI. Evaluation

Our evaluation compares the performance of our task-aware coding approach and other benchmark approaches (as in Sec. V) over a few standard ML datasets, including MNIST [29], CIFAR-10, CIFAR-100 [30] and SAT-6 [31]. For MNIST, each



(a) Observations at node 1/2 & demands at node 3/4.

(b) Utility comparison for MNIST.

(c) Task loss, MNIST.

(d) Task loss, CIFAR-10.

(e) Task loss, CIFAR-100.

(f) Task loss, SAT-6.

Fig. 4. Evaluation setup and result with MNIST, CIFAR-10, CIFAR-100 and the SAT-6 satellite imagery dataset.

data sample is a 28×28 handwritten digit image, and we let x be a horizontally-concatenated image (28×56) of two images. Node 1 and 2 observe the upper and the lower half part of the concatenated image (both 14×56) respectively. Task matrices $K^{(3)}$ and $K^{(4)}$ are formulated as follows: we pre-train a convolutional neural network (CNN) to classify original MNIST digits by their labels. Task matrix $K^{(3)}$ requires both the reconstruction of the feature map (i.e., the output of the first layer of CNN) of the left MNIST digit in the concatenated image, and the reconstruction of the concatenated image itself:

$$K^{(3)} = [\ \underbrace{\gamma \tilde{K}^{(3)\top}}_{\text{recon. of left feature map}}\ ,\ \underbrace{(1-\gamma)I}_{\text{recon. of concatenated image}}\ ]^{\top}, \quad (18)$$

where $\tilde{K}^{(3)}$ represents the mapping between x and the feature map of the left MNIST digit, and γ is a weight coefficient. Here we use $\gamma = 0.9$. Task matrix $K^{(4)}$ is formulated similarly while the feature map of the right MNIST digit is considered instead. For CIFAR-10/CIFAR-100/SAT-6, each data sample is a 32×32 or 28×28 colored image with 3 or 4 channels and we let x represent the original image. We similarly let node 1 and 2 observe the upper and the lower half part of the image respectively, and let node 3 and node 4 require the reconstruction of the left and the right half part respectively. The setup is illustrated in Fig. 4(a).

The evaluation result is shown in Fig. 4. In Fig. 4(c)-4(f), we plot the overall task loss $\mathcal{L}_{\text{total}}$ under different edge capacity Z. In these figures, we see task-aware coding and coding benchmark approach outperform task-aware no coding and task-agnostic coding approach, and the overall task loss $\mathcal{L}_{\text{total}}$ of our task-aware coding approach is the closet to $\mathcal{L}_{\text{total, lb}}$. The maximum improvements of overall task loss $\mathcal{L}_{\text{total}}$ for task-aware coding approach are 26.1%, 26.4%, 25.3% and 17.1% respectively, compared to task-agnostic coding approach; and are 9.1%, 103.3%, 97.8% and 28.4% respectively, compared to task-aware no coding approach. We also notice that, task-agnostic coding approach doesn't always outperform task-aware no coding approach, and vice versa. Therefore, it is beneficial to combine network coding and task-awareness.

In Fig. 4(b), we compare the utilities of $\phi^{(5,6)}, \phi^{(1,3)}$ and $\phi^{(2,4)}$ in terms of minimizing the overall task loss $\mathcal{L}_{\text{total}}$ when $Z = 10$, $\gamma = 0.9$. Here we consider the utility of $\phi^{(5,6)}$ in the absence of $\phi^{(1,3)}$ and $\phi^{(2,4)}$, and the utility of $\phi^{(1,3)}$ and $\phi^{(2,4)}$ in the presence of $\phi^{(5,6)}$. We observe that the coding benchmark approach outperforms other approaches with respect to the utility of $\phi^{(5,6)}$, but underperforms our task-aware coding approach by 4.2% with respect to the overall task loss $\mathcal{L}_{\text{total}}$. This is because coding benchmark approach greedily determines the encoder parameters associated with edge $(5, 6)$ first which however could not guarantee optimality. On the other hand, our task-aware coding approach tunes all the encoding and decoding parameters jointly and achieves a lower $\mathcal{L}_{\text{total}}$.

VII. Conclusion

This paper considers task-aware network coding over a butterfly network in real-coordinate space. We prove a lower bound $\mathcal{L}_{\text{total,lb}}$ of the total loss, as well as conditions for achieving $\mathcal{L}_{\text{total,lb}}$. We also provide a machine learning algorithm in the general setting. Experimental results demonstrate that our task-aware coding approach outperforms the benchmark approaches under various settings. Regarding future extensions, although the butterfly network is a representative topology in network coding, it is worth extending the analysis of the task-aware network coding problem to general networks. A similar $\mathcal{L}_{\text{total,lb}}$ can still be derived, yet the associated necessary and sufficient conditions for achieving $\mathcal{L}_{\text{total,lb}}$ depend on the specific network topology in a manner that needs further investigation to be fully understood.

REFERENCES

[1] D. Vasisht, J. Shenoy, and R. Chandra, "L2d2: Low latency distributed downlink for leo satellites," in *Proceedings of the 2021 ACM SIGCOMM 2021 Conference*, 2021, pp. 151–164.

[2] R. Ahlswede, N. Cai, S.-Y. Li, and R. W. Yeung, "Network information flow," *IEEE Transactions on information theory*, vol. 46, no. 4, pp. 1204–1216, 2000.

[3] S.-Y. Li, R. W. Yeung, and N. Cai, "Linear network coding," *IEEE transactions on information theory*, vol. 49, no. 2, pp. 371–381, 2003.

[4] R. Koetter and M. Médard, "An algebraic approach to network coding," *IEEE/ACM transactions on networking*, vol. 11, no. 5, pp. 782–795, 2003.

[5] R. Dougherty, C. Freiling, and K. Zeger, "Insufficiency of linear coding in network information flow," *IEEE transactions on information theory*, vol. 51, no. 8, pp. 2745–2759, 2005.

[6] S. Jaggi, P. Sanders, P. A. Chou, M. Effros, S. Egner, K. Jain, and L. M. Tolhuizen, "Polynomial time algorithms for multicast network code construction," *IEEE Transactions on Information Theory*, vol. 51, no. 6, pp. 1973–1982, 2005.

[7] T. Ho, M. Médard, R. Koetter, D. R. Karger, M. Effros, J. Shi, and B. Leong, "A random linear network coding approach to multicast," *IEEE Transactions on Information Theory*, vol. 52, no. 10, pp. 4413–4430, 2006.

[8] M. Chen, M. Ponec, S. Sengupta, J. Li, and P. A. Chou, "Utility maximization in peer-to-peer systems," *ACM SIGMETRICS Performance Evaluation Review*, vol. 36, no. 1, pp. 169–180, 2008.

[9] A. S. Avestimehr and T. Ho, "Approximate capacity of the symmetric half-duplex gaussian butterfly network," in *2009 IEEE Information Theory Workshop on Networking and Information Theory*. IEEE, 2009, pp. 311–315.

[10] P. Parag and J.-F. Chamberland, "Queueing analysis of a butterfly network for comparing network coding to classical routing," *IEEE Transactions on Information Theory*, vol. 56, no. 4, pp. 1890–1908, 2010.

[11] A. Soeda, Y. Kinjo, P. S. Turner, and M. Murao, "Quantum computation over the butterfly network," *Physical Review A*, vol. 84, no. 1, p. 012333, 2011.

[12] Y. Wu and S. Verdú, "Rényi information dimension: Fundamental limits of almost lossless analog compression," *IEEE Transactions on Information Theory*, vol. 56, no. 8, pp. 3721–3748, 2010.

[13] G. H. Dunteman, *Principal components analysis*. Sage, 1989, no. 69.

[14] L. Liu, A. Solomon, S. Salamatian, and M. Médard, "Neural network coding," in *ICC 2020-2020 IEEE International Conference on Communications (ICC)*. IEEE, 2020, pp. 1–6.

[15] J. Whang, A. Acharya, H. Kim, and A. G. Dimakis, "Neural distributed source coding," *arXiv preprint arXiv:2106.02797*, 2021.

[16] 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.

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

[18] C. Shannon, "The zero error capacity of a noisy channel," *IRE Transactions on Information Theory*, vol. 2, no. 3, pp. 8–19, 1956.

[19] D. Slepian and J. Wolf, "Noiseless coding of correlated information sources," *IEEE Transactions on information Theory*, vol. 19, no. 4, pp. 471–480, 1973.

[20] R. Ahlswede and J. Korner, "Source coding with side information and a converse for degraded broadcast channels," *IEEE Transactions on Information Theory*, vol. 21, no. 6, pp. 629–637, 1975.

[21] A. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Transactions on information Theory*, vol. 22, no. 1, pp. 1–10, 1976.

[22] J. Korner and K. Marton, "How to encode the modulo-two sum of binary sources (corresp.)," *IEEE Transactions on Information Theory*, vol. 25, no. 2, pp. 219–221, 1979.

[23] Y. Blau and T. Michaeli, "Rethinking lossy compression: The rate-distortion-perception tradeoff," in *International Conference on Machine Learning*. PMLR, 2019, pp. 675–685.

[24] M. Nakanoya, S. Chinchali, A. Anemogiannis, A. Datta, S. Katti, and M. Pavone, "Co-design of communication and machine inference for cloud robotics," *Robotics: Science and Systems XVII, Virtual Event*, 2021.

[25] Y. Dubois, B. Bloem-Reddy, K. Ullrich, and C. J. Maddison, "Lossy compression for lossless prediction," *arXiv preprint arXiv:2106.10800*, 2021.

[26] G. Zhang, J. Qian, J. Chen, and A. Khisti, "Universal rate-distortion-perception representations for lossy compression," *arXiv preprint arXiv:2106.10311*, 2021.

[27] J. Cheng, M. Pavone, S. Katti, S. P. Chinchali, and A. Tang, "Data sharing and compression for cooperative networked control," in *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

[28] C. E. Shannon *et al.*, "Coding theorems for a discrete source with a fidelity criterion," *IRE Nat. Conv. Rec*, vol. 4, no. 142-163, p. 1, 1959.

[29] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[30] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[31] S. Basu, S. Ganguly, S. Mukhopadhyay, R. DiBiano, M. Karki, and R. Nemani, "Deepsat: a learning framework for satellite imagery," in *Proceedings of the 23rd SIGSPATIAL international conference on advances in geographic information systems*, 2015, pp. 1–10.