Source Coding for Dependent Sources

M. Langberg
The Open University of Israel
mikel@openu.ac.il

M. Effros
California Institute of Technology
effros@caltech.edu

Abstract—In this work, we address the capacity region of multi-source multi-terminal network communication problems, and study the change in capacity when one moves form independent to dependent source information. Specifically, we ask whether the trade off between capacity and source independence is of continuous nature. We tie the question at hand to that of edge removal which has seen recent interest.

I. INTRODUCTION

The network coding paradigm has seen significant interest in the last decade (see e.g., [1], [2], [3], [4], [5] and references therein). In the multiple source network coding problem, it is common to assume that the information available at different source nodes is independent. Under this assumption, several aspects of network coding have been explored, including the study of the capacity region of the multi-source multi-terminal network coding problem, e.g. [6], [7], [8], [9], [10], [11].

In this work we focus our study on the “independence” property of source information, and ask whether it is of significant importance in the study of network coding capacities. Loosely speaking, we consider the following question. All notions that appear below will be defined rigorously in the upcoming Section II.

Question 1: Given an instance to the multi-source multi-terminal network coding problem, does the capacity region differ significantly when one removes the requirement that the source information is independent?

Clearly, if the source information is highly dependent, it is not hard to devise instances to the network coding problem in which the corresponding capacity region differs significantly from that in which source information is independent. However, what is the trade off between independence and capacity? Can it be that the trade off is not continuous?

The main contribution of this work is in a connection we make between the question at hand and the recently studied “edge removal” problem [12], [13], [14] in which one asks to quantify the loss in capacity when removing a single edge from a given network. For all but a few special cases of networks, the effect on capacity of single edge removal is not fully understood. We show that Question 1 is closely related to the edge removal problem. In particular, we show that quantifying the rate loss in the former will imply a quantification for the latter and vice-versa.

As the edge removal problem is open, our connection does not progress in answering Question 1, but rather puts it in a broader perspective. For example, as a corollary of our equivalence, we show that removing an edge of vanishing capacity (in the block length) from a network will have a vanishing effect on the capacity of the network if and only if the trade off in Question 1 is continuous (as before, rigorous definitions will be given in Section II). Using recent results of ours from [14], this implies a similar equivalence between Question 1 and the advantage in network coding capacity when one allows an \( \varepsilon > 0 \) error in communication as opposed to zero error communication.

II. MODEL

A \( k \)-source coding problem \((I, n, X)\) is defined by (a) a network instance \( I \), (b) a block length \( n \), and (c) a vector of random variables \( X \).

1) Instance \( I = (G, S, T, C, D) \) describes a directed, acyclic graph \( G(V, E) \), a multiset of source nodes \( S = \{s_1, \ldots, s_k\} \subset V \), a set of terminal nodes \( T = \{t_1, \ldots, t_{|T|}\} \subset V \), a capacity vector \( C = [c_e : e \in E] \), and a demand matrix \( D = [d_{i,j} : (i, j) \in [k] \times [|T|]] \), where for any real number \( N \geq 1 \), \([N] = \{1, \ldots, [N]\}\).

Without loss of generality, we assume that each source \( s \in S \) has no incoming edges and that each terminal \( t \in T \) has no outgoing edges. Capacity vector \( C \) describes the capacity \( c_e \) for each edge \( e \in E \). Binary demand matrix \( D \) specifies which sources are required at each terminal; namely, \( d_{i,j} = 1 \) if and only if terminal \( t_j \) requires the information originating at source \( s_i \).

2) The block length \( n \) specifies the number of available network uses.

3) The source vector \( X = (X_1, \ldots, X_k) \) describes the random source variable \( X_i \) available at each source \( s_i \). Random variables \( X_1, \ldots, X_k \) may be independent or dependent.

A solution \( X_n = \{X_e\} \) to the \( k \)-source coding problem \((I, n, X)\) assigns a random variable \( X_e \) to each edge \( e \in E \) with the following two properties:

1) Functionality: for \( e = (u, v) \), \( X_e \) is a deterministic function \( f_e \) of the random variables \( X_{e'} \) corresponding to incoming edges of node \( u \). Equivalently, setting \( In(u) \) to be the set of incoming edges of \( u \), \( Out(u) \) the set of outgoing edges from \( u \), and setting \( X_{In(u)} = \{X_{e'} : e' \in In(u)\} \), \( X_{Out(u)} = \{X_{e'} : e' \in Out(u)\} \); then

This work was supported in part by NSF grant CCF-1018741, ISF grant 480/08 and the Open University of Israel’s research fund (grant no. 46114). Work done at part while the first author was visiting Caltech.
dependent sources is also every vector of random variables $R$ with respect to a problem edge removal $R$ and (b) for all $t$ exists a block length $n$ of such that if there exists a network source code $X_n$ of independent sources is said to be $R$-feasible on independent sources? For example, it is true that in this case $X$ is $R_3$-feasible on independent sources if and only if there exists a block length $n$ of random variables $X = (X_1, X_2, X_3, X_4)$ such that for any fixed $X_1, X_2, X_3, X_4$ the $k$-source coding problem $(I, n, X)$ is feasible.

In what follows we address the following question:

**Question 2:** If $I$ is $R = (R_1, R_2)$-feasible on $\delta$-dependent sources, what can be said about its feasibility on independent sources? For example, is it true that in this case $I$ is $R_3 = (R_1 - \delta, R_2 - \delta)$ feasible on independent sources?

While we do not resolve Question 2, we show a strong connection between this question and the edge removal problem studied in [12], [13], [14].

**Remark 1:** It is natural to also define the notion of feasibility with respect to all $\delta$-dependent sources. Namely, a network instance $I$ is said to be strongly $R$-feasible on $\delta$-dependent sources if and only if there exists a block length $n$ such that for every vector of random variables $X = (X_1, X_2, X_3, X_4)$ for which (a) $\sum_{i=1}^k H(X_i) - H(X) \leq \delta n$ and (b) for all $i$, $H(X_i) \geq R_i n$; such that the $k$-source coding problem $(I, n, X)$ is feasible.

Under this definition it is not hard to verify that any instance $I$ which is strongly $R = (R_1, R_2)$ feasibility for $\delta$-dependent sources is also $R_{\delta/k} = (R_1 - \delta/k, R_2 - \delta/k)$ feasible for independent sources. To see this connection, consider a set of $\delta$-dependent sources $X_1, X_2, X_3, X_4$ where each $X_i$ is equal to the pair $(Y_i, Z)$ where for blocklength $n$ the $Y_i$’s are uniform in $[2^{\delta n/k}]$ and also independent of the $Y_i$’s. It follows that $X = X_1, X_2, X_3, X_4$ is $\delta$-dependent. Now, if $I$ is feasible on source information $X$ with block length $n$, then it is feasible with source information $X'_i = (Y_i, z)$ for any fixed value $z \in [2^{\delta n/k}]$. However, the random variables $X'_i$ are now independent and of rate $H(X'_i) = (R_1 - \delta/k) n$. We conclude that $I$ is $R_{\delta/k} = (R_1 - \delta/k, . . . , R_k - \delta/k)$ feasible on independent sources.

**A. The edge removal proposition**

The edge removal proposition compares the rates achievable on a given instance $I$ before and after an edge $e$ of capacity $c_e$ is removed from the network $G$.

**Proposition 1 (Edge removal):** Let $I = (G, S, T, C, D)$ be a network instance. Let $e \in G$ be an edge of capacity $\delta$. Let $I^e = (G^e, S, T, C, D)$ be the network obtained by replacing $G$ with the network $G^e$ in which edge $e$ is removed. Let $c > 0$ be some constant. Let $R = (R_1, . . . , R_k)$, and let $R_{\delta} = (R_1 - c\delta, . . . , R_k - c\delta)$. There exists a universal constant $c$, such that if $I$ is $R$-feasible on independent sources then $I^e$ is $R_{\delta}$-feasible on independent sources.

**B. The source coding proposition**

Addressing Questions 1 and 2 we consider the following proposition:

**Proposition 2 (Source coding):** Let $I = (G, S, T, C, D)$ be a network instance. Let $c > 0$ be some constant. Let $\delta > 0$. Let $R = (R_1, . . . , R_k)$, and let $R_{\delta} = (R_1 - c\delta, . . . , R_k - c\delta)$. There exists a universal constant $c$, such that if $I$ is $R$-feasible on $\delta$-dependent sources then $I$ is $R_{\delta}$-feasible on independent sources.

**III. Main Result**

Our main result shows that the two propositions above are equivalent.

**Theorem 1:** Proposition 1 is true if and only if Proposition 2 is true.

To prove Theorem 1 we will use the following two lemmas proven in Sections V and VI respectively.

**Lemma 1 (Edge removal lemma):** Let $I = (G, S, T, C, D)$ be a network instance. Let $e \in G$ be an edge of capacity $\delta$. Let $I^e = (G^e, S, T, B)$ be the instance obtained by replacing $G$ with the network $G^e$ in which edge $e$ is removed. Let $R = (R_1, . . . , R_k)$, and let $R_{\delta} = (R_1 - c\delta, . . . , R_k - c\delta)$. There exists a universal constant $c$, such that if $I$ is $R$-feasible on independent sources then $I^e$ is $R_{\delta}$-feasible on $\delta$-dependent sources.

**Lemma 2 (Collocated source coding):** Consider a network instance $I = (G, S, T, C, D)$ in which all sources are collocated at a single node in $G$ (i.e., each $s_i \in S$ equals the same vertex $s \in V$). Let $c > 0$ be some constant. Let $\delta > 0$. Let $R = (R_1, . . . , R_k)$, and let $R_{\delta} = (R_1 - c\delta, . . . , R_k - c\delta)$. There exists a universal constant $c$, such that if $I$ is $R$-feasible on $\delta$-dependent sources then $I$ is $R_{\delta}$-feasible on independent sources.

The following corollary of Lemma 2 is also proven in Section VI:
Corollary 1 (Super source): Let \( \mathcal{I} = (G, S, T, C, D) \) be a network instance. Let \( c > 0 \) be some (sufficiently large) constant. Let \( \delta > 0 \). Assume there is a vertex \( s \in V \) (so-called a “super source”) which has knowledge of all source information \( X = (X_1, \ldots, X_k) \), and in addition has outgoing edges of capacity \( c \delta \) to each and every source node in \( S \). Let \( R = (R_1, \ldots, R_k) \), and let \( R_{c,\delta} = (R_1 - c \delta, \ldots, R_k - c \delta) \). There exists a universal constant \( c \), such that if \( \mathcal{I} \) is \( R \)-feasible on \( \delta \)-dependent sources then \( \mathcal{I} \) is \( R_{c,\delta} \)-feasible on independent sources.

IV. PROOF OF THEOREM 1

We now present the proof of Theorem 1 using Lemma 1 and Corollary 1. Our proof will have two directions, each given in a separate subsection below.

A. Proposition 1 implies Proposition 2

In what follows, we show for the constant \( c \) in Corollary 1 that Proposition 1 is true with constant \( c_1 \) implies Proposition 2 with constant \( c_2 = c + c_1 \).

\[ \text{Proof: } \] Let \( \mathcal{I} = (G, S, T, C, D) \) be a network instance which is \( R \)-feasible on \( \delta \)-dependent sources. We show that \( \mathcal{I} \) is \( R_{c,\delta} \)-feasible on independent sources for \( c_2 = c + c_1 \).

We consider 2 additional instances \( \mathcal{I}_1 = (G_1, S_1, T, C, D) \) and \( \mathcal{I}_2 = (G_2, S_2, T, C, D) \) similar to those considered in [9], [14]. We start by defining the network \( G_2 \); network \( G_1 \) is then obtained from network \( G_2 \) by a single edge removal.

Network \( G_2 \) is obtained from \( G \) by adding \( k \) new source nodes \( s'_1, \ldots, s'_k \), a new “super-node” \( s \), and a relay node \( r \). For each \( s_i \in G \), there is a capacity-\( R_i \) edge \( (s'_i, s_i) \) from new source \( s'_i \) to old source \( s_i \). For each \( s'_i \in G_2 \), there is a capacity-\( R_i \) edge \( (s'_i, s) \) from new source \( s'_i \) to super-node \( s \). Let \( c \) be the constant from Corollary 1. There is a capacity-\( c \delta \) edge \( (s, r) \) from the super-source \( s \) to the relay \( r \); this edge is the network bottleneck. Finally, the relay \( r \) is connected to each source node \( s_i \) by an edge \( (r, s_i) \) of capacity \( c \delta \). The new source set \( S_2 \) is \( \{s'_1, \ldots, s'_k\} \). For \( \mathcal{I}_2 \), we set \( S_1 = S_2 \), and \( G_1 = G_2 \) apart from the removal of the bottleneck edge \( (s, r) \) of capacity \( c \delta \).

Our assertion now follows from the following arguments.

First note that \( \mathcal{I}_2 \) is \( R \)-feasible on \( \delta \)-dependent sources. This follows directly from our construction. Similarly, \( \mathcal{I}_2 \) is also \( R \)-feasible on \( \delta \)-dependent sources. Now, by Corollary 1, instance \( \mathcal{I}_2 \) is \( R_{c,\delta} \)-feasible on independent sources. Using Proposition 1, we have that \( \mathcal{I}_1 \) is \( R_{c,\delta} \)-feasible on independent sources. (Here \( c_1 \) is the universal constant from Proposition 1.) Finally, we conclude that \( \mathcal{I} \) is also \( R_{c,\delta} \)-feasible on independent sources.

B. Proposition 2 implies Proposition 1

We now prove that Proposition 2 is true with constant \( c_2 \) implies Proposition 1 is true with constant \( c_1 = 1 + c_2 \).

\[ \text{Proof: } \] Let \( \mathcal{I} = (G, S, T, C, D) \) be a network instance. Let \( e \in G \) be an edge (of capacity \( \delta \)). Let \( \mathcal{I}^e = (G^e, S, T, C, D) \) be the instance obtained by replacing \( G \) with the network \( G^e \) in which the edge \( e \) is removed. Assume that \( \mathcal{I} \) is \( R \)-feasible on independent sources. Lemma 1 implies that \( \mathcal{I}^e \) is \( R_0 \)-feasible on \( \delta \)-dependent sources. Now, using Proposition 2, it holds that \( \mathcal{I}^e \) is \( R_{1 + c_2} \)-feasible on independent sources. This suffices to complete our proof with \( c_1 = 1 + c_2 \).

V. PROOF OF LEMMA 1

We start with the following definition.

Definition 2: Let \( m \) be an integer. A vector \((h_\alpha)_{\alpha \in [m]} \) is indexed by all subsets of \([m]\) is said to be entropic if there exist a vector of random variables \((X_1, \ldots, X_m) \) such that \( h_\alpha \) is the joint entropy of the random variables \( \{X_i \mid i \in \alpha \} \). Let \( \Gamma_m^\alpha \) be the set of all entropic vectors representing \( m \) random variables.

The only property we will need from \( \Gamma^\alpha \) in this work is that it is closed with respect to linear combinations over the positive integers, namely:

\[ \text{Fact 1 (e.g., [115], p. 366): For } \{h_1\}_{i=1}^m \subset \Gamma_m^\alpha \text{ and positive integers } \{a_i\}_{i=1}^m \text{ it holds that } \sum a_i h_i \in \Gamma_m^\alpha. \]

We now turn to prove Lemma 1. Let \( \mathcal{I} = (G, S, T, C, D) \) be a network instance. Let \( R = (R_1, \ldots, R_k) \). Let \( e \in G \) be an edge of capacity \( \delta \). Let \( \mathcal{I}^e = (G^e, S, T, C, D) \) be the instance obtained by replacing \( G \) with the network \( G^e \) in which edge \( e \) is removed. Let \( R_\delta = (R_1 - \delta, \ldots, R_k - \delta) \). Assume that \( \mathcal{I} \) is \( R \)-feasible on independent sources. Thus, there exists a block length \( n \), and a code \( X_n \) realizing the feasibility of the \( k \)-source coding problem \((\mathcal{I}, n, X) \) with \( X = \{X_1, \ldots, X_k\} \) in each \( X_i \) is uniformly and independently distributed over \([2^{R_i}]\). Let \( X = \{X_1, \ldots, X_k, X_n\} = (X_1, \ldots, X_k, \{X_e \}_{e \in E}) \).

Consider the entropic vector \( h \) corresponding to \( X \). Let \( \Sigma \) be the union of the supports of the random variables \((X_1, \ldots, X_k, X_n) \). It holds that \( |\Sigma| \) is finite. This follows from the fact that in our setting the probability distribution governing \( X \) (and in particular the source random variables \( X_1, \ldots, X_k \)) has finite support of size \( \prod_{i=1}^k 2^{R_i} \). In what follows, we denote the support size \( \Pi_{i=1}^k 2^{R_i} \) by \( N \).

Notice, that all events of the form \( X_e = \sigma \) for \( e \in E \) have probability which is an integer multiple of \( 1/N \). We will use this fact shortly in our analysis.

Let \( e \) be the edge that we are removing, and assume that \( e_\delta = \delta \). For any value \( \sigma \in \Sigma \), consider the vector of random variables drawn from the conditional distribution on \( X \) given \( X_e = \sigma \); we denote this random variable by \( X^\sigma = \{(X_e \mid X_e = \sigma)\}_{e \in E \cup \{1, \ldots, k\}} \). Let \( h^\sigma \) be the entropic vector in \( \Gamma^\sigma \) corresponding to \( X^\sigma \), and consider the convex combination \( h^\sigma = \sum_\sigma \Pr[X_e = \sigma] h^\sigma \). As \( \Gamma^\sigma \) is not convex, the vector \( h^\sigma \) is not necessarily in \( \Gamma^\sigma \). However, as noted above, there exist integers \( n_\sigma \) such that \( \Pr[X_e = \sigma] = n_\sigma/N \).

Thus, by Fact 1,

\[ N \cdot h^\sigma = N \cdot \sum_\sigma \Pr[X_e = \sigma] h^\sigma = \sum_\sigma n_\sigma h^\sigma \in \Gamma^\alpha. \]

Let \( X^\sigma = (X_1^\sigma, X_2^\sigma, \ldots, X_k^\sigma, \{X_e^\sigma \}_{e \in E}) \) be the random variables corresponding to \( N \cdot h^\sigma \). In what follows, we show via the code \( X^\sigma \) that the problem \((\mathcal{I}^e, N, X^\sigma) = (R_1 - \delta, \ldots, R_k - \delta) \) feasible on \( \delta \)-dependent sources.
Effectively, the random variables in $X^c$ correspond to the variables in $X$ conditioned on $X_e$. For any subset $\alpha \subseteq E \cup \{1, 2, \ldots, k\}$ let $X_\alpha = \{X_e' \mid e' \in \alpha\}$. Similarly define $X^c_\alpha$. Then,

$$H(X^c_\alpha) = N \cdot h^c_\alpha = N \cdot \sum_{\sigma} \Pr[X_e = \sigma] h^\sigma_\alpha$$

$$= N \cdot \sum_{\sigma} \Pr[X_e = \sigma] H(X_\alpha | X_e = \sigma)$$

$$= NH(X_\alpha | X_e) = NH(X_\alpha, X_e) - H(X_e).$$

We conclude that for each subset $\alpha$ (and in particular for $\alpha = \{i\}$ corresponding to a certain source $s_i$) it holds that

$$NH(X_\alpha) \geq H(X^c_\alpha) \geq NH(X_\alpha) - \delta n.$$  

This implies that $X^c_1, \ldots, X^c_k$ are $\delta$-dependent with $H(X^c_e) \geq NH(R_1 - \delta)$. Namely, for $S = \{1, \ldots, k\}$,

$$H(X^c_S) = NH(X_S | X_e) = NH(X_S | H(X_e))$$

$$= N \left( \sum_{i=1}^k H(X_i) - \delta n \right)$$

$$\geq N \left( \sum_{i=1}^k H(X_i | X_e) - \delta n \right)$$

$$= \sum_i H(X^c_i) - \delta N n.$$  

In addition, we have that $H(X^c_e) = N \cdot H(X_e | X_e) = 0$, and thus throughout we may consider the value of $X^c_e$ to be a constant. Setting $X_e$ to a constant corresponds to communication over the graph $G^c$ (that does not contain the edge $e$ at all).

We now turn to analyze the Functionality and Capacity constraints with respect to $X^c$. For Functionality, let $u$ be a vertex in $G$, and let $Out(u)$ and $In(u)$ be its set of outgoing and incoming edges. For every vertex $u$ that is not the head of $e$, we have that $H(X_{Out(u)}, In(u) | X_{In(u)}) = H(X_{Out(u)} | X_{In(u)}) = 0$, and

$$0 \leq H(X^c_{Out(u)}, In(u) | X^c_{In(u)})$$

$$= H(X^c_{Out(u)}, In(u)) - H(X^c_{In(u)})$$

$$= N(H(X_{Out(u)}, In(u) | X_e) - H(X^c_{In(u)}))$$

$$= N(H(X_{Out(u)}, In(u), X_e) - H(X^c_{In(u)}))$$

$$\leq N \cdot H(X_{Out(u)} | X_{In(u)}) = 0.$$  

The third equality follows from the chain rule. The final inequality follows since conditioning reduces entropy.

When $u$ is the head of $e$, the set $In(u)$ in $G$ differs from the set $In^c(u)$ in $G^c$. Specifically, $In(u) = In^c(u) \cup \{e\}$. In this case,

$$0 \leq H(X^c_{Out(u)}, In^c(u)) - H(X^c_{In^c(u)})$$

$$= N(H(X_{Out(u)}, In^c(u) | X_e) - H(X^c_{In^c(u)}))$$

$$= N(H(X_{Out(u)}, In^c(u), X_e) - H(X^c_{In^c(u)}))$$

$$= N(H(X_{Out(u)}, In(u), X_e) - H(X^c_{In(u)}))$$

$$\leq N \cdot H(X_{Out(u)} | X_{In(u)}) = 0.$$  

For the Capacity constraints, for each edge $e' \in E$ it holds that $H(X^c_{e'}) = NH(X^c_{e'} | X_e) \leq NH(X^c_{e'}) \leq NH(X^c_{e'}).$

Finally, to show that $(\mathcal{E}, N, (X_1^c, \ldots, X_k^c))$ is $R_\delta$ $(R_1 - \delta, \ldots, R_k - \delta)$ feasible on $\delta$-dependent sources, we need to show for any terminal $t$ that requires source information $i$ that $H(X^c_{t_i}) = 0$. This follows similarly to the previous arguments based on the fact that $X$ satisfies $H(X_t | X_{In(t)}) = 0$. Specifically,

$$0 \leq H(X^c_{t_i} | X^c_{In(t)})$$

$$= NH(X_{In(t), t_i} | X_e) - H(X^c_{In(t), t_i})$$

$$= NH(X_{In(t), t_i}, X_e)$$

$$\leq NH(X_{In(t), t_i}) = 0.$$  

VI. Proof of Lemma 2 and Corollary 1

The proof that follows uses two intermediate claims proven in Section VI-A.

Claim 1: Let $\mathcal{I} = (G, S, T, C, D)$ be a network instance. Let $\delta > 0$. Let $R = (R_1, \ldots, R_k)$, and let $R_\delta = (R_1 - \delta, \ldots, R_k - \delta)$. If $\mathcal{I}$ is $R$-feasible on $\delta$-dependent sources $(X_1, \ldots, X_k)$ then $\mathcal{I}$ is $R_\delta$-feasible on $\delta$-dependent sources $(X_1, \ldots, X_k)$ such that $X_i$ is distributed over an alphabet $X_i$ of size at most $2^{\delta N n}$. Using Claim 1 we may assume throughout this section that our source random variables are $\delta$-dependent, and that each $X_i$ has entropy $H(X_i) \geq R_i n - \delta n$ and support $X_i$ of size at most $2^{\delta N n}$. For ease of presentation, we set $H(X_i) = R_i n$ and consider $X = (X_1, \ldots, X_k)$ to be $\delta$ dependent. Thus the constants $c$ for Lemma 2 and Corollary 1 need to be increased accordingly with respect to the constants $c$ computed below.

Claim 2: Let $\delta \geq 0$. Let $X = (X_1, \ldots, X_k)$ be a set of random variables over alphabets $(X_1, \ldots, X_k)$ such that (a) $\sum_{i=1}^k H(X_i) - H(X) \leq \delta n$, (b) for each $i$ the marginal distribution satisfies $H(X_i) = R_i n$, and (c) each $X_i$ has support $X_i$ of size at most $2^{\delta N n}$. There exist a constant $c$, such that for each $i$, there exists a partition $P_i = P_{i1}, \ldots, P_{ir_i}$ of $X_i$ in which (a) each $r_i$ is at least $2^{\delta N n - cn}$, (b) for each $i, j$, and $j'$, the size of $P_{ij}$ equals that of $P_{ij'}$ and (c) for every $(j_1, \ldots, j_k)$ with $j_i \in [r_i]$ the product space

$$P(j_1, \ldots, j_k) = P_{1j_1} \times P_{2j_2} \times \ldots P_{kj_k}$$

satisfies

$$\Pr[(X_1, \ldots, X_k) \in P(j_1, \ldots, j_k)] > 0.$$  

We now prove Lemma 2 using Claim 2 above. We then prove Corollary 1. Our proof follows the line of proof appearing in [9], [14]. Let $\mathcal{E} = (G, S, T, C, D)$ be a network instance in which all sources are collocated at a single node in $G$. Let $R = (R_1, \ldots, R_k)$. We assume that $\mathcal{E}$ is $R$ feasible on $\delta$ dependent sources. The following argument shows that there exists a constant $c$ such that $\mathcal{E}$ is $R_\delta$ feasible on independent sources.

Let $n$ be a block length such that there exists random source variables $X = (X_1, \ldots, X_k)$ with $H(X_i) = R_i n$ and $H(X) \geq \sum_i R_i n - \delta n$ such that the corresponding communication problem is feasible. The general idea is conceptually
simple. As $X_1, \ldots, X_k$ satisfy the condition in Claim 2, we may define the corresponding partitions $P_{i,j}$, product sets $P(j_i, j_k)$, and values $r_i$ appearing in the claim. Consider the random variables $X_1, \ldots, X_k$ where each $X_i$ is uniformly distributed over the set $[r_i]$. As $r_i$ is at least $2^{h_i, n - \delta n}$, $H(X) \geq \sum_i R_i n - c \delta n$. As

$$Pr[(X_1, \ldots, X_k) \in P(j_i, \ldots, j_k)] > 0$$

for all $j_i \in [r_i]$, we can identify (at least) a single k-tuple $(x_1, j_1, \ldots, x_k, j_k) \in P(j_i, \ldots, j_k)$ for which the communication of $(x_1, \ldots, x_k)$ over $I$ is successful (yields correct decoding). This enables the following communication scheme over $I$ with $X$ as input. For every k-tuple $j_1, \ldots, j_k \in [r_1] \times [r_2] \times \cdots \times [r_k]$, the (single) source node computes the corresponding k-tuple $(x_1, j_1, \ldots, x_k, j_k) \in P(j_i, \ldots, j_k)$ and continues the communication as if it were communicating with the original source information $X$. Since communication is successful, in this case every terminal that decodes some $x_{i,j}$ can deduce $j_i$ simply by finding the element of $P_i = P_{i,1}, \ldots, P_{i,r_i}$ that contains $x_{i,j}$. Lemma 2 follows.

The proof of Corollary 1 is very similar and also follows ideas of [9], [14]. As $r_i$ of Claim 2 is (at least) of size $2^{h_i, n - \delta n}$ and $|X| \leq 2^{R_{i, n} + \delta n}$, the size of $P_{i,j}$ is (at most) $2^{(c+1) \delta n}$. We conclude that a super source that has access to $X_1, \ldots, X_k$ and has rate $(c+1)\delta$ edges to every source node $s_i$ in $G$, when given a k-tuple $j_1, \ldots, j_k \in [r_1] \times [r_2] \times \cdots \times [r_k]$, can send the location of $x_{i,j}$ in $P_{i,j}$. This allows source $s_i$, which has $j_i$, to deduce $x_{i,j}$ based on the information it receives from the super source, source $s_i$ thus continues the protocol described in the proof to Lemma 2, sending $x_{i,j}$ as defined above.

A. Proof of Claims 1 and 2

1) Proof of Claim 1: The proof is based on standard typicality arguments and is omitted due to space limitations.

2) Proof of Claim 2: For simplicity of presentation, we present our proof for the case $k = 2$. A similar proof can be given for general $k$. We first note that the size of the support of $X = (X_1, X_2)$ is at least $2^H(X) \geq 2^{R_1 + R_2 n - \delta n}$, and that of $X_1$ in the range $[2^{R_{1, n}} 2^{R_{2, n} - \delta n}]$. Here, a pair $(x_1, x_2)$ is in the support of $(X_1, X_2)$ if it has positive probability. Thus for an average $x_1$ there are at least $2^{R_{2, n} - \delta n}$ values $x_2$ for which $p(x_1, x_2) > 0$. Consider the set $A_{i,j} \subset X_1$ such that there are at least $2^{R_{2, n} - \delta n}$ values $x_2$ for which $p(x_1, x_2) > 0$. It holds that the support $\Gamma$ of $(X_1, X_2)$ when $X_1$ is restricted to $A_{i,j}$ is at least of size $|A_{i,j}|2^{R_{2, n} - \delta n - 1}$. In addition, the size of $A_{i,j}$ is at least $2^{R_{1, n} - \delta n - 1}$. The latter follows from the fact that $|A_{i,j}|2^{R_{2, n} - \delta n - 1 + 2^{R_{2, n} + \delta n} - (1 - \delta n)} \geq \frac{2^{R_{1, n} + \delta n} - (1 - \delta n)}{(1 - \delta n)}$.

Let $\epsilon' \geq 1$ be a constant to be determined later in the proof. Now, consider a random partition $P_1$ of $A_{i,j}$ into $r_1 = |A_{i,j}|2^{\delta n}$ sets $P_{i,1}$, each of size $2^{\delta n}$. Consider also a random partition $P_2$ of $X_2$ into $r_2 = |X_2|2^{\delta n}$ sets $P_{2,1}$, each of size $2^{\delta n}$. Here we assume that $|A_{i,j}|$ and $|X_2|$ are divisible by $2^{\epsilon' \delta n}$, minor modifications in the proof are needed otherwise. For any $j_1, j_2$ we compute the probability that $P_{i,1} \times P_{2,1}$ intersects $\Gamma$ defined above. Let $x_1 \in P_{i,1}$. As $x_1$ is in $A_{i,j}$, the probability that there exists an $x_2$ such that $(x_1, x_2) \in \Gamma$ is at least (for sufficiently large $n$)

$$1 - \left(\frac{2^{R_{2, n} - 2^{R_{2, n} - \delta n - 1}}}{2^{\epsilon n}}\right) \approx 1 - e^{-2^{(c' - 2)\delta n}}$$

Using the union bound, this implies that there exist partitions $P_1$ and $P_2$ for which each and every “cell” $P_{i,1} \times P_{2,1}$ intersect $\Gamma$. Dividing the remaining elements $x_1$ among the sets $P_{i,1}$ evenly, and noticing that $r_1 = |A_{i,j}|2^{\delta n} \geq 2^{R_{1, n} - (c' - 2)\delta n - 1}$ suffices to prove our assertion for $c' = 3$ while the constant $c$ in the claim equals $c' + 2 = 5$.

REFERENCES


