Distributed Computing of Functions of Structured Sources with Helper Side Information

Derya Malak

Communication Systems Department, EURECOM, France

Abstract—In this work, we consider the problem of distributed computing of functions of structured sources, focusing on the classical setting of two correlated sources and one user that seeks the outcome of the function while benefiting from low-rate side information provided by a helper node. Focusing on the case where the sources are jointly distributed according to a very general mixture model, we here provide an achievable coding scheme that manages to substantially reduce the communication cost of distributed computing by exploiting the nature of the joint distribution of the sources, the side information, as well as by exploiting the symmetry enjoyed by the desired functions. Our scheme — which can readily apply in a variety of real-life scenarios including learning, combinatorics, and graph neural network applications — is here shown to provide substantial reductions in the communication costs, while simultaneously providing computational savings by reducing the exponential complexity of joint decoding techniques to a complexity that is merely linear.

I. INTRODUCTION

The past few years have seen a rising need to speed up computationally-intensive tasks, as well as have witnessed an ever increasing necessity for new parallel processing techniques that efficiently distribute computations across groups of servers. This necessary transition to distributed computing though, has also introduced a variety of challenges that involve accuracy [1], computing scalability [2], and straggler mitigation [3]. Key among these challenges comes in the form of the crippling communication bottleneck of distributed computing, brought about by the astronomical communication costs often required for implementing computing in a distributed manner. This bottleneck is central to our work here.

Following the seminal work in [4] on the communication complexity of distributed computing, several other works have sought to minimize the associated communication cost, with some of these works including the recent breakthroughs in distributed linearly separable computation [5], distributed matrix product computation [6], and others that followed in the footsteps of seminal works on distributed source compression for function computation. Such seminal works include [7] and [8], that consider the case of function computation over networks, as well as [8] that consider the generalization to functional rate-distortion. This same communication cost was naturally also affected by the nature of the computed function. With this in mind, we studied distributed computation using structured distributions, see e.g., [9].

The above motivate us to explore the joint effect of having structure in data and in sources, and to explore how this structure can help us reduce the aforementioned communication bottleneck of distributed computing. Toward this, we

here consider a partially distributed computation problem with two sources of jointly distributed data, with one user that seeks the outcome of a function of the sources, and with one low-rate helper that mediates the computation of the function by providing a small amount of side information on the instantaneous matching of the sources. The sources are here modeled by a mixture distribution. From the perspective of Bayesian inference, such models can accurately capture the behavior of data distributed according to a mixing distribution.

Our aim will be to exploit the structure of the data and of the desired functions, in order to reduce communication (as well as computational/decoding) costs. To do that, while capturing the structure of the partially distributed sources, we will seek to decompose the joint source distribution into a convex combination of integral matchings, and we will do so by exploiting the well-known *Birkhoff-von Neumann Theorem* in [10], as well as *Sinkhorn's Theorem* in [11]. We will here focus on the case where the nonmatched distributions correspond to low-probability events, and the case where the side information captures this matching behavior.

We next describe generalizations of the Birkhoff-von Neumann statistical multiplexing approaches that have been successful in signal processing, wireless and networking applications, including switching theory, e.g., the study of the rate region of flows to compute a switch schedule using a graph-theoretic formulation [12], and best-effort switching services [13]. Fast converging algorithms for delay-sensitive applications with sparse switching configurations by leveraging Frank-Wolfe methods were also studied [14]. In [15], the Birkhoff-von Neumann theorem was used for the analysis of gossip algorithms. There exist applications in multi-sensor data in signal processing [16], matrix analysis [17], and stochastic matrix optimization for power amplifiers [18]. For instance, in latency-constrained applications, it is possible to precompute an approximated Birkhoff-von Neumann decomposition offline and then select a permutation matrix at random with probability proportional to its coefficient [19].

All the above will allow us to propose an achievable coding scheme that captures the aforementioned structure in data and functions as well as exploits the side information, to reduce the overall communication cost. It is worth noting that the proposed scheme is different than the existing zero-error distributed coding schemes [20] whose operating rates are often limited by the network topology. The proposed scheme, as we will see, can attain gains of approximately 40% over distributed schemes that do not exploit the above structure. In addition, the helper-based approach will be shown

to dramatically decrease the decoding complexity of joint distributions versus minimum-entropy decoding (which is NP-hard in general) used in Slepian-Wolf source coding [21].

The rest of the paper is structured as follows. In Sect. II, we give a primer on graph compression for modeling the computation rate region for two sources jointly distributed according to a mixture. In Sect. III, we detail the rate region with a helper that can extract the matching information of the sources for (i) perfectly matched, and (ii) maximally matched sources. We conclude in Sect. IV with an example to numerically evaluate the complexity of our technique.

Notation. For a random variable X with a finite alphabet $\mathcal{X}$, then P_X will denote its probability mass function (PMF). Similarly, for variables X_1 and X_2, then P_{X_1, X_2} will denote the joint PMF. Let the entropy function of a PMF $\mathbf{p}$ be $h(\mathbf{p}) = -\sum_i p_i \log p_i$ where the logarithm is in base 2, $h(p)$ be the binary entropy function with parameter p, and $H(X) = \mathbb{E}[-\log P_X(X)]$ be the Shannon entropy of X drawn from P_X. We denote by $\mathbf{X}_1^n = X_{11}, X_{12}, \ldots, X_{1n} \in \mathcal{X}_1^n$ the length n sequence of X_1 sampled from an n-fold finite alphabet $\mathcal{X}_1^n$. We let $[N] = \{1, 2, \ldots, N\}$, $N \in \mathbb{Z}^+$.

II. MODEL, DEFINITIONS, & RELATED EXISTING RESULTS

We consider the problem of computing functions of two sources of partially distributed data according to a mixture distribution, and one user that aims to recover the function outcome. Our model relies on the availability of the side information that captures the matching information extracted from the mixture, which we will elaborate later on, just below, at the source sites, which we leverage via one helper possibly with a limited rate. The user can recover both sources in an error-free manner. Our main purpose is, for sources and functions with described structures, to characterize an achievable rate region for asymptotically lossless computation. Given the existence of schemes, e.g., [22] and [23], that aim to recover data or a function of data in a similar manner and have higher sum rates than our scheme, the two advantages of our model over these pertinent ones are (i) a tighter rate region, which is critical in modern applications because of symmetries, and (ii) complexity of decoding, not only alleviating the distributed computing systems from performing calculations beyond what the user seeks to compute but also eliminating the prohibitive complexity of minimum-entropy decoding.

In what follows, we describe some of the fundamental ingredients that we will use, first with respect to some basic principles behind the graph entropy approach that will allow us to capture some of the interdependencies between the sources and the computed function, and then we will discuss the mixture model that captures the joint source distribution.

A. Representation of Source Characteristic Graphs

In this section, we review the fundamental limits of asymptotically lossless compression for computation, which can be realized using the concept of *characteristic graphs*. Source one builds a characteristic graph $G_{X_1} = (V_{X_1}, E_{X_1})$ [7] for computing $f(X_1, X_2)$ to distinguish the source outcomes that

yield a different output for any value of X_2, and similarly for source two. Vertices are the sample values, i.e., $V_{X_1} = \mathcal{X}_1$, and the edges are determined as follows. Given two arbitrary vertices $(x_1^{k_1}, x_1^{k_2}) \in \mathcal{X}_1^2$ in G_{X_1} such that $k_1 \neq k_2$, if $\exists$ a vertex $x_2^l \in \mathcal{X}_2$ in G_{X_2} such that $P_{X_1, X_2}(x_1^{k_1}, x_2^l) P_{X_1, X_2}(x_1^{k_2}, x_2^l) > 0$ and the function satisfies $f(x_1^{k_1}, x_2^l) \neq f(x_1^{k_2}, x_2^l)$, then $(x_1^{k_1}, x_1^{k_2}) \in E_{X_1}$. Otherwise, $(x_1^{k_1}, x_1^{k_2}) \notin E_{X_1}$.

We let $c_{G_{X_1}}(X_1)$ be a valid coloring of G_{X_1}, where a valid coloring is such that any two vertices of G_{X_1} that share an edge are assigned distinct colors. As one would expect, the characteristic graphs built by the sources are correlated.

For simultaneous encoding of multiple instances of a source, source $m \in [2]$ similarly builds the n-th power of G_{X_m}, i.e., $G_{\mathbf{X}_m}^n$. We note that $G_{\mathbf{X}_m}^n = (V_{X_m}^n, E_{X_m}^n)$ is an OR graph such that $V_{X_m}^n = \mathcal{X}_m^n$ and if $(x_{mi}^{k_1}, x_{mi}^{k_2}) \in E_{X_m}$ for some coordinate $i \in [n]$, then $(\mathbf{x}_m^{k_1}, \mathbf{x}_m^{k_2}) \in E_{X_m}^n$.

The characteristic graph entropy of X_1 is given by [7]

$$H_{G_{X_1}}(X_1) = \lim_{n \to \infty} \min_{c_{G_{\mathbf{X}_1}^n}} \frac{1}{n} H(c_{G_{\mathbf{X}_1}^n}(\mathbf{X}_1)) \, ,$$

where the minimization is over the set of all valid colorings $c_{G_{\mathbf{X}_1}^n}(\mathbf{X}_1)$ of $G_{\mathbf{X}_1}^n$. Similarly, conditional graph entropy [24] and joint graph entropy [8] can be determined.

B. Mixture Distribution in the Discrete Parameter Domain

As we outlined above, the two sources are partially distributed based on a mixture distribution. Generally, such mixture distribution results from assuming that a random variable Y is distributed according to some parametrized distribution $P_{Y|\Theta}$ with an unknown parameter Θ with the latent distribution P_Θ [25]. In this context, the unconditional distribution P_Y results from marginalizing over P_Θ:

$$P_Y(y) = \sum_\theta P_\Theta(\theta) \, P_{Y|\Theta}(y|\theta) = \sum_{l \in [L]} q_l \, P_{Y|\Theta}(y|\theta_l) \, , \quad (1)$$

where $\Theta \sim (q_1, \ldots, q_L)$ is distributed over $(\theta_1, \ldots, \theta_L)$.

Our interest in mixture distributions stems from two main properties. The first has to do with the fact that such mixture distributions span a broad range of scenarios that are of particular interest in modern science. Mixture distributions arise in Bayesian inference where a hypothesis is updated as more information becomes available. The second aspect relates to the fact that albeit broad, this mixture structure is endowed with properties that are easier to analyze and expand. One such property relates to the renowned Birkhoff's algorithm.

Birkhoff's algorithm is a greedy algorithm that receives as input a bistochastic matrix $W = (w_{ij}) \in \mathbb{R}_{\geq 0}^{N \times N}$ (meaning that each of whose rows and columns sums to 1), and returns a Birkhoff-von Neumann decomposition of W, instantiated by the Birkhoff–von Neumann Theorem [10], where W is a sum of permutation matrices with non-negative weights [26]. The run-time complexity of Birkhoff's algorithm is $O(N^2)$. This algorithm is useful in our particular case with partially distributed sources, in providing a decomposition for the mixture distribution, which we will detail in Sect. III-A.

Birkhoff's algorithm has applications in fair random assignment, alleviates the problem of solving challenging linear systems with unstructured and indefinite coefficient matrices, and improves efficiency and scalability of parallel computing [27]. It also helps us in our goal to derive the rate region in Sect. III for distributed computing of functions.

III. PROBLEM STATEMENT AND RESULTS

We here consider a pair of partially distributed sources X_1, X_2 that accept a joint PMF $P_{X_1,X_2}(x_1,x_2) > 0$ for all (x_1, x_2). Our goal is to derive the rate region for the distributed computation of $f(X_1, X_2)$ exploiting the fact that (X_1, X_2) is distributed according to a mixture PMF (cf. 1). Such mixture assumption implies that the sources are matched with a high probability. Equivalently, we can see that for $\mathcal{B}_{X_1,X_2} = (\mathcal{X}_1, \mathcal{X}_2, E)$ being the bipartite graph representation of P_{X_1,X_2} — with vertex sets $\mathcal{X}_1$ and $\mathcal{X}_2$ representing the individual source outcomes, and edge set E whose weights capture P_{X_1,X_2} — then the vertices of $\mathcal{B}_{X_1,X_2}$ are non-matched with a vanishing probability. To derive the sought rate region, we next detail how to represent and then exploit this structural decomposition of mixture distributions.

A. Representation of Matchings

To now explore and exploit the implications of having integral matchings, let us consider X_2. Let $X_2 = \pi(X_1, \Theta)$ such that the permutation index $l \in [L]$ leads to a matching configuration described by the mixture parameter $\Theta = \theta_l$, as introduced in (1). We let $\mathbf{q} = (q_1, \ldots, q_L) \in \mathbb{R}^L_{>0}$ where $q_l = P_\Theta(\theta_l)$ such that $\sum_{l \in [L]} q_l = 1$, for an integer $L \in O(N^2)$ [10], and a set of $N \times N$ permutation matrices $\{B_l, l \in [L]\}$. Let $\pi_l : \mathcal{X}_1 \rightarrow \mathcal{X}_1$ be a one-to-one and onto function that permutes the elements of $\mathcal{X}_1$ to provide a perfect matching where every vertex is adjacent to exactly one edge in $\mathcal{B}_{X_1,X_2}$, which simply means that $X_2 = \pi_l(X_1)$. Given the perfect integral matching structure in B_l given $l \in [L]$ and the value of X_1, say $X_1 = x_1^k$, the source X_2 takes the value $\pi_l(x_1^k)$ and the function $f(X_1, X_2)$ in takes the form

$$f(x_1^k, \pi_l(x_1^k)) , \quad \forall k \in [N] , \qquad (2)$$

with probability $q_l \mathbb{P}(X_1 = x_1^k | \theta_l)$. We note that the outcome of $f(X_1, X_2)$ is completely determined by B_l and X_1.

We next state our lemma, which allows a standard form for any joint PMF with positive entries.

Lemma 1. *Provided that $P_{X_1,X_2}(x_1, x_2) > 0$ for all (x_1, x_2), the joint PMF P_{X_1,X_2} can be written as a weighted sum of generalized permutation matrices:*

$$P_{X_1,X_2}(x_1,x_2) = \sum_{l \in [L]} q_l\, P_{X_1,X_2|\Theta}(x_1, x_2|\theta_l) . \qquad (3)$$

Proof. Under our assumption that $P_{X_1,X_2}(x_1, x_2) > 0$ for all (x_1, x_2), then we first employ Sinkhorn's Theorem [11] to note that P_{X_1,X_2} can be mapped into a doubly stochastic $W = D_1 P_{X_1,X_2} D_2$ for unique diagonal matrices $D_1 = (d_{1,ij})$ and $D_2 = (d_{2,ij})$ in $\mathbb{R}^{N \times N}_{>0}$. Then we proceed to employ the famous Birkhoff–von Neumann Theorem [10] to note that $W = \sum_{l \in [L]} q_l B_l$, and P_{X_1,X_2} is a weighted sum of generalized permutation matrices, thus taking the form in (3), where $P_{X_1,X_2|\Theta}(\cdot, \cdot|\theta_l) = D_1^{-1} B_l D_2^{-1}$ is a generalized permutation matrix obtained from $\{B_l, l \in [L]\}$ where $L \in O(N^2)$, and the matrices D_1^{-1} and D_2^{-1} are such that $D_1^{-1} = (d_{1,ij}^{-1})$ and $D_2^{-1} = (d_{2,ij}^{-1})$ in $\mathbb{R}^{N \times N}_{>0}$, and the mapping $\pi_l : \mathcal{X}_1 \rightarrow \mathcal{X}_1$ from X_1 to X_2 captures $P_{X_1,X_2|\Theta}(x_1, x_2|\theta_l)$. □

B. Identifying Structures of $\mathcal{B}_{X_1,X_2}$ via Matchings

There exist techniques to exploit the combinatorial structure of $\mathcal{B}_{X_1,X_2}$ when it has multiple connected components, see e.g., the Gács-Körner-Witsenhausen common information (GKW-CI) [28]. However, if $\mathcal{B}_{X_1,X_2}$ has only one bipartition, GKW-CI cannot be extracted, i.e., GKW-CI is zero.

A helper-based scheme can still provide a low-complexity distributed coding technique even when $\mathcal{B}_{X_1,X_2}$ has one bipartition, despite being sub-optimal in terms of its operating rate. We are motivated to contemplate a matching-based helper as maximum matching in graphs can be determined using polynomial time algorithms [29], which can be exploited to facilitate the Birkhoff-von Neumann decomposition. When P_{X_1,X_2} is a mixture distribution satisfying (3) that accepts a matching-based decomposition, a helper can be used to distinguish the matched and non-matched vertices of $\mathcal{B}_{X_1,X_2}$. In the following, we devise a helper-based model to extract the matching information from $\mathcal{B}_{X_1,X_2}$, and denote this helper variable by K_{M}. For the described partial distributed setting, Theorem 1 provides the rate region for computing (2).

Theorem 1. *(A matching-based computation sum rate.) For computing $f(X_1, X_2)$, where P_{X_1,X_2} satisfies the mixture distribution in (3), with a helper that extracts the perfect matching information between the sources, there exists a low-complexity zero-error encoding and decoding of X_1 and Θ that operates at rates*

$$R_H \geq H(K_{\mathrm{M}}) = H(\Theta), \quad R_1 \geq H_{G_{X_1}}(X_1|\Theta) ,$$

yielding the following total rate to compute $f(X_1, X_2)$:

$$R_M \geq H(K_{\mathrm{M}}) + H_{G_{X_1}}(X_1|\Theta) , \qquad (4)$$

where $H_{G_{X_1}}(X_1|\Theta)$ denotes the conditional characteristic graph entropy of X_1 for computing $f(X_1, X_2)$ given Θ.

Proof. In the partially distributed setting, to be able to compute $f(\mathbf{X}_1^n, \mathbf{X}_2^n)$ accurately, the perfect matching information Θ^n where $\Theta_i \sim \mathbf{q}$ for all $i \in [n]$ should be made available to the user (or via side information), which requires an asymptotic rate $R_H \geq H(K_M) = H(\Theta)$. Knowledge of Θ^n determines the jointly typical coloring sequence pairs $(c_{G_{X_1}}(\mathbf{X}_1^n), c_{G_{X_2}}(\mathbf{X}_2^n))$ to be compressed asymptotically.

Given that the perfect matching information $\Theta \sim \mathbf{q}$ is known at the user, such that $X_2 = \pi(X_1, \Theta)$, it is sufficient for the user if only one source transmits. Let us assume source one is selected. Source one builds the characteristic graph G_{X_1} to compute $f(\mathbf{X}_1^n, \mathbf{X}_2^n) = f(\mathbf{X}_1^n, \pi(\mathbf{X}_1^n, \Theta^n))$ given Θ^n. This

requires, following the notion of the conditional graph entropy, as detailed in [24], an asymptotic rate of $H_{G_{X_1}}(X_1|\Theta)$.

The rate needed from source one to compute $f(X_1, X_2)$ is

$$R_1 \geq \sum_{l \in [L]} q_l H(f(X_1, X_2) \mid X_2 = \pi_l(X_1)) , \qquad (5)$$

enabling the partial distributed computation of $f(X_1, X_2)$ at an asymptotic rate upper bounded by $H(f(X_1, X_2), \Theta)$. $\square$

The rate region in Theorem 1 is encompassed by that of optimal distributed functional compression given in [8]. On the other hand, when P_{X_1, X_2} accepts the decomposition in (3), a helper-based model to extract the matching information has advantages over Slepian-Wolf coding. Not only it eliminates the complexity associated with joint typicality decoding [30] with exponential complexity, but also provides an almost lossless compression asymptotically, given by Theorem 1.

We next impose additional structure on the distributed sources by exploiting the maximal coupling construction.

C. Extracting Matched versus Non-Matched Vertices

A maximal coupling between a pair (X_1, X_2) maximizes $\mathbb{P}(X_1 = X_2)$ subject to the marginal PMFs $X_1 \sim \mathbf{p}_1$ and $X_2 \sim \mathbf{p}_2$. Let $C(\mathbf{p}_1, \mathbf{p}_2)$ be the set of all joint PMFs of $X_1 \sim \mathbf{p}_1$ and $X_2 \sim \mathbf{p}_2$. Elements $M = [m_{ij}] \in \mathbb{R}_{\geq 0}^{|\mathcal{X}| \times |\mathcal{X}|}$ of $C(\mathbf{p}_1, \mathbf{p}_2)$ are couplings of $\mathbf{p}_1$ and $\mathbf{p}_2$:

$$C(\mathbf{p}_1, \mathbf{p}_2) \triangleq \left\{ [m_{ij} : \sum_{j \in |\mathcal{X}|} m_{ij} = p_{1i}, \sum_{i \in |\mathcal{X}|} m_{ij} = p_{2j}] \right\} .$$

A coupling of (X_1, X_2) that maximizes $\mathbb{P}(X_1 = X_2)$ is called a maximal coupling which is formally stated next.

Lemma 2. *Maximal coupling (X_1, X_2) subject to the marginal distributions $X_1 \sim \mathbf{p}_1$ and $X_2 \sim \mathbf{p}_2$ satisfies*

$$\|\mathbf{p}_1 - \mathbf{p}_2\|_{\mathrm{TV}} = \inf \left[\mathbb{P}(X_1 \neq X_2) : C(\mathbf{p}_1, \mathbf{p}_2) \right] , \quad (6)$$

where the measure $\|\mathbf{p}_1 - \mathbf{p}_2\|_{\mathrm{TV}}$ is the total variation distance between the PMFs of X_1 and X_2.

Maximal coupling of the pair (X_1, X_2) becomes relevant when the discrepancy between $\mathbf{p}_1$ and $\mathbf{p}_2$ is bounded above, e.g., the PMFs coincide in the first two decimal points. Intuitively, this coupling can be exploited to further reduce the communication cost for distributed computing.

Recall that Theorem 1 derives a schedule-based lower bound on the sum rate given by Lemma 1, irrespective of the coupling of (X_1, X_2). Next, Theorem 2 provides a lower bound on the sum rate by assuming an additional structural correlation between X_1 and X_2 through their maximal coupling.

Theorem 2. *(A maximally-coupled computation sum rate.) Assume that the sources X_1 and X_2 are maximally coupled. Then, there exists a low-complexity zero-error encoding and decoding of $f(X_1, X_2)$ and Θ with a helper that extracts the perfect matching information and operates at rates*

$$R_H + \sum_{m \in [2]} R_m \geq H(\Theta) + \sum_{l \in [L]} q_l \big(h(\delta_l) + \delta_l^c H(T_l) \big)$$
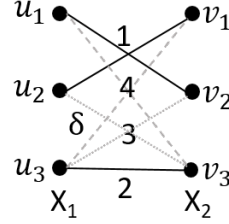

Fig. 1. Bipartite graph for Example 1 where the function outcomes are indicated on the edges, the bipartite vertex sets denote the outcomes $\mathcal{X}_1$ and $\mathcal{X}_2$, respectively, and each dotted or dashed edge has a probability δ.

$$+ \delta_l [H(V_l) + H(W_l)]) , \qquad (7)$$

where V_l, W_l, and T_l are independent integer-valued variables with distributions given according to (8) with δ is substituted with $\delta_l = \mathbb{P}(\pi_l(X_1) \neq X_2)$ for given $l \in [L]$, and $\delta_l^c = 1 - \delta_l$.

Proof. To prove this result, we provide a construction for maximal coupling from [31, Ch. 4.12]. There exists a pair X_1' and X_2' having the same marginals as X_1 and X_2 such that $\mathbb{P}(X_1' = X_2') = 1 - \|\mathbf{p}_1 - \mathbf{p}_2\|_{\mathrm{TV}} = 1 - \delta$. Let $U \sim \mathrm{Bern}(1 - \delta)$ and V, W, T be independent integer-valued variables with respective masses at $k = 1, \ldots, |\mathcal{X}|$:

$$\mathbb{P}(T = k) \triangleq \frac{\min[p_{1k}, p_{2k}]}{(1 - \delta)}, \qquad (8)$$

$$\mathbb{P}(V = k) \triangleq \frac{[p_{1k} - p_{2k}]^+}{\delta}, \quad \mathbb{P}(W = k) \triangleq \frac{[p_{2k} - p_{1k}]^+}{\delta} .$$

Then, the random variables X_1' and X_2' defined as

$$X_1' \triangleq UT + (1 - U)V, \quad X_2' \triangleq UT + (1 - U)W \quad (9)$$

have the required marginals, and are maximally coupled such that $\mathbb{P}(X_1' = X_2') = \mathbb{P}(U = 1) = 1 - \delta$. Furthermore, X_1' and X_2' are independent when $X_1' \neq X_2'$ as their sets are disjoint. For the proof, we refer the reader to [31, Ch. 4.12].

Let us consider the possible matchings described by $\Theta \sim \mathbf{q}$, where the matched vertices are invariant up to permutations. For each $l \in [L]$ with probability $q_l = P_\Theta(\theta_l)$, we have

$$\pi(X_1, \theta_l) = \pi_l(X_1) = U_l T_l + (1 - U_l) V_l ,$$
$$X_2 = U_l T_l + (1 - U_l) W_l , \quad l \in [L] , \quad (10)$$

where $U_l \sim \mathrm{Bern}(1 - \delta_l)$, and V_l, W_l, and T_l are independent integer-valued variables with respective masses at $k \in [|\mathcal{X}|]$ according to (8) where δ being replaced by δ_l.

A maximal coupling of X_1 and X_2 results in

$$\mathbb{P}(X_2 = \pi(X_1, \Theta)) = \sum_{l \in [L]} q_l \mathbb{P}(U_l = 1) = \sum_{l \in [L]} q_l \cdot \delta_l^c ,$$

where $\delta_l^c = 1 - \delta_l$. The parts corresponding to $U_l = 1$ versus $U_l = 0$ denote the matched and the non-matched components, respectively. This yields the sum rate needed from the helper and both sources to compute $f(X_1, X_2)$, as given in (7). $\square$

To contrast our model with the state-of-the-art, we next consider an example. Due to limited space, we deferred several examples to the extended version of the draft [32].

IV. Example with Structured Sources

We next study an example where a helper leverages the structure of $\mathcal{B}_{X_1,X_2}$, and provides the necessary rate to distinguish the matching information, K_M, in $\mathcal{B}_{X_1,X_2}$. Extracting this information alleviates the complexity of distributed computing provided that the non-matched distributions correspond to low-probability events. Given K_M, one source (or both) needs to send a refinement to identify the function outcome.

In the following, we denote by F the table of function outcomes, where the coordinates match the coordinates of P_{X_1,X_2} with entries ordered in an increasing fashion.

Example 1. *Consider the probability matrix and the table of function outcomes given as follows:*

$$P_{X_1,X_2} = \frac{1}{3}\begin{bmatrix} 0 & 1-\delta & \delta \\ 1-\delta & 0 & \delta \\ \delta & \delta & 1-2\delta \end{bmatrix}, \quad F = \begin{bmatrix} X & 1 & 4 \\ 1 & X & 3 \\ 4 & 3 & 2 \end{bmatrix},$$

where $\delta \in (0, 0.5)$. The bipartite graph $\mathcal{B}_{X_1,X_2}$ is shown in Fig. 1 where the vertices are listed in the presented order.

Fully distributed coding. *The entropy of the function for given P_{X_1,X_2} and F is $H(f(X_1,X_2)) = h\left(\frac{2-2\delta}{3}, \frac{1-2\delta}{3}, \frac{2\delta}{3}, \frac{2\delta}{3}\right)$. A trivial rate upper bound is $H_{G_{X_1}}(X_1) + H_{G_{X_2}}(X_2) = 2\log 3$. The sum rate required to compute the function in the case of no helper is $H_{G_{X_1}}(X_1) + H_{G_{X_2}}(X_2|X_1) = \log(3) + \frac{2}{3}h(\delta) + \frac{1}{3}h(\delta, \delta, 1-2\delta)$.*

Partially distributed coding via extracting matchings. *The helper decomposes $\mathcal{B}_{X_1,X_2}$ into a perfect matching and a non-matched graph. More specifically, P_{X_1,X_2} can be described by the following mixture distribution:*

$$P_{X_1,X_2} = q_1 \begin{bmatrix} 0 & \frac{1-\delta}{3-4\delta} & 0 \\ \frac{1-\delta}{3-4\delta} & 0 & 0 \\ 0 & 0 & \frac{1-2\delta}{3-4\delta} \end{bmatrix} + q_2 \begin{bmatrix} 0 & 0 & \frac{1}{4} \\ 0 & 0 & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & 0 \end{bmatrix},$$

where $\delta \in (0, 0.5)$, the first matrix describes a perfect matching, denoted by $K_M = 0$ with probability $q_1 = 1 - \frac{4\delta}{3}$, and the second matrix describes a low-probability event, denoted by $K_M = 1$ with $q_2 = \frac{4\delta}{3}$. Hence, the rate required from the helper to distinguish between these two is $H(K_M) = h(q_1)$. Given $K_M = 0$, only 1 source needs to transmit, which requires a rate of $h\left(\frac{2-2\delta}{3-4\delta}\right)$ to determine the function outcome. Given $K_M = 1$, source 1, X_1, needs 2 colors for u_1 and u_2 to distinguish the outcomes 4 and 3 each with a probability $\frac{1}{4}$, and u_3, which has a probability $\frac{1}{2}$, does not need to be distinguished from $\{u_1, u_2\}$ (no shared edges between $\{u_1, u_2\}$ and u_3). Similarly for X_2. Hence, given $K_M = 1$, each source is required to send at an asymptotic rate of $h\left(\frac{3}{4}\right)$ bits per use. Hence, the sum rate required to compute F with a helper that exploits the matching information of $\mathcal{B}_{X_1,X_2}$ is

$$R_M \geq H(K_M) + \mathbb{P}(K_M = 0)H_{G_{X_1}}(X_1|K_M = 0)$$
$$+ \mathbb{P}(K_M = 1)(H_{G_{X_1}}(X_1|K_M = 1) + (H_{G_{X_2}}(X_2|K_M = 1))$$
$$= h(q_1) + q_1 h\left(\frac{1-2\delta}{3-4\delta}\right) + q_2\left(h\left(\frac{3}{4}\right) + h\left(\frac{3}{4}\right)\right).$$

From above, our matching-based approach has the best approximation for small δ, where the gain of our model over the fully distributed coding setting that exploits the structure of the function but not the source is %42 and the loss versus the fundamental limit $H(f(X_1,X_2))$ is at most %26.

References

[1] J. Wang, Z. Jia, and S. A. Jafar, "Price of precision in coded distributed matrix multiplication: A dimensional analysis," in *Proc., IEEE Inf. Theory Wksh.*, Virtual Conference, Oct. 2021, pp. 1–6.

[2] M. Soleymani, H. Mahdavifar, and S. Avestimehr, "Analog lagrange coded computing," *IEEE J. Sel. Areas Inf. Theory*, vol. 2, Feb. 2021.

[3] W. Li, Z. Chen, Z. Wang, S. A. Jafar, and H. Jafarkhani, "Flexible constructions for distributed matrix multiplication," in *Proc., IEEE ISIT*, Virtual Conference, Jul. 2021, pp. 1576–1581.

[4] A. C.-C. Yao, "Some complexity questions related to distributive computing (preliminary report)," in *Proc. ACM Symp. Theory of Computing*, Atlanta, GA, Apr. 1979, pp. 209–213.

[5] A. Khalesi and P. Elia, "Multi-user linearly-separable distributed computing," *arXiv preprint arXiv:2206.11119*, Jun. 2022.

[6] Z. Jia and S. Jafar, "On the capacity of secure distributed batch matrix multiplication," *IEEE Trans. Inf. Theo.*, vol. 67, pp. 7420–37, Sep. 2021.

[7] J. Körner, "Coding of an information source having ambiguous alphabet and the entropy of graphs," in *Proc., 6th Prague Conf. Inf. Theory*, Prague, Czech Republic, Sep. 1973, pp. 411–425.

[8] S. Feizi and M. Médard, "On network functional compression," *IEEE Trans. Inf. Theory*, vol. 60, no. 9, pp. 5387–5401, Sep. 2014.

[9] D. Malak and M. Médard, "A distributed computationally aware quantizer design via hyper binning," *IEEE Trans. Signal Proces.,*, vol. 71, pp. 76–91, Jan. 2023.

[10] R. Caron, X. Li, P. Mikusiński, H. Sherwood, and M. Taylor, "Nonsquare doubly stochastic matrices," *Lecture Notes*, pp. 65–75, Jan. 1996.

[11] R. Sinkhorn, "A relationship between arbitrary positive matrices and doubly stochastic matrices," *Ann. Math. Stat.*, vol. 35, no. 2, pp. 876–879, Jun. 1964.

[12] J. Sundararajan, S. Deb, and M. Medard, "Extending the Birkhoff-von Neumann switching strategy for multicast - on the use of optical splitting in switches," *IEEE J. Sel. A. Commun.*, vol. 25, pp. 36–50, Aug. 2007.

[13] C.-S. Chang, W.-J. Chen, and H.-Y. Huang, "Birkhoff-von Neumann input buffered crossbar switches," in *Proc., IEEE Int. Conf. Comp. Commun.*, vol. 3, Mar. 2000, pp. 1614–1623.

[14] V. Valls, G. Iosifidis, and L. Tassiulas, "Birkhoff's decomposition revisited: Sparse scheduling for high-speed circuit switches," *IEEE/ACM Trans. Netw.*, vol. 29, no. 6, pp. 2399–2412, Aug. 2021.

[15] A. G. Dimakis, S. Kar, J. M. Moura, M. G. Rabbat, and A. Scaglione, "Gossip algorithms for distributed signal processing," *Proc. of the IEEE*, vol. 98, no. 11, pp. 1847–1864, 2010.

[16] B. Uçar, "Partitioning, matching, and ordering: Combinatorial scientific computing with matrices and tensors," Ph.D. dissertation, ENS de Lyon, 2019.

[17] J. A. Tropp, "Acm 204: Matrix analysis," 2022.

[18] K. Safa, M. S. Hassan, F. Jardel, and P. Sehier, "Low PAPR probabilistically controlled transitions scheme," in *Proc., IEEE Wireless Commun. and Netw. Conf. (WCNC)*, Austin, TX, Apr. 2022, pp. 2184–2189.

[19] A. Hoyos-Idrobo, "Approximate Birkhoff-von-Neumann decomposition: a differentiable approach," 2020.

[20] H. Witsenhausen, "The zero-error side information problem and chromatic numbers (corresp.)," *IEEE Trans. Inf. Theory*, vol. 22, no. 5, pp. 592–593, Sep. 1976.

[21] T. P. Coleman, M. Médard, and M. Effros, "Towards practical minimum-entropy universal decoding," in *Proc., Data Compression Conf.*, Snowbird, UT, 2005, pp. 33–42.

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

[23] J. Körner and A. Orlitsky, "Zero-error information theory," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2207–2229, Oct. 1998.

[24] A. Orlitsky and J. R. Roche, "Coding for computing," *IEEE Trans. Inf. Theory*, vol. 47, no. 3, p. 903–917, Mar. 2001.

[25] C. Röver and T. Friede, "Discrete approximation of a mixture distribution via restricted divergence," *J. Comput. Graph. Stat.*, vol. 26, no. 1, pp. 217–222, Jan. 2017.

[26] H. Aziz, "Simultaneously achieving ex-ante and ex-post fairness," in *Proc., Int. Conf. Web and Internet Economics*. Springer, Dec. 2020.

[27] M. Benzi, J. C. Haws, and M. Tuma, "Preconditioning highly indefinite and nonsymmetric matrices," *SIAM J. Sci. Comput.*, vol. 22, no. 4, pp. 1333–53, 2000.

[28] P. Gács and J. Körner, "Common information is far less than mutual information," *Prob. Control and Inf. Theory*, vol. 2, pp. 149–162, 1973.

[29] J. Edmonds, "Paths, trees, and flowers," *Canadian Journal of Mathematics*, vol. 17, pp. 449–467, 1965.

[30] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, 2012.

[31] G. R. Grimmett and D. Stirzaker, *Probability and Random Processes*. Oxford university press, 2001.

[32] D. Malak, "Distributed computing of functions of structured sources with helper side information," 2023.