

Separation of source-network coding and channel coding in wireline networks

Shirin Jalali, and Michelle Effros, *Fellow, IEEE*

Abstract

In this paper we prove the separation of source-network coding and channel coding in wireline networks. For the purposes of this work, a wireline network is any network of independent, memoryless, point-to-point, finite-alphabet channels used to transmit dependent sources either losslessly or subject to a distortion constraint. In deriving this result, we also prove that in a general memoryless network with dependent sources, lossless and zero-distortion reconstruction are equivalent provided that the conditional entropy of each source given the other sources is non-zero. Furthermore, we extend the separation result to the case of continuous-alphabet, point-to-point channels such as additive white Gaussian noise (AWGN) channels.

I. INTRODUCTION

In his seminal work [1], Shannon separates the problem of communicating a memoryless source across a single noisy, memoryless channel into separate lossless source coding and channel coding problems. The corresponding result for lossy coding in point-to-point channels is also proven in the same work. For a single point-to-point channel, separation holds under a wide variety of source and channel distributions (see, for example, [2] and the references therein). Unfortunately, separation does not necessarily hold in network systems. Even in very small networks like the multiple access channel [3], separation can fail when statistical dependencies between the sources at different network locations are useful for increasing the rate across the channel. Since source codes tend to destroy such dependencies, joint source-channel codes can achieve better performance than separate source and channel codes in these scenarios.

This paper proves the separation between source-network coding and channel coding in networks of independent noisy, discrete, memoryless channels (DMC); these networks are called *wireline networks* in this work. Roughly, we show that the vector of achievable distortions in delivering a family of dependent sources across such a network \mathcal{N} equals the vector of achievable distortions for delivering the same sources across a distinct network $\hat{\mathcal{N}}$. Network $\hat{\mathcal{N}}$ is built by replacing each channel $p(y|x)$ in \mathcal{N} by a noiseless, point-to-point bit-pipe of the corresponding

This paper was presented in part at IEEE International Symposium on Information Theory, Austin, Texas, 2010, and Information Theory and Applications (ITA), San Diego, CA, 2011.

S. Jalali is with the Center for the Mathematics of Information, California Institute of Technology, Pasadena, CA 91125 USA (e-mail: shirin@caltech.edu),

M. Effros is with the Department of Electrical Engineering, California Institute of Technology, Pasadena, CA 91125 USA (e-mail: effros@caltech.edu),

capacity $C = \max_{p(x)} I(X; Y)$. Thus a code that applies source-network coding across links that are made almost lossless through the application of independent channel coding across each link asymptotically achieves the optimal performance across the network as a whole.

Note that the operations of network source coding and network coding are not separable, as shown in [4] and [5] for lossless source coding in non-multicast and multicast networks, respectively. As a result, a joint network-source code is required, and only the channel code can be separated. While the achievability of a separated strategy is straightforward, the converse is more difficult since preserving statistical dependence between codewords transmitted across distinct edges of a network of noisy links improves the end-to-end network performance in some networks [6], [7], [8].

The results derived here are consistent with those of [9], [10], [8], which prove the separation between network coding and channel coding for multicast [9], [10] and general demands [6], [8], respectively, under the assumption that messages transmitted to different subsets of users are independent. The shift here is from independent sources to dependent sources and from reliable information delivery to both lossy and lossless data descriptions.

After hearing about our work, the author of [11] pointed us to his unpublished work from the 90s, which proves the separation of lossy network source coding and channel coding in three specific network structures, namely, the Slepian-Wolf configuration, the multiple description configuration, and Yamamoto's cascade network. In these cases, [11] proves separation without requiring the single-letter characterizations of the distortion regions. Our result generalizes this result to any network configuration that consists of point-to-point noisy channels. The strategy underlying our proof follows that of [6], [8], but the details differ significantly, both due to the inclusion of dependent sources and lossy reconstruction and in the focus on discrete-alphabet channels.

The organization of this paper is as follows. Sections II and III describe the notation and problem set-up, respectively. Section IV describes a tool from [8] called a stacked network that allows us to employ, in later arguments, typicality across copies of a network rather than typicality across time. Section V proves the separation of lossy source-network coding and channel coding. Section VI proves the equivalence of zero-distortion and lossless reconstruction in general memoryless channels. Section VII shows that the separation of source-network coding and channel coding continues to hold for well-behaved continuous channels such as AWGN channels under input power constraints. Section VIII concludes the paper.

The first part of the results presented in this paper, showing the separation of lossy source-network coding and channel coding in a wireline network was first presented at ISIT 2010 [12]. A similar result by other authors was presented at the same ISIT [13], where they prove that, in the same setup and under the finite source and channel alphabet assumption, if each source is required only at one other node, or at multiple other nodes but at the same distortion level, then separation of source-network coding and channel coding is optimal. For the general case, under a restricted set of distortion measures, they prove approximate optimality of separation strategy.

II. NOTATION AND DEFINITIONS

Finite sets are denoted by script letters such as \mathcal{X} and \mathcal{Y} . The size of a finite set \mathcal{A} is denoted by $|\mathcal{A}|$. Random variables are denoted by upper case letters such as X and Y . Bold face letters represent vectors. The alphabet of a random variable X is denoted by \mathcal{X} . Random vectors are represented by upper case bold letters like \mathbf{X} and \mathbf{Y} . The length of a vector is implied in the context. The ℓ^{th} element of a vector \mathbf{X} is denoted by X_ℓ . A vector $\mathbf{x} = (x_1, \dots, x_n)$ or $\mathbf{X} = (X_1, \dots, X_n)$ is sometimes represented as x^n or X^n . For $1 \leq i \leq j \leq n$, $x_i^j = (x_i, x_{i+1}, \dots, x_j)$. For a set $\mathcal{A} \subseteq \{1, 2, \dots, n\}$, $\mathbf{x}_{\mathcal{A}} = (x_i)_{i \in \mathcal{A}}$, where the elements are sorted in ascending order of their indices.

For two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^r$, $\mathbf{x} \leq \mathbf{y}$ iff $x_i \leq y_i$ for all $1 \leq i \leq r$. The ℓ_1 distance between two vectors \mathbf{x} and \mathbf{y} of the same length r is denoted by $\|\mathbf{x} - \mathbf{y}\|_1 = \sum_{i=1}^r |x_i - y_i|$. If \mathbf{x} and \mathbf{y} represent pmfs, i.e., $\sum_{i=1}^r x_i = \sum_{i=1}^r y_i = 1$ and $x_i, y_i \geq 0$ for all $i \in \{1, \dots, r\}$, then the total variation distance between \mathbf{x} and \mathbf{y} is defined as $\|\mathbf{x} - \mathbf{y}\|_{\text{TV}} = 0.5\|\mathbf{x} - \mathbf{y}\|_1$.

Definition 1: The empirical distribution of a sequence $x^n \in \mathcal{X}^n$ is defined as

$$\pi(x|x^n) \triangleq \frac{|\{i : x_i = x\}|}{n},$$

for all $x \in \mathcal{X}$. Similarly, the joint empirical distribution of a sequence $(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$ is defined as

$$\pi(x, y|x^n, y^n) \triangleq \frac{|\{i : (x_i, y_i) = (x, y)\}|}{n},$$

for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$.

Definition 2: For a random variable $X \sim p(x)$ and a constant $\epsilon > 0$, the set $\mathcal{T}_\epsilon^{(n)}(X)$ of ϵ -typical sequences¹ of length n is defined as

$$\mathcal{T}_\epsilon^{(n)}(X) \triangleq \{x^n : |\pi(x|x^n) - p(x)| \leq \epsilon p(x) \text{ for all } x \in \mathcal{X}\}.$$

For $(X, Y) \sim p(x, y)$, the set $\mathcal{T}_\epsilon^{(n)}(X, Y)$ of jointly ϵ -typical sequences is defined as

$$\mathcal{T}_\epsilon^{(n)}(X, Y) \triangleq \{(x^n, y^n) : |\pi(x, y|x^n, y^n) - p(x, y)| \leq \epsilon p(x, y), \text{ for all } (x, y) \in \mathcal{X} \times \mathcal{Y}\}.$$

We shall use $\mathcal{T}_\epsilon^{(n)}$ instead of $\mathcal{T}_\epsilon^{(n)}(X)$ or $\mathcal{T}_\epsilon^{(n)}(X, Y)$ when the random variable(s) are clear from the context.

For $x^n \in \mathcal{T}_\epsilon^{(n)}$, let

$$\mathcal{T}_\epsilon^{(n)}(Y|x^n) \triangleq \{y^n : (x^n, y^n) \in \mathcal{T}_\epsilon^{(n)}\}.$$

III. THE PROBLEM SETUP

Consider a multiterminal network \mathcal{N} consisting of m nodes interconnected via a collection of point-to-point, independent DMCs. The network structure is represented by a directed graph G with node set \mathcal{V} and edge

¹In this paper we only consider strong typicality, and use the definition introduced in [14].

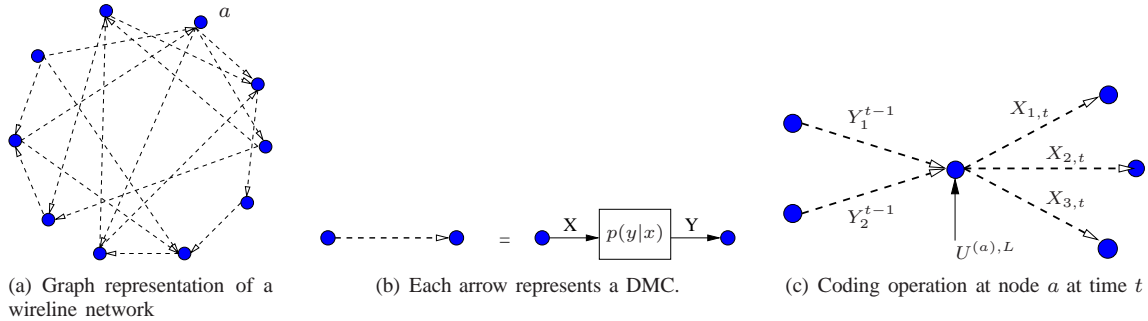


Fig. 1. Noisy wired network model

set \mathcal{E} . Each directed edge $e = (a, b) \in \mathcal{E}$ represents an independent point-to-point DMC $(\mathcal{X}_e, p(y_e|x_e), \mathcal{Y}_e)$ between nodes a (input) and b (output). For the channel represented by the edge e , the transition probabilities are $\{p(y_e|x_e)\}_{(x_e, y_e) \in \mathcal{X}_e \times \mathcal{Y}_e}$. The channels are independent by assumption, together giving a multiterminal channel $(\prod_{e \in \mathcal{E}} \mathcal{X}_e, \prod_{e \in \mathcal{E}} p(y_e|x_e), \prod_{e \in \mathcal{E}} \mathcal{Y}_e)$. The channel input at each node $a \in \mathcal{V}$ is $x^{(a)} = (x_{(a,v)} : (a, v) \in \mathcal{E})$. The channel output at node a is $y^{(a)} = (y_{(v,a)} : (v, a) \in \mathcal{E})$.

Each node a observes some source process $\mathbf{U}^{(a)} = \{U_k^{(a)}\}_{k=1}^\infty$ and is interested in reconstructing the processes observed by a subset of the other nodes. The alphabet $\mathcal{U}^{(a)}$ of source $\mathbf{U}^{(a)}$ can be either scalar- or vector-valued. A vector-valued source $\mathbf{U}^{(a)}$ denotes a collection of sources available at node a . In a block coding framework, source output symbols are divided into non-overlapping blocks of length L . Each block is described separately. At the beginning of the j^{th} coding period, each node a observes a length- L block of the process $\mathbf{U}^{(a)}$, i.e., $U_{(j-1)L+1}^{(a), jL} = (U_{(j-1)L+1}^{(a)}, \dots, U_{jL}^{(a)})$. The blocks $\{U_{(j-1)L+1}^{(a), jL}\}_{a \in \mathcal{V}}$ observed at the nodes $a \in \mathcal{V}$ are described over n uses of the network. The rate

$$\kappa \triangleq \frac{L}{n}$$

is a parameter of the code. At each time $t \in \{1, \dots, n\}$, each node a generates its next channel inputs as a function of its source observation $U_{(j-1)L+1}^{(a), jL}$ and its observed channel outputs $Y^{(a), t-1} = (Y_1^{(a)}, \dots, Y_{t-1}^{(a)})$ up to time $t-1$ using encoder

$$X_t^{(a)} : (\mathcal{Y}^{(a)})^{t-1} \times \mathcal{U}^{(a), L} \rightarrow \mathcal{X}^{(a)}. \quad (1)$$

Note that each node might have more than one incoming channel and more than one outgoing channel. Thus, $X_t^{(a)}$ and $Y_t^{(a)}$ are vectors with dimensions equal to the outdegree and indegree of node a , respectively. The reconstruction at node b of the source vector observed at node a is denoted by $\hat{U}^{(a \rightarrow b), L}$. This reconstruction is determined using a decoder with inputs equal to the source and channel outputs observed at node b . Thus, $\hat{U}^{(a \rightarrow b), L} = \hat{U}^{(a \rightarrow b)}(Y^{(b), n}, U^{(b), L})$, where

$$\hat{U}^{(a \rightarrow b)} : \mathcal{Y}^{(b), n} \times \mathcal{U}^{(a), L} \rightarrow \hat{\mathcal{U}}^{(a \rightarrow b), L}. \quad (2)$$

The performance of a given code is the vector of expected average distortions between the sources $\{\mathbf{U}^{(a)}\}_{a \in \mathcal{V}}$ and reconstructions $\{\hat{\mathbf{U}}^{(a \rightarrow b)}\}_{a, b \in \mathcal{V}}$. For each $a, b \in \mathcal{V}$,

$$\mathbb{E}[d_L^{(a \rightarrow b)}(U^{(a), L}, \hat{U}^{(a \rightarrow b), L})] \triangleq \mathbb{E} \left[\frac{1}{L} \sum_{k=1}^L d^{(a \rightarrow b)}(U_k^{(a)}, \hat{U}_k^{(a \rightarrow b)}) \right],$$

where $d^{(a \rightarrow b)} : \mathcal{U}^{(a)} \times \hat{\mathcal{U}}^{(a \rightarrow b)} \rightarrow \mathbb{R}^+$ is a per-letter distortion measure. As mentioned before $\mathcal{U}^{(a)}$ and $\hat{\mathcal{U}}^{(a \rightarrow b)}$ may be either scalar or vector-valued. This allows the case where node a observes multiple sources and node b is interested in reconstructing a subset of them. Let

$$d_{\max} \triangleq \max_{\substack{(a, b) \in \mathcal{V}^2 \\ \alpha \in \mathcal{U}^{(a)}, \beta \in \hat{\mathcal{U}}^{(a \rightarrow b)}}} d^{(a \rightarrow b)}(\alpha, \beta) < \infty.$$

The $|\mathcal{V}| \times |\mathcal{V}|$ distortion matrix D is said to be achievable at rate κ , if for any $\epsilon > 0$, and any L large enough, there exists a blocklength- (L, n) coding scheme such that

$$\frac{L}{n} \geq \kappa - \epsilon,$$

and

$$\mathbb{E}[d_L^{(a \rightarrow b)}(U^{(a), L}, \hat{U}^{(a \rightarrow b), L})] \leq D(a, b) + \epsilon, \quad (3)$$

for every $(a, b) \in \mathcal{V}^2$. Let $\mathcal{D}(\kappa, \mathcal{N})$ denote the set of achievable distortion matrices at rate κ in network \mathcal{N} .

Remark 1: While here we are assuming that all sources have a fixed rate κ , in general, the rate κ can vary for different sources. Our results continue to hold in that case as well. However, for notational simplicity, we assume that κ is fixed among all sources.

Throughout the paper, for any network \mathcal{N} of noisy point-to-point channels described by directed graph G , let the network \mathcal{N}_b denote a network of noiseless point-to-point channels described by the same directed graph G . Precisely, network \mathcal{N}_b replaces each noisy DMC $(\mathcal{X}_e, p(y_e|x_e), \mathcal{Y}_e)$, $e \in \mathcal{E}$, by a noiseless bit pipe of the same finite capacity $C_e = \max_{p(x_e)} I(X_e; Y_e)$. A bit pipe of capacity C_e is an error-free, point-to-point communication channel that delivers, in n channel uses, $\lfloor nC_e \rfloor$ bits from the transmitter to the receiver, for any $n \geq 1$. The timing of the delivery of these bits has no impact on the set of achievable distortion matrices. This result is shown for the network capacity problem in [6]; the same argument goes through immediately for the case of lossy reconstruction.

Example 1: Fig. 2 demonstrates a simple example of the kind of networks we study in this paper. The graph of the network, shown in Fig. 2(a), consists of two edges and three nodes. Each edge models a point-to-point DMC. Fig. 2(b) shows a specific realization of such a network with sources available at nodes 1 and 2. It also shows how the encoding and decoding operations are performed on network of Fig. 2(a). The decoder reconstructs both sources U_1 and U_2 . For $i \in \{1, 2\}$, let C_i denote the noisy capacity of channel i . Fig. 2(c) shows the equivalent noiseless model. At coding rate κ , choosing $n = \lfloor \kappa^{-1} L \rfloor$, $W_i \in \{1, \dots, 2^{\lfloor nC_i \rfloor}\}$. For this special example, [11] proves that at $\kappa = 1$ separation is optimal and the set of achievable distortions on both networks are equal. In this

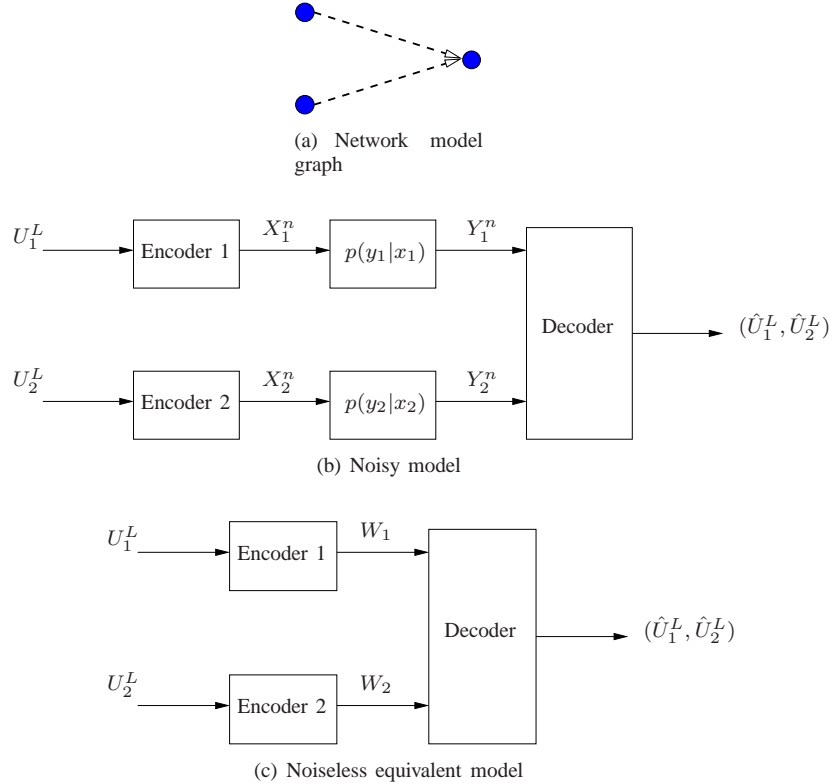


Fig. 2. Example 1

paper, we extend this result to general networks of point-to-point noisy channels, at arbitrary coding rate κ .

IV. STACKED NETWORK

The stacked network is a tool introduced in [6] for proving separation results. The key underlying observation is that by taking multiple copies of the same network and applying the same code to that network in each copy, we create i.i.d. copies of the input and output of a given channel at each time t . This allows us to later employ typicality arguments to our channel inputs and outputs across copies of the network and not across time. Applying typicality arguments across time is problematic since the inputs to the channel at different times t need not be i.i.d.

For a given network \mathcal{N} , the corresponding N -fold stacked network $\underline{\mathcal{N}}$ is defined as N copies of the original network [6]. That is, for each node $a \in \mathcal{V}$ and each edge $e \in \mathcal{E}$ in \mathcal{N} , there are N copies of node a and N copies of edge e in $\underline{\mathcal{N}}$. At each time instance, each node has access to the data available at all copies of node a , and each may use this extra information in generating the channel inputs for future time instances. Likewise, in decoding, all N copies of a node can collaborate in reconstructing the source vectors. This is made more precise in the following two definitions. The encoder $\mathbf{X}_t^{(a)}$ for node a at time t in N -fold stacked network $\underline{\mathcal{N}}$ is a mapping

$$\mathbf{X}_t^{(a)} : \mathcal{Y}^{(a), N(t-1)} \times \mathcal{U}^{(a), NL} \rightarrow \mathcal{X}^{(a), N}, \quad (4)$$

and the node- b decoder $\hat{U}^{(a \rightarrow b), NL}$ for signal $U^{(a), NL}$ of node a is a mapping

$$\hat{U}^{(a \rightarrow b), NL} : \mathcal{Y}^{(b), nN} \times \mathcal{U}^{(b), NL} \rightarrow \hat{\mathcal{U}}^{(a \rightarrow b), NL}. \quad (5)$$

These definitions correspond to (1) and (2) in network \mathcal{N} . In (4), network input $\mathbf{X}_t^{(a)}$ is a vector with N components denoted by $\mathbf{X}_t^{(a)} = (\mathbf{X}_t^{(a)}(1), \dots, \mathbf{X}_t^{(a)}(N))$.

In the N -fold stacked network, the distortion between the source originating at node a and its reconstruction at node b is defined as

$$D_N(a, b) = \mathbb{E} \left[d_{NL}^{(a \rightarrow b)}(U^{(a), NL}, \hat{U}^{(a \rightarrow b), NL}) \right],$$

for any $(a, b) \in \mathcal{V} \times \mathcal{V}$.

A distortion matrix D is said to be achievable at rate κ in the stacked version of network \mathcal{N} , if for any given $\epsilon > 0$, there exist n , L , and N such that distortion D and rate κ are achievable in the N -fold stacked network; that is, $L/n \geq \kappa - \epsilon$ and $D_N(a, b) \leq D(a, b) + \epsilon$ for all $a, b \in \mathcal{V}$ on N -fold stacked network $\underline{\mathcal{N}}$. Let $\mathcal{D}_s(\kappa, \underline{\mathcal{N}})$ denote the set of achievable distortion matrices at rate κ in the stacked network $\underline{\mathcal{N}}$. Note that the depth N of the stacked network $\underline{\mathcal{N}}$ on which each distortion matrix $D \in \mathcal{D}_s(\kappa, \mathcal{N})$ is achievable may vary with D .

Note that the dimension of the distortion matrices in both single layer and multi-layer networks is $m \times m$. The following theorem establishes the relationship between the two sets.

Theorem 1: At any rate κ ,

$$\mathcal{D}(\kappa, \mathcal{N}) = \mathcal{D}_s(\kappa, \underline{\mathcal{N}}). \quad (6)$$

Proof of Theorem 1:

- i. $\mathcal{D}(\kappa, \mathcal{N}) \subseteq \mathcal{D}_s(\kappa, \underline{\mathcal{N}})$: This is obvious, because the stacked network is a generalization of the original network. In fact, choosing $N = 1$, any distortion matrix D that is achievable on \mathcal{N} is also achievable on its stacked version too.
- ii. $\mathcal{D}_s(\kappa, \underline{\mathcal{N}}) \subseteq \mathcal{D}(\kappa, \mathcal{N})$: The proof is very similar to the proof of the analogous part of Lemma 1 in [8], but, for completeness, we present the proof in Appendix A. ■

V. REPLACING A NOISY CHANNEL WITH A BIT PIPE

We assume that the sources are independent and identically distributed (i.i.d.) according to some distribution $p(u^{(1)}, u^{(2)}, \dots, u^{(m)})$. That is, for any $k \geq 1$,

$$\begin{aligned} \mathbb{P} \left((U^{(1), k}, \dots, U^{(m), k}) = (u^{(1), k}, \dots, u^{(m), k}) \right) \\ = \prod_{i=1}^k p(u_i^{(1)}, \dots, u_i^{(m)}). \end{aligned}$$

For the given i.i.d. source assumption, Theorem 2 proves that the space of achievable distortions for networks \mathcal{N} and \mathcal{N}_b are identical. The proof follows the proof strategy of [6, Theorem 3], showing that any code for network $\underline{\mathcal{N}}_b$ can be applied across network \mathcal{N} with the aid of a channel code and any code for \mathcal{N} can be applied across network $\underline{\mathcal{N}}_b$ with the aid of an “emulation code”. Just as a channel code enables us to emulate a noiseless bit pipe across a noisy channel, an emulation code enables us to emulate a noisy channel across a noiseless bit pipe. The result proves the optimality of separate source-network codes and channel codes on networks of point-to-point DMCs. Notice, however, that separate codes are here applied in the manner described in the proof of Theorem 1 rather than the more conventional direct application across time.

Theorem 2: For a network \mathcal{N} of independent point-to-point DMCs with memoryless sources,

$$\mathcal{D}(\kappa, \mathcal{N}) = \mathcal{D}(\kappa, \mathcal{N}_b), \quad (7)$$

for any $\kappa > 0$.

Proof of Theorem 2: By Theorem 1, the achievable region of a network \mathcal{N} is equal to the achievable region of its stacked network $\underline{\mathcal{N}}$. Hence, $\mathcal{D}(\kappa, \mathcal{N}) = \mathcal{D}_s(\kappa, \underline{\mathcal{N}})$ and $\mathcal{D}(\kappa, \mathcal{N}_b) = \mathcal{D}_s(\kappa, \underline{\mathcal{N}}_b)$, and therefore, it suffices to prove that $\mathcal{D}_s(\kappa, \underline{\mathcal{N}}) = \mathcal{D}_s(\kappa, \underline{\mathcal{N}}_b)$.

- i. $\mathcal{D}_s(\kappa, \underline{\mathcal{N}}_b) \subseteq \mathcal{D}_s(\kappa, \underline{\mathcal{N}})$: Note that \mathcal{N} and \mathcal{N}_b are identical except that for each $e \in \mathcal{E}$, DMC $(\mathcal{X}_e, p(y_e|x_e), \mathcal{Y}_e)$ in \mathcal{N} is replaced by a bit pipe of capacity $C = \max_{p(x_e)} I(X_e; Y_e)$ in \mathcal{N}_b . We next show that any code for network $\underline{\mathcal{N}}_b$ can be operated on $\underline{\mathcal{N}}$ with a similar expected distortion. Fix any code of source blocklength LN , channel blocklength nN and expected distortion matrix D for N -fold stacked network $\underline{\mathcal{N}}_b$. Now consider a pN -fold stacked network $\tilde{\underline{\mathcal{N}}}_b$. By partitioning the pN layers into p stacks, each consisting of N layers, and then applying the code independently to these stacks, we can construct a code for network $\tilde{\underline{\mathcal{N}}}_b$, which has the same expected distortion matrix D . Consider a pM -fold stacked network $\underline{\mathcal{N}}$, with $M > N$. Using the mentioned strategy to construct a code for the pN -fold stacked network from the code given for the N -fold network, at each time step t , each bit pipe e in $\tilde{\underline{\mathcal{N}}}_b$ sends a message of at most $p\lfloor NC_e \rfloor$ bits across the pN copies of edge e in $\tilde{\underline{\mathcal{N}}}_b$. To operate the same code on network $\underline{\mathcal{N}}$, we need to send the same information across the pM copies of DMC $(\mathcal{X}_e, p(y_e|x_e), \mathcal{Y}_e)$ in $\underline{\mathcal{N}}$. To achieve this goal, we use a channel code of blocklength pM operating at rate $R_e < C_e$. By choosing $M = \lceil NC_e/R_e \rceil$, we guarantee that $pMR_e \geq pNC_e$. Hence the M copies of DMC $(\mathcal{X}_e, p(y_e|x_e), \mathcal{Y}_e)$ in $\underline{\mathcal{N}}$ carry the same information as the N copies of bit pipe e in $\underline{\mathcal{N}}_b$. Since the capacity of DMC $(\mathcal{X}_e, p(y_e|x_e), \mathcal{Y}_e)$ equals C_e , R_e can be made arbitrarily close to C_e . The rate of the code for $\underline{\mathcal{N}}$ is

$$\frac{pNL}{pMn} = \frac{NL}{\lceil NC_e/R_e \rceil n},$$

which can be made arbitrary close to $\kappa = L/n$.

Let $P_e^{(M)}$ denote the maximal probability of error for the channel code of blocklength pM used over the pM copies of DMC $(\mathcal{X}_e, p(y_e|x_e), \mathcal{Y}_e)$ in $\underline{\mathcal{N}}$. Let $P_{\max}^{(M)} = \max_{e \in \mathcal{E}} P_e^{(M)}$. The code for each channel e is used n times – once for each $t \in \{1, \dots, n\}$. Errors in the channel code for e increase the distortion achieved

by applying the code for $\tilde{\mathcal{N}}_b$ across \mathcal{N} . We can bound this increase in the expected average distortion using the union bound. More precisely, let \mathcal{R} denote the event that there is a decoding error in at least one of the channels $e \in \mathcal{E}$ at some time step $t \in \{1, 2, \dots, n\}$. Since the sources and channel codes are independent,

$$\begin{aligned} \mathbb{E}[d_{pNL}(U^{(a),pNL}, \hat{U}^{(a \rightarrow b),pNL})] &= \mathbb{E}[d_{pNL}(U^{(a),pNL}, \hat{U}^{(a \rightarrow b),pNL}) | \mathcal{R}^c] \mathbb{P}(\mathcal{R}^c) \\ &\quad + \mathbb{E}[d_{pNL}(U^{(a),pNL}, \hat{U}^{(a \rightarrow b),pNL}) | \mathcal{R}] \mathbb{P}(\mathcal{R}) \\ &\leq D(a, b) + n|\mathcal{E}|P_{\max}^{(M)}d_{\max}, \end{aligned}$$

for each $(a, b) \in \mathcal{V}^2$. Therefore, for fixed n and N , letting $p \rightarrow \infty$, $|\mathbb{E}[d_{pNL}(U^{(a),pNL}, \hat{U}^{(a \rightarrow b),pNL})] - D(a, b)|$ can be made arbitrarily small for each $(a, b) \in \mathcal{V}^2$.

- ii. $\mathcal{D}_s(\kappa, \mathcal{N}) \subseteq \mathcal{D}_s(\kappa, \mathcal{N}_b)$: Let $D \in \mathcal{D}(\kappa, \mathcal{N})$. We prove that $D \in \mathcal{D}_s(\kappa, \mathcal{N}_b)$. Consider a code defined on \mathcal{N} with source blocklength L , channel blocklength n , and an expected distortion matrix that is component-wise upper-bounded by $D + \epsilon \cdot \mathbf{1}$. Applying this code independently in each layer of N -fold stacked network \mathcal{N} gives a code for \mathcal{N} with $D_N(a, b) \leq D(a, b) + \epsilon$, for all $(a, b) \in \mathcal{V}^2$. Throughout the rest of the proof, n and L , corresponding to the source and channel blocklengths of the mentioned code, are fixed. To simulate the performance of this code on the stacked version of \mathcal{N}_b , we let the number of layers N go to infinity. As shown in [8] any code for an N -fold stacked network can be unraveled across time to give a single-layer code with the same performance. The blocklength for that code goes to infinity as N grows without bound.

We first show that for identically distributed memoryless sources, the performance of the code given the realization of $(\mathbf{X}_{e,1}, \mathbf{Y}_{e,1})$ depends only on the empirical distribution $\{\pi(x_e, y_e | \mathbf{X}_{e,1}, \mathbf{Y}_{e,1})\}_{(x_{e,1}, y_{e,1}) \in \mathcal{X}_e \times \mathcal{Y}_e}$ of $(\mathbf{X}_{e,1}, \mathbf{Y}_{e,1})$. Here the subscript 1 refers to time $t = 1$. After establishing this, we use the result proved in [15] and show that at time $t = 1$ we can emulate the behavior of the noisy link across a bit pipe of the same capacity. For the rest of the proof, let $\mathbf{U} = \{U_t\}$ denote an i.i.d. source observed at some node in $a \in \mathcal{V}$ and $\hat{\mathbf{U}} = \{\hat{U}_t\}$ denote its reconstruction at some other node $b \in \mathcal{V} \setminus \{a\}$.

In network \mathcal{N} , the expected distortion between source vector U^L and its reconstruction \hat{U}^L is

$$\begin{aligned} \mathbb{E}[d_L(U^L, \hat{U}^L)] &= \\ &\sum_{(x_{e,1}, y_{e,1}) \in \mathcal{X}_e \times \mathcal{Y}_e} \mathbb{E} \left[d_L(U^L, \hat{U}^L) | (X_{e,1}, Y_{e,1}) = (x_{e,1}, y_{e,1}) \right] \\ &\quad \times \mathbb{P}((X_{e,1}, Y_{e,1}) = (x_{e,1}, y_{e,1})). \end{aligned} \tag{8}$$

In the N -fold stacked network \mathcal{N} , the reconstruction of the corresponding N independent copies of U^L by

reproduction \hat{U}^{NL} satisfies

$$\begin{aligned}
& \mathbb{E} \left[d_{NL}(U^{NL}, \hat{U}^{NL}) \right] \\
&= \mathbb{E} \left[\frac{1}{N} \sum_{\ell=1}^N d_L \left(U_{(\ell-1)L+1}^{\ell L}, \hat{U}_{(\ell-1)L+1}^{\ell L} \right) \times \right. \\
&\quad \left. \sum_{(x_{e,1}, y_{e,1}) \in \mathcal{X}_e \times \mathcal{Y}_e} \mathbb{1}_{(\mathbf{X}_{e,1}(\ell), \mathbf{Y}_{e,1}(\ell)) = (x_{e,1}, y_{e,1})} \right] \\
&= \frac{1}{N} \sum_{(x_{e,1}, y_{e,1}) \in \mathcal{X}_e \times \mathcal{Y}_e} \sum_{\ell=1}^N \\
&\quad \mathbb{E} \left[d_L \left(U_{(\ell-1)L+1}^{\ell L}, \hat{U}_{(\ell-1)L+1}^{\ell L} \right) \mathbb{1}_{(\mathbf{X}_{e,1}(\ell), \mathbf{Y}_{e,1}(\ell)) = (x_{e,1}, y_{e,1})} \right]. \tag{9}
\end{aligned}$$

For any random variables A and B , $\mathbb{E}[A\mathbb{1}_{B=b}] = \sum_{a,b'} a\mathbb{1}_{b'=b}p(a,b) = \sum_a ap(a,b) = p(b)\mathbb{E}[A|B=b]$. Using this equality, and since the code used on $\underline{\mathcal{N}}$ applies the solution for \mathcal{N} independently in each layer of stacked network $\underline{\mathcal{N}}$, it follows that

$$\begin{aligned}
& \mathbb{E} \left[d_L \left(U_{(\ell-1)L+1}^{\ell L}, \hat{U}_{(\ell-1)L+1}^{\ell L} \right) \mathbb{1}_{(\mathbf{X}_{e,1}(\ell), \mathbf{Y}_{e,1}(\ell)) = (x_{e,1}, y_{e,1})} \right] \\
&= \mathbb{E}[d_L(U_{(\ell-1)L+1}^{\ell L}, \hat{U}_{(\ell-1)L+1}^{\ell L}) | (\mathbf{X}_{e,1}(\ell), \mathbf{Y}_{e,1}(\ell)) = (x_{e,1}, y_{e,1})] \mathbb{P}((\mathbf{X}_{e,1}(\ell), \mathbf{Y}_{e,1}(\ell)) = (x_{e,1}, y_{e,1})) \\
&= \mathbb{E}[d_L(U^L, \hat{U}^L) | (\mathbf{X}_{e,1}(\ell), \mathbf{Y}_{e,1}(\ell)) = (x_{e,1}, y_{e,1})] \mathbb{P}((\mathbf{X}_{e,1}(\ell), \mathbf{Y}_{e,1}(\ell)) = (x_{e,1}, y_{e,1})) \tag{10}
\end{aligned}$$

where each conditional expectation of $d_L(U^L, \hat{U}^L)$ in (10) equals the corresponding conditional expectation in (8). Combining (9) and (10) yields

$$\begin{aligned}
\mathbb{E} \left[d_{NL}(U^{NL}, \hat{U}^{NL}) \right] &= \sum_{(x_{e,1}, y_{e,1}) \in \mathcal{X}_e \times \mathcal{Y}_e} \mathbb{E} \left[d_L(U^L, \hat{U}^L) | (X_{e,1}, Y_{e,1}) = (x_{e,1}, y_{e,1}) \right] \times \\
&\quad \frac{1}{N} \sum_{\ell=1}^N \mathbb{P}((\mathbf{X}_{e,1}(\ell), \mathbf{Y}_{e,1}(\ell)) = (x_{e,1}, y_{e,1})) \\
&= \sum_{(x_{e,1}, y_{e,1}) \in \mathcal{X}_e \times \mathcal{Y}_e} \mathbb{E} \left[d_L(U^L, \hat{U}^L) | (X_{e,1}, Y_{e,1}) = (x_{e,1}, y_{e,1}) \right] \times \\
&\quad \mathbb{E}[\pi(x_{e,1}, y_{e,1} | \mathbf{X}_{e,1}, \mathbf{Y}_{e,1})]. \tag{11}
\end{aligned}$$

Equations (8) and (11) differ only in their distributions on $\mathcal{X}_e \times \mathcal{Y}_e$. Since each conditional expectation is finite (in particular, all are bounded by d_{\max}), we can replace channel $(\mathcal{X}_e, p(y_e|x_e), \mathcal{Y}_e)$ by a bit pipe of capacity C_e at time $t=1$, if we can find a coding scheme across the layers of the stack for which,

$$\left| \mathbb{P}((X_{e,1}, Y_{e,1}) = (x_{e,1}, y_{e,1})) - \mathbb{E}[\pi(x_{e,1}, y_{e,1} | \mathbf{X}_{e,1}, \mathbf{Y}_{e,1})] \right|, \tag{12}$$

can be made arbitrary small, for all $(x_{e,1}, y_{e,1}) \in \mathcal{X}_e \times \mathcal{Y}_e$.

To prove that this is possible, consider a channel with input drawn i.i.d. from some distribution $p(x_{e,1})$. We wish to build an emulation code with an encoder that maps N source symbols, $\mathbf{X}_{e,1} \in \mathcal{X}_e^N$, to a message

of NR bits and a decoder that converts these NR bits into a reconstruction block $\mathbf{Y}_{e,1} \in \mathcal{Y}_e^N$. We aim to use this code to emulate the DMC with transition probabilities $\{p(y_{e,1}|x_{e,1})\}_{(x_{e,1},y_{e,1}) \in \mathcal{X}_e \times \mathcal{Y}_e}$ when the channel input is an i.i.d. process drawn according to $p(x_{e,1})$. The codebook, $\mathcal{C}^{(N)}$, of this emulation code consists of 2^{NR} codewords, $\{\mathbf{Y}_{e,1}[1], \mathbf{Y}_{e,1}[2], \dots, \mathbf{Y}_{e,1}[2^{NR}]\}$, each drawn independently i.i.d. according to $p(y_{e,1}) = \sum_{x_{e,1} \in \mathcal{X}_e} p(x_{e,1})p(y_{e,1}|x_{e,1})$. The encoder assigns message $M \in \{1, \dots, 2^{NR}\}$ to input sequence $\mathbf{X}_{e,1}$, if $(\mathbf{X}_{e,1}, \mathbf{Y}_{e,1}[M]) \in \mathcal{T}_\epsilon^{(N)}(X_{e,1}, Y_{e,1})$. If there are multiple such messages in the codebook, the encoder chooses the one with the smallest index. If there exist no codewords in $\mathcal{C}^{(N)}$ that are jointly typical with $\mathbf{X}_{e,1}$, then the encoder assigns message $M = 1$ to $\mathbf{X}_{e,1}$. After receiving message M , the decoder outputs $\mathbf{Y}_{e,1}[M]$. Let $\{\pi(x_{e,1}, y_{e,1}|\mathbf{X}_{e,1}, \mathbf{Y}_{e,1})\}_{(x_{e,1}, y_{e,1}) \in \mathcal{X}_e \times \mathcal{Y}_e}$ be the joint empirical distribution between the channel input and channel output induced by running the emulation code across the N copies of the bit pipe at time $t = 1$. In [15], it is shown that, the described code can emulate channel $(\mathcal{X}_e, p(y_{e,1}|x_{e,1}), \mathcal{Y}_e)$ by a bit pipe of rate R , provided that $R > I(X_{e,1}; Y_{e,1})$. The given emulation ensures that the total variation between $\pi(x_{e,1}, y_{e,1}|\mathbf{X}_{e,1}, \mathbf{Y}_{e,1})$ and $p(x_{e,1}, y_{e,1}) = p(x_{e,1})p(y_{e,1}|x_{e,1})$ can be made arbitrarily small as the blocklength N grows without bound. In other words, there exists a sequence of codes over the bit pipe such that

$$\|\pi - \mathbf{p}\|_{\text{TV}} \xrightarrow{N \rightarrow \infty} 0, \quad (13)$$

almost surely. (Here π and \mathbf{p} are vectors describing distributions $(\pi(x_{e,1}, y_{e,1}|\mathbf{X}_{e,1}, \mathbf{Y}_{e,1}) : (x_{e,1}, y_{e,1}) \in \mathcal{X}_e \times \mathcal{Y}_e)$ and $(p(x_{e,1}, y_{e,1}) : (x_{e,1}, y_{e,1}) \in \mathcal{X}_e \times \mathcal{Y}_e)$ respectively.) Although Theorem 3 in [15] only guarantees convergence of π to \mathbf{p} in probability, we can also prove almost sure convergence of π to \mathbf{p} using Borel-Cantelli Lemma. Let $\gamma = R - I(X_{e,1}; Y_{e,1})$. Let $\mathbf{Y}_{e,1}(\mathbf{X}_{e,1})$ denote the codeword in $\mathcal{C}^{(N)}$ that is assigned to $\mathbf{X}_{e,1}$ by the emulation encoder. For $\epsilon > 0$, define the error event

$$\mathcal{E}^{(N)} = \{(\mathbf{X}_{e,1}, \mathbf{Y}_{e,1}(\mathbf{X}_{e,1})) : \|\pi - \mathbf{p}\|_{\text{TV}} > \epsilon\}.$$

Breaking the error event into two parts and then applying the union bound, Hoeffding's inequality, and the joint typicality lemma from [16] gives

$$\begin{aligned} \mathbb{P}(\mathcal{E}^{(N)}) &\leq \mathbb{P}(\mathbf{X}_{e,1} \notin \mathcal{T}_\epsilon^{(N)}(X_{e,1})) + \left[\mathbb{P}((\mathbf{X}_{e,1}, \mathbf{Y}_{e,1}[1]) \notin \mathcal{T}_\epsilon^{(N)}(X_{e,1}, Y_{e,1})) \right]^{2^{NR}} \\ &\leq \sum_{x_e \in \mathcal{X}_e} \mathbb{P}(|\pi(x_e|\mathbf{X}_{e,1}) - p(x_e)| > p(x_e)\epsilon) + e^{-2^{N(\gamma-\delta(\epsilon))}} \\ &\leq \sum_{x_e \in \mathcal{X}_e} 2e^{-2N\epsilon^2 p^2(x_e)} + e^{-2^{N(\gamma-\delta(\epsilon))}} \\ &\leq 2|\mathcal{X}_e|e^{-2N\epsilon^2 \min_{x_e \in \mathcal{X}_e} p^2(x_e)} + e^{-2^{N(\gamma-\delta(\epsilon))}}, \end{aligned} \quad (14)$$

where $\delta(\epsilon) = \epsilon(H(Y_{e,1}) + H(Y_{e,1}|X_{e,1})) \rightarrow 0$, as $\epsilon \rightarrow 0$. Therefore,

$$\sum_{N=1}^{\infty} \mathbb{P}(\mathcal{E}^{(N)}) < \infty,$$

and hence (13) holds almost surely, by the Borel-Cantelli Lemma.

We next combine the emulation code with the code for $\underline{\mathcal{N}}$. The code emulates channel $p(y_e|x_e)$ at time $t = 1$ across the N layers of stacked network $\underline{\mathcal{N}}'_b$ that replaces $p(y_e|x_e)$ by a link of capacity $R > C$, only at time $t = 1$. The given code for $\underline{\mathcal{N}}$ can be run across $\underline{\mathcal{N}}'_b$ with expected distortion bounded as

$$\begin{aligned} & \mathbb{E} \left[d_{NL}(U^{NL}, \hat{U}^{NL}) \right] \\ &= \sum_{(x_e, y_e) \in \mathcal{X}_e \times \mathcal{Y}_e} \mathbb{E} \left[d_L(U^L, \hat{U}^L) | (X_{e,1}, Y_{e,1}) = (x_e, y_e) \right] \times \\ & \quad \mathbb{E}[\pi(x_e, y_e | \mathbf{X}_{e,1}, \mathbf{Y}_{e,1})] \\ &\leq \sum_{(x_e, y_e) \in \mathcal{X}_e \times \mathcal{Y}_e} \mathbb{E} \left[d_L(U^L, \hat{U}^L) | (X_{e,1}, Y_{e,1}) = (x_e, y_e) \right] (p(x_e, y_e) + \epsilon) \\ &\leq \mathbb{E}[d_L(U^L, \hat{U}^L)] + \epsilon d_{\max}. \end{aligned}$$

Thus we can replace the noisy link by a bit-pipe at time $t = 1$. We use induction to extend this result to the next $n - 1$ time steps. Note that in the original network

$$\begin{aligned} & \mathbb{E}[d_L(U^L, \hat{U}^L)] = \\ & \quad \sum_{(x_e^n, y_e^n) \in \mathcal{X}_e^n \times \mathcal{Y}_e^n} \mathbb{E} \left[d_L(U^L, \hat{U}^L) | (X_e^n, Y_e^n) = (x_e^n, y_e^n) \right] \\ & \quad \times \mathbb{P}((X_e^n, Y_e^n) = (x_e^n, y_e^n)). \end{aligned} \tag{15}$$

On the other hand, using the same analysis used in deriving (11), in the N -fold stacked network,

$$\begin{aligned} & \mathbb{E} \left[d_{NL}(U^{NL}, \hat{U}^{NL}) \right] \\ &= \sum_{(x_e^n, y_e^n) \in \mathcal{X}_e^n \times \mathcal{Y}_e^n} \mathbb{E} \left[d_L(U^L, \hat{U}^L) | (X_e^n, Y_e^n) = (x_e^n, y_e^n) \right] \\ & \quad \times \mathbb{E}[\pi(x_e^n, y_e^n | \mathbf{X}_e^n, \mathbf{Y}_e^n)]. \end{aligned} \tag{16}$$

Here $\mathbf{X}_e^n = (\mathbf{X}_{e,1}, \mathbf{X}_{e,2}, \dots, \mathbf{X}_{e,n})$ and $\mathbf{Y}_e^n = (\mathbf{Y}_{e,1}, \mathbf{Y}_{e,2}, \dots, \mathbf{Y}_{e,n})$ refer to the inputs and outputs of channel e in the N layers of the stacked network, for times $t = 1, 2, \dots, n$, while $\mathbf{X}_e^n(\ell)$ and $\mathbf{Y}_e^n(\ell)$ correspond to the inputs and outputs of the emulated channel at layer ℓ for times $t = 1, 2, \dots, n$, and

$$\pi(x_e^n, y_e^n | \mathbf{X}_e^n, \mathbf{Y}_e^n) = \frac{|\{\ell : (\mathbf{X}_e^n(\ell), \mathbf{Y}_e^n(\ell)) = (x_e^n, y_e^n)\}|}{N}.$$

Therefore, we need to show that by appropriate coding over the bit-pipes,

$$|\mathbb{P}((X_e^n, Y_e^n) = (x_e^n, y_e^n)) - \pi(x_e^n, y_e^n | \mathbf{X}_e^n, \mathbf{Y}_e^n)| \quad (17)$$

can be made arbitrarily small. Note that

$$\begin{aligned} \mathbb{P}((X_e^n, Y_e^n) = (x_e^n, y_e^n)) &= \\ \prod_{t=1}^n \mathbb{P}((X_{e,t}, Y_{e,t}) = (x_{e,t}, y_{e,t}) | (X_e^{t-1}, Y_e^{t-1}) = (x_e^{t-1}, y_e^{t-1})), & \end{aligned} \quad (18)$$

and

$$\pi(x_e^n, y_e^n | \mathbf{X}_e^n, \mathbf{Y}_e^n) = \prod_{t=1}^n \frac{\pi(x_e^t, y_e^t | \mathbf{X}_e^t, \mathbf{Y}_e^t)}{\pi(x_e^{t-1}, y_e^{t-1} | \mathbf{X}_e^{t-1}, \mathbf{Y}_e^{t-1})}, \quad (19)$$

where for $t = 1$

$$\pi(x_e^{t-1}, y_e^{t-1} | \mathbf{X}_e^{t-1}, \mathbf{Y}_e^{t-1}) = 1.$$

We have already proven that we can make the first term in the product in (19) converge to the first term in the product in (18) with probability one. We next prove by induction that the same result is true for each subsequent term in (18) and (19). Since all of the terms in (19) are positive and upper-bounded by 1, so too is their product. Thus, the Dominated Convergence Theorem (see, for example, [17]) shows that (17) can be made arbitrarily small provided that each term converges almost surely.

To apply induction, assume that there exist $t - 1$ emulation codes whose application makes the first $t - 1$ terms in (19) each converge to the corresponding term in (18) almost surely. Using this inductive hypothesis, we prove that the t^{th} term in (19) converges to the t^{th} term in (18) as well.

Given the inductive hypothesis that

$$\frac{\pi(x_e^{t'}, y_e^{t'} | \mathbf{X}_e^{t'}, \mathbf{Y}_e^{t'})}{\pi(x_e^{t'-1}, y_e^{t'-1} | \mathbf{X}_e^{t'-1}, \mathbf{Y}_e^{t'-1})} \rightarrow p(x_{e,t'}, y_{e,t'} | x_e^{t'-1}, y_e^{t'-1}) \quad (20)$$

almost surely, for all $(x_e^{t'}, y_e^{t'})$ and all $t' \leq t - 1$, it follows that

$$\pi(x_e^{t-1}, y_e^{t-1} | \mathbf{X}_e^{t-1}, \mathbf{Y}_e^{t-1}) \rightarrow p(x_e^{t-1}, y_e^{t-1}) \quad (21)$$

almost surely, for all (x_e^{t-1}, y_e^{t-1}) . Since the two networks apply precisely the same deterministic code to the channel outputs at time $t - 1$ to create the channel inputs at time t , this bound implies

$$\pi(x_e^t, y_e^{t-1} | \mathbf{X}_e^t, \mathbf{Y}_e^{t-1}) \rightarrow p(x_e^t, y_e^{t-1}) \quad (22)$$

almost surely, for all (x_e^t, y_e^{t-1}) as well. We now show that if the emulation code used at time t is generated

independently of the codes used at times $1, 2, \dots, t-1$, then for each (x_e^t, y_e^t) ,

$$\frac{\pi(x_e^t, y_e^t | \mathbf{X}_e^t, \mathbf{Y}_e^t)}{\pi(x_e^t, y_e^{t-1} | \mathbf{X}_e^t, \mathbf{Y}_e^{t-1})} \rightarrow p(y_{e,t} | x_{e,t}) \quad (23)$$

almost surely, where $p(y_{e,t} | x_{e,t}) = P(Y_e = y_{e,t} | X_e = x_{e,t})$. Note that

$$\begin{aligned} & P(\mathbf{Y}_{e,t}(1) = y_{e,t} | (\mathbf{X}_e^t(1), \mathbf{Y}_e^{t-1}(1)) = (x_e^t, y_e^{t-1})) \\ &= \sum_{s^N \in \mathcal{X}_e^N: s_1 = x_{e,1}} P(\mathbf{Y}_{e,t}(1) = y_{e,t}, \mathbf{X}_{e,t}(2:N) = s_2^N | (\mathbf{X}_e^t(1), \mathbf{Y}_e^{t-1}(1)) = (x_e^t, y_e^{t-1})) \\ &= \sum_{s^N \in \mathcal{X}_e^N: s_1 = x_{e,1}} P(\mathbf{X}_{e,t}(2:N) = s_2^N | (\mathbf{X}_e^t(1), \mathbf{Y}_e^{t-1}(1)) = (x_e^t, y_e^{t-1})) \\ &\quad \times P(\mathbf{Y}_{e,t}(1) = y_{e,t} | \mathbf{X}_{e,t} = s^N, (\mathbf{X}_e^{t-1}(1), \mathbf{Y}_e^{t-1}(1)) = (x_e^t, y_e^{t-1})) \\ &= \sum_{s^N \in \mathcal{X}_e^N: s_1 = x_{e,1}} P(\mathbf{X}_{e,t}(2:N) = s_2^N | (\mathbf{X}_e^t(1), \mathbf{Y}_e^{t-1}(1)) = (x_e^t, y_e^{t-1})) \\ &\quad \times P(\mathbf{Y}_{e,t}(1) = y_{e,t} | \mathbf{X}_{e,t} = s^N), \end{aligned} \quad (24)$$

where the last equality holds because $(\mathbf{X}_e^{t-1}, \mathbf{Y}_e^{t-1}) \rightarrow \mathbf{X}_{e,t} \rightarrow \mathbf{Y}_{e,t}$ since the emulation code maps $\mathbf{X}_{e,t}$ to $\mathbf{Y}_{e,t}$ independently of all prior channel inputs and outputs.

Since each network layer independently operates an identical code, and codewords in the emulation codebook are generated according to an i.i.d. distribution, it follows that

$$P(\mathbf{Y}_{e,t}(1) = y_{e,t} | \mathbf{X}_{e,t} = s^N) = P(\mathbf{Y}_{e,t}(\ell) = y_{e,t} | \mathbf{X}_{e,t} = s^N)$$

for any ℓ such that $s_\ell = x_{e,t}$ under the operation of a random emulation code. Therefore,

$$\begin{aligned} & P(\mathbf{Y}_{e,t}(1) = y_{e,t} | \mathbf{X}_{e,t} = s^N) \\ &= \frac{1}{N\pi(x_{e,t} | s^N)} \sum_{\ell: s_\ell = x_{e,t}} P(\mathbf{Y}_{e,t}(\ell) = y_{e,t} | \mathbf{X}_{e,t} = s^N) \\ &= \frac{1}{N\pi(x_{e,t} | s^N)} \sum_{\ell: s_\ell = x_{e,t}} E[\mathbb{1}_{\mathbf{Y}_{e,t}(\ell) = y_{e,t}} | \mathbf{X}_{e,t} = s^N] \\ &= E \left[\frac{1}{N\pi(x_{e,t} | s^N)} \sum_{\ell: s_\ell = x_{e,t}} \mathbb{1}_{\mathbf{Y}_{e,t}(\ell) = y_{e,t}} \middle| \mathbf{X}_{e,t} = s^N \right] \\ &= E \left[\frac{\pi(x_{e,t}, y_{e,t} | \mathbf{X}_{e,t}, \mathbf{Y}_{e,t})}{\pi(x_{e,t} | \mathbf{X}_{e,t})} \middle| \mathbf{X}_{e,t} = s^N \right]. \end{aligned} \quad (25)$$

By our inductive assumption and an argument similar to the one used in Remark 1, if $s^N \in \mathcal{T}_e^{(N)}(X_{e,t})$, for

N large enough

$$\left| \mathbb{E} \left[\frac{\pi(x_{e,t}, y_{e,t} | \mathbf{X}_{e,t}, \mathbf{Y}_{e,t})}{\pi(x_{e,t} | \mathbf{X}_{e,t})} \middle| \mathbf{X}_{e,t} = s^N \right] - p(y_{e,t} | x_{e,t}) \right| < \epsilon. \quad (26)$$

Combining (24), (25) and (26), it follows that

$$\begin{aligned} & \mathbb{P}(\mathbf{Y}_{e,t}(1) = y_{e,t} | (\mathbf{X}_e^t(1), \mathbf{Y}_e^{t-1}(1)) = (x_e^t, y_e^{t-1})) \\ &= \sum_{s^N \in \mathcal{T}_\epsilon^{(N)}(X_{e,t}) : s_1 = x_{e,1}} \mathbb{P}(\mathbf{X}_{e,t}(2:N) = s_2^N | (\mathbf{X}_e^t(1), \mathbf{Y}_e^{t-1}(1)) = (x_e^t, y_e^{t-1})) \\ & \quad \times \mathbb{P}(\mathbf{Y}_{e,t}(1) = y_{e,t} | \mathbf{X}_{e,t} = s^N) \\ &+ \sum_{s^N \notin \mathcal{T}_\epsilon^{(N)}(X_{e,t}) : s_1 = x_{e,1}} \mathbb{P}(\mathbf{X}_{e,t}(2:N) = s_2^N | (\mathbf{X}_e^t(1), \mathbf{Y}_e^{t-1}(1)) = (x_e^t, y_e^{t-1})) \\ & \quad \times \mathbb{P}(\mathbf{Y}_{e,t}(1) = y_{e,t} | \mathbf{X}_{e,t} = s^N) \\ &\leq (p(y_{e,t} | x_{e,t}) + \epsilon) \mathbb{P}(\mathbf{X}_{e,t} \in \mathcal{T}_\epsilon^{(N)}(X_{e,t}) | (\mathbf{X}_e^t(1), \mathbf{Y}_e^{t-1}(1)) = (x_e^t, y_e^{t-1})) \\ & \quad + \mathbb{P}(\mathbf{X}_{e,t} \notin \mathcal{T}_\epsilon^{(N)}(X_{e,t}) | (\mathbf{X}_e^t(1), \mathbf{Y}_e^{t-1}(1)) = (x_e^t, y_e^{t-1})). \end{aligned} \quad (27)$$

Similarly,

$$\begin{aligned} & \mathbb{P}(\mathbf{Y}_{e,t}(1) = y_{e,t} | (\mathbf{X}_e^t(1), \mathbf{Y}_e^{t-1}(1)) = (x_e^t, y_e^{t-1})) \\ &\geq (p(y_{e,t} | x_{e,t}) - \epsilon) \mathbb{P}(\mathbf{X}_{e,t} \in \mathcal{T}_\epsilon^{(N)}(X_{e,t}) | (\mathbf{X}_e^t(1), \mathbf{Y}_e^{t-1}(1)) = (x_e^t, y_e^{t-1})). \end{aligned} \quad (28)$$

But, if $\mathbb{P}((\mathbf{X}_e^t(1), \mathbf{Y}_e^{t-1}(1)) = (x_e^t, y_e^{t-1})) \neq 0$, then

$$\begin{aligned} & \mathbb{P}(\mathbf{X}_{e,t} \notin \mathcal{T}_\epsilon^{(N)}(X_{e,t}) | (\mathbf{X}_e^t(1), \mathbf{Y}_e^{t-1}(1)) = (x_e^t, y_e^{t-1})) \\ &= \frac{\mathbb{P}(\mathbf{X}_{e,t} \notin \mathcal{T}_\epsilon^{(N)}(X_{e,t}), (\mathbf{X}_e^t(1), \mathbf{Y}_e^{t-1}(1)) = (x_e^t, y_e^{t-1}))}{\mathbb{P}((\mathbf{X}_e^t(1), \mathbf{Y}_e^{t-1}(1)) = (x_e^t, y_e^{t-1}))} \\ &\leq \frac{\mathbb{P}(\mathbf{X}_{e,t} \notin \mathcal{T}_\epsilon^{(N)}(X_{e,t}))}{\mathbb{P}((\mathbf{X}_e^t(1), \mathbf{Y}_e^{t-1}(1)) = (x_e^t, y_e^{t-1}))} \rightarrow 0 \end{aligned} \quad (29)$$

as $N \rightarrow \infty$, and hence $\mathbb{P}(\mathbf{X}_{e,t} \in \mathcal{T}_\epsilon^{(N)}(X_{e,t}) | (\mathbf{X}_e^t(1), \mathbf{Y}_e^{t-1}(1)) = (x_e^t, y_e^{t-1})) \rightarrow 1$, as $N \rightarrow \infty$. Therefore, combining (27), (28), and (29), it follows that, for each (x_e^t, y_e^t) ,

$$\mathbb{P}(\mathbf{Y}_{e,t}(1) = y_{e,t} | (\mathbf{X}_e^t(1), \mathbf{Y}_e^{t-1}(1)) = (x_e^t, y_e^{t-1})) \rightarrow p(y_{e,t} | x_{e,t}), \quad (30)$$

almost surely, as N grows to infinity.

This concludes the proof, because it shows that, for each $\ell \in \{1, 2, \dots, N\}$, as the number of layers N grows, $\mathbf{Y}_{e,t}(\ell)$ becomes independent of $(\mathbf{X}_e^{t-1}(\ell), \mathbf{Y}_e^{t-1}(\ell))$ conditioned on $\mathbf{X}_{e,t}(\ell)$, and its conditional distribution

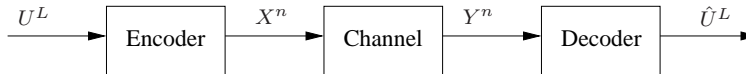


Fig. 3. Simple point-to-point channel

converges to $p(y_{e,t}|x_{e,t})$ corresponding to the transition probability of channel e . ■

Remark 2: The first part of the proof of Theorem 2 is not specific to DMCs, and shows that $\mathcal{D}(\kappa, \mathcal{N}_b) \subseteq \mathcal{D}(\kappa, \mathcal{N})$ for all networks \mathcal{N} of (discrete or continuous) point-to-point channels.

VI. CONTINUITY: ZERO-DISTORTION VERSUS LOSSLESS

The distortion criteria for lossless source coding and lossy source coding with a distortion constraint of zero are different. In lossless coding, we require that the probability of error in reconstructing a vector of source symbols goes to zero as the blocklength of that vector grows without bound. In lossy coding, we require that the per symbol distortion between the source vector and its reconstruction approach zero for sufficiently long blocklengths. As a result, even under the Hamming distortion measure, distortion 0 reconstructions do not necessarily meet the lossless source reconstruction criterion. Before investigating the relationship between these problems in a generic network \mathcal{N} of the form defined in Section III, we consider some special cases where the relationship is known. Consider the simple point-to-point network shown in Fig. 3. Let the source U be i.i.d. and distributed according to $p(u)$, and let $C = \max_{p(x)} I(X; Y)$ denote the capacity of the point-to-point channel connecting the source and the destination. The minimal required rate for describing the source U at distortion D is [18] $R(D) = \min_{p(\hat{u}|u): \mathbb{E}[d(U, \hat{U})] \leq D} I(U; \hat{U})$. In such point-to-point networks separation of source coding and channel coding is known to be optimal [1]. Hence to describe the source at distortion D , we need $C \geq \kappa R(D)$. Evaluating $R(D)$ at $D = 0$ gives

$$R(0) = \min_{p(\hat{u}|u): \mathbb{E}[d(U, \hat{U})] = 0} I(U; \hat{U}) = I(U; U) = H(U),$$

where $H(U)$ is the entropy rate of the source U . Since the minimal rate for lossless reconstruction of the source U is also the entropy rate, the zero-distortion and lossless reconstruction rate regions coincide in this simple network. Explicit characterizations of the multi-dimensional rate-distortion regions for general multiuser networks are unknown. Therefore, proving or disproving the equivalence of zero-distortion and lossless reconstruction rate-regions in such networks requires more elaborate analysis. In his Ph.D. thesis, W.H. Gu proved that in noiseless networks consisting of point-to-point bit-pipes, zero-distortion and lossless reconstruction rate regions coincide [19].

In this section, we prove the equivalence of zero-distortion reconstruction and lossless reconstruction in general networks described by multiuser discrete memoryless channels (mDMCs) with statistically dependent sources. More precisely, we prove that in any mDMC with independent or dependent sources, lossless reconstruction is achievable if and only if zero-distortion reconstruction is achievable.

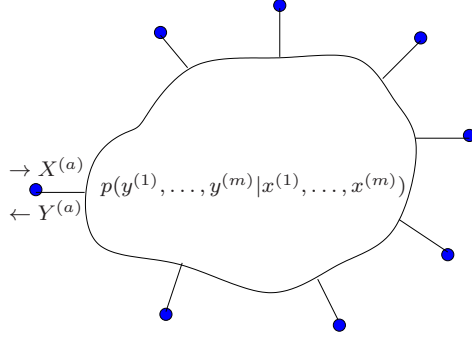


Fig. 4. General multiuser discrete memoryless channel (mDMC)

Consider network \mathcal{N} shown in Fig. 4, which consists of a general mDMC described by

$$p(y^{(1)}, \dots, y^{(m)} | x^{(1)}, \dots, x^{(m)}).$$

Let $\mathcal{V} \triangleq \{1, \dots, m\}$. Node $a \in \mathcal{V}$ observes source process $U^{(a)}$ and is interested in reconstructing sources observed by the other nodes. The coding operations are very similar to the case of wired networks. Each node observes a block of length L of its own source symbols and describes them to the other nodes in n channel uses. As before, the coding rate κ is defined as $\kappa = L/n$. At each each time $t = 1, \dots, n$, node a generates channel input $X_t^{(a)}$ as a function its own source block $U^{(a),L}$ and its received channel outputs up to time $t - 1$, i.e., $Y^{(a),t-1}$. In other words, $X_t^{(a)} = X_t^{(a)}(U^{(a),L}, Y^{(a),t-1})$. The set of achievable distortion matrices on network \mathcal{N} at rate κ is denoted by $\mathcal{D}(\kappa, \mathcal{N})$. Throughout this section we assume that for any $(a, b) \in \mathcal{V}^2$, $d^{(a \rightarrow b)}(u, \hat{u}) = 0$ if and only if $u = \hat{u}$.

Given any $D \in \mathcal{D}(\kappa, \mathcal{N})$, let

$$\mathcal{L}(D) \triangleq \{(a, b) : D(a, b) = 0\}.$$

Theorem 3: Fix any non-negative matrix $D = (D(a, b) : (a, b) \in \mathcal{V}^2)$ with $|\mathcal{L}(D)| > 0$. For any $(a, b) \in \mathcal{L}(D)$, assume that $H(U^{(a)} | (U^{(c)})_{c \in \mathcal{V} \setminus a}) > 0$. Then $D \in \mathcal{D}(\kappa, \mathcal{N})$ if and only if, for any $\epsilon > 0$ there exists integers L and $n \leq L/(\kappa - \epsilon)$, for which we can design a code of source blocklength L and channel blocklength n that satisfies

$$\mathbb{P}(U^{(a),L} \neq \hat{U}^{(a \rightarrow b),L}) \leq \epsilon,$$

for all $(a, b) \in \mathcal{L}(D)$ and

$$\mathbb{E}[d_L(U^{(a),L}, \hat{U}^{(a \rightarrow b),L})] \leq D(a, b) + \epsilon,$$

for all $(a, b) \in \mathcal{V}^2 \setminus \mathcal{L}(D)$.

Proof of Theorem 3:

For the forward result, fix a sequence of codes at rate $L/n \rightarrow \kappa$, distortion $\mathbb{E}[d_L(U^{(a),L}, \hat{U}^{(a \rightarrow b),L})] \rightarrow D(a, b)$ for all $(a, b) \notin \mathcal{L}(D)$ and error probability $\mathbb{P}(U^{(a),L} \neq \hat{U}^{(a \rightarrow b),L}) \rightarrow 0$ for all $(a, b) \in \mathcal{L}(D)$. For each $(a, b) \in \mathcal{L}(D)$,

the given sequence of codes satisfies

$$\begin{aligned}
& \mathbb{E}[d_L(U^{(a),L}, \hat{U}^{(a \rightarrow b),L})] \\
&= \mathbb{E} \left[d_L(U^{(a),L}, \hat{U}^{(a \rightarrow b),L}) \mid U^{(a),L} \neq \hat{U}^{(a \rightarrow b),L} \right] \mathbb{P}(U^{(a),L} \neq \hat{U}^{(a \rightarrow b),L}) \\
&+ \mathbb{E} \left[d_L(U^{(a),L}, \hat{U}^{(a \rightarrow b),L}) \mid U^{(a),L} = \hat{U}^{(a \rightarrow b),L} \right] \mathbb{P}(U^{(a),L} = \hat{U}^{(a \rightarrow b),L}) \\
&\leq d_{\max} \mathbb{P}(U^{(a),L} \neq \hat{U}^{(a \rightarrow b),L}).
\end{aligned}$$

Since the given bound approaches 0 as $\mathbb{P}(U^{(a),L} \neq \hat{U}^{(a \rightarrow b),L}) \rightarrow 0$, the sequence of codes achieves zero-distortion reconstruction of source a at node b , which is the the desired result.

To prove the converse, fix any $D \in \mathcal{D}(\kappa, \mathcal{N})$ with $|\mathcal{L}(D)| > 0$ and any $\epsilon > 0$. By the definition of $\mathcal{D}(\kappa, \mathcal{N})$, for any $\epsilon > 0$, there exists a code with source blocklength L and channel blocklength $n \leq L/(\kappa - \epsilon)$ such that

$$\mathbb{E}[d_L(U^{(a),L}, \hat{U}^{(a \rightarrow b),L})] \leq D(a, b) + \epsilon \quad (31)$$

for each $(a, b) \in \mathcal{V}^2$. Specifically, for any (a, b) such that $D(a, b) = 0$,

$$\mathbb{E}[d_L(U^{(a),L}, \hat{U}^{(a \rightarrow b),L})] \leq \epsilon.$$

We now prove that with an asymptotically negligible increase in number of channel uses n , node a can send node b sufficient information to improve node b 's reconstruction of node a 's data from a zero-distortion reproduction to a lossless reconstruction. We further show that this change preserves the quality of all other reconstructions.

The following argument builds a code of source blocklength NL and channel blocklength $n(N + N')$, for some integer N' to be defined shortly, from the given code of source blocklength L and channel blocklength n .

Each node $a \in \mathcal{V}$ breaks its incoming source block of length NL into N non-overlapping blocks of length L , given by

$$U^{(a),L}, U_{L+1}^{(a),2L}, \dots, U_{(N-1)L+1}^{(a),NL}.$$

Each node then applies the blocklength- L code N times to independently code each of these blocks. In total, this requires Nn channel uses. Independently decoding each L -block with the blocklength- L decoder achieves, for each $a, b \in \mathcal{V}$, a reconstruction of length NL such that

$$\mathbb{E}[d_L(U_{(\ell-1)L+1}^{(a),\ell L}, \hat{U}_{(\ell-1)L+1}^{(a \rightarrow b),\ell L})] \leq D(a, b) + \epsilon, \quad (32)$$

for each $\ell = 1, 2, \dots, N$.

For $(a, b) \in \mathcal{L}(D)$ and each $\ell \in \{1, \dots, N\}$, denote the input of node a in session ℓ as

$$U^L(\ell) \triangleq U_{(\ell-1)L+1}^{(a),\ell L},$$

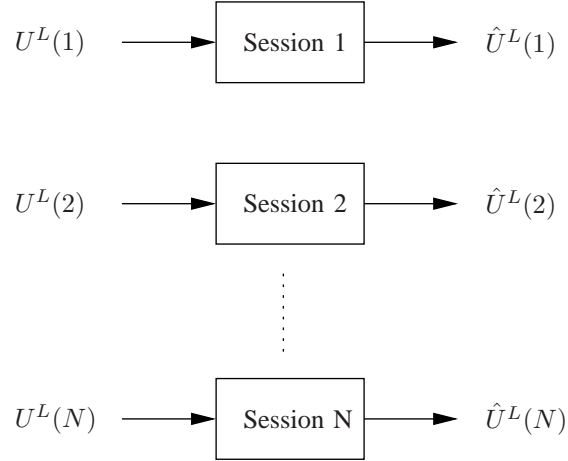


Fig. 5. Source U and its reconstruction at L parallel sessions.

and the corresponding output at node b as

$$\hat{U}^L(\ell) \triangleq \hat{U}_{(\ell-1)L+1}^{(a \rightarrow b), \ell L}.$$

By assumption,

$$\mathbb{E}[d_L(U^L(\ell), \hat{U}^L(\ell))] \leq \epsilon.$$

Thus

$$\begin{aligned} \epsilon &\geq \mathbb{E}[d_L(U^L(\ell), \hat{U}^L(\ell))] \\ &= \frac{1}{L} \sum_{i=1}^L \mathbb{E}[d(U_i(\ell), \hat{U}_i(\ell))] \\ &\geq \frac{1}{L} \sum_{i=1}^L d_{\min} \mathbb{P}(U_i(\ell) \neq \hat{U}_i(\ell)), \end{aligned} \quad (33)$$

where $d_{\min} \triangleq \min_{(u, \hat{u}) \in \mathcal{U} \times \hat{\mathcal{U}}: u \neq \hat{u}} d(u, \hat{u})$. Since all alphabets are assumed to be finite, and $d(u, \hat{u}) = 0$ if and only if $u = \hat{u}$, $d_{\min} > 0$ by assumption. Therefore,

$$\frac{1}{L} \sum_{i=1}^L \mathbb{P}(U_i(\ell) \neq \hat{U}_i(\ell)) \leq \frac{\epsilon}{d_{\min}}$$

for all $\ell \in \{1, 2, \dots, N\}$.

Recall that all sources and channels are memoryless by assumption and that the same code is used independently on each L -vector. Therefore, $\{U^L(\ell), \hat{U}^L(\ell)\}_{\ell=1}^N$ is an i.i.d. sequence. (See Fig. 5.) Our goal in the argument that follows is to losslessly describe $U^L(1), \dots, U^L(N)$ to a decoder that knows $\hat{U}^L(1), \dots, \hat{U}^L(N)$. We treat this as a problem of lossless source coding with receiver side information, as shown in Fig. 6. From [20], rate

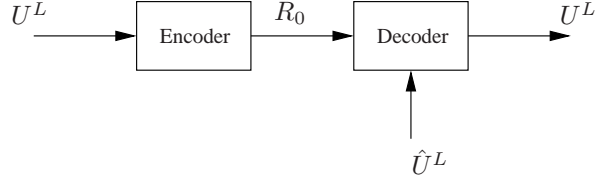


Fig. 6. Slepian-Wolf coding for converting zero-distortion reconstruction into lossless reconstruction.

$R_0^{(k)} = H(U^L|\hat{U}^L)$ suffices for losslessly reconstructing U^L at a receiver that knows \hat{U}^L . Here lossless coding means that the reconstruction \tilde{U}^{LN} at the receiver has an error probability $\mathbb{P}(U^{LN} \neq \tilde{U}^{LN})$ that can be made arbitrarily small, which is precisely the criterion needed for our proof. Therefore, for any $\epsilon < d_{\min}/2$, using Fano's inequality [18], Jensen's inequality, and the concavity of the entropy function, we have

$$\begin{aligned}
R_0^{(k)} &= H(U^L|\hat{U}^L) = \sum_{i=1}^L H(U_i|U^{i-1}, \hat{U}^L) \\
&\leq \sum_{i=1}^L H(U_i|\hat{U}_i) \\
&\leq \sum_{i=1}^L H(U_i, \mathbb{1}_{U_i=\hat{U}_i}|\hat{U}_i) \\
&\leq \sum_{i=1}^L [h(\mathbb{P}(U_i \neq \hat{U}_i)) + \log |\mathcal{U}| \mathbb{P}(U_i \neq \hat{U}_i)] \\
&\leq Lh\left(\frac{1}{L} \sum_{i=1}^L \mathbb{P}(U_i \neq \hat{U}_i)\right) + \log |\mathcal{U}| \sum_{i=1}^L \mathbb{P}(U_i \neq \hat{U}_i) \\
&\leq L\left(h\left(\frac{\epsilon}{d_{\min}}\right) + \frac{\log |\mathcal{U}| \epsilon}{d_{\min}}\right) \\
&\triangleq Lf(\epsilon),
\end{aligned} \tag{34}$$

where for any $0 \leq p \leq 1$, $h(p) = -p \log p - (1-p) \log(1-p)$, and $f(\epsilon) \triangleq h(\frac{\epsilon}{d_{\min}}) + \frac{\log |\mathcal{U}| \epsilon}{d_{\min}}$. Note that $f(\epsilon) \rightarrow 0$ as $\epsilon \rightarrow 0$.

For each $(a, b) \in \mathcal{L}(D)$, we send the rate- $R_0^{(k)}$ description of U^L from node a to node b by treating the random mapping from U^L to \hat{U}^L that results from applying the given code across the given network as a noisy channel. Specifically, we order the source-receiver pairs $(a, b) \in \mathcal{L}(D)$ lexicographically and send the description for the k -th pair (a, b) using N'_k dummy source vectors $U^L(N + \sum_{k'=1}^{k-1} N'_{k'} + 1), \dots, U^L(N + \sum_{k'=1}^{k-1} N'_{k'} + N'_k)$, thereby creating N'_k uses of a channel $p(\hat{u}^L|u^L)$ through which we can reliably transmit the lossless description of $U^L(1), \dots, U^L(N)$ for (a, b) to the decoder. The decoder's distortion- ϵ reconstructions $\hat{U}^L(1), \dots, \hat{U}^L(N)$ of source vectors $U^L(1), \dots, U^L(N)$ are treated as side information known only by the decoder.

The following discussion describes the approach precisely and investigates its performance. The code used to losslessly describe $U^L(1), \dots, U^L(N)$ from node a to the node b employs fixed source values $U^{(v),L}(N +$

$\sum_{k'=1}^{k-1} N'_{k'} + 1) = \dots = U^{(v),L}(N + \sum_{k'=1}^{k-1} N'_{k'} + N'_k) = u^{(v),L}$ for all nodes $v \in \mathcal{V} \setminus a$ in the network. The value transmitted by each node $v \in \mathcal{V} \setminus \{a\}$ is chosen as follows.

Since distortion is non-negative by assumption,

$$\begin{aligned} \epsilon &\geq \mathbb{E}[d_L(U^L, \hat{U}^L)] \\ &\geq \mathbb{E}[d_L(U^L, \hat{U}^L) | U^{(-a),L} \in \mathcal{T}_\delta^{(L)}] \mathbb{P}(U^{(-a),L} \in \mathcal{T}_\delta^{(L)}), \end{aligned}$$

where $U^{(-a),L} \triangleq (U^{(v),L})_{v \in \mathcal{V} \setminus a}$. For any $\delta > 0$ and all L large enough, $\mathbb{P}(U^{(-a),L} \in \mathcal{T}_\delta^{(L)}) > 1 - \delta$, which implies that

$$\mathbb{E}[d_L(U^L, \hat{U}^L) | U^{(-a),L} \in \mathcal{T}_\delta^{(L)}] \leq \frac{\epsilon}{1 - \delta}.$$

Hence, there exists $u^{(-a),L} \in \mathcal{T}_\delta^{(L)}$ such that

$$\mathbb{E}[d_L(U^L, \hat{U}^L) | U^{(-a),L} = u^{(-a),L}] \leq \frac{\epsilon}{1 - \delta}. \quad (35)$$

Fix any such $u^{(-a),L}$. To bound the capacity of the resulting channel, we first bound the conditional entropy of U^L given \hat{U}^L , when $U^{(-a),L} = u^{(-a),L}$. Here, following steps similar to those in (33) and (34), but here conditioning on $U^{(-a),L} = u^{(-a),L}$, we conclude that

$$H(U^L | \hat{U}^L, U^{(-a),L} = u^{(-a),L}) \leq Lf\left(\frac{\epsilon}{1 - \delta}\right).$$

To finish our capacity calculation, we next bound the entropy of U^L given $U^{(-a),L} = u^{(-a),L}$. Since $u^{(-a),L} \in \mathcal{T}_\delta^{(L)}$, for any $u^L \in \mathcal{T}_\delta^{(L)}(U | u^{(-a),L})$,

$$p(u^L | u^{(-a),L}) \leq 2^{-(1-\delta)LH(U|U^{(-a)})}$$

by [16]. Hence, for L large enough,

$$\begin{aligned} &H(U^L | U^{(-a),L} = u^{(-a),L}) \\ &= \sum_{u^L} -p(u^L | u^{(-a),L}) \log p(u^L | u^{(-a),L}) \\ &\geq \sum_{u^L \in \mathcal{T}_\delta^{(L)}(U | u^{(-a),L})} -p(u^L | u^{(-a),L}) \log p(u^L | u^{(-a),L}) \\ &\geq (1 - \delta)LH(U | U^{(-a)}) \mathbb{P}(U^L \in \mathcal{T}_\delta^{(L)} | U^{(-a),L} = u^{(-a),L}) \\ &\geq (1 - \delta)^2 LH(U | U^{(-a)}), \end{aligned}$$

where the last line follows since, for L large enough, $\mathbb{P}(U^L \in \mathcal{T}_\delta^{(L)} | U^{(-a),L} = u^{(-a),L}) > 1 - \delta$.

Hence, fixing $U^{(-a),L} = u^{(-a),L}$ yields a channel $p(\hat{u}^L | u^L, U^{(-a),L} = u^{(-a),L})$, with capacity

$$C_0^{(k)} \geq (1 - \delta)^2 LH(U|U^{(-a)}) - Lf\left(\frac{\epsilon}{1 - \delta}\right). \quad (36)$$

Thus the rate required to losslessly describe U^{LN} to a decoder with reproduction \hat{U}^{LN} of U^{LN} is at most $R_0^{(k)}N$, and the capacity of the channel over which we wish to describe U^{LN} is at least $C_0^{(k)}$ bits per L network uses. We can therefore achieve the desired lossless description provided that $N'_k C_0^{(k)} > NR_0^{(k)}$, giving $N'_k > NR_0^{(k)}/C_0^{(k)}$. Thus the total number of sessions required to send first the lossy description and then the lossless incremental description is

$$N + N' = N + \sum_{k=1}^{|\mathcal{L}(D)|} N'_k > N \left(1 + \sum_{k=1}^{|\mathcal{L}(D)|} R_0^{(k)}/C_0^{(k)} \right).$$

Here

$$\begin{aligned} \frac{R_0^{(k)}}{C_0^{(k)}} &\leq \frac{Lf(\epsilon)}{(1 - \delta)^2 LH(U|U^{(-a)})Lf\left(\frac{\epsilon}{1 - \delta}\right)} \\ &= \frac{f(\epsilon)}{(1 - \delta)^2 H(U|U^{(-a)}) - f\left(\frac{\epsilon}{1 - \delta}\right)}, \end{aligned}$$

which approaches zero as ϵ approaches zero and δ approaches zero. Repeating this process for every $(a, b) \in \mathcal{L}(D)$, the resulting coding rate can be bounded as

$$\frac{\kappa}{1 + \sum_{k=1}^{|\mathcal{L}(D)|} R_0^{(k)}/C_0^{(k)}} \leq \kappa' \leq \kappa.$$

Since $|\mathcal{L}(D)| < |\mathcal{V}|^2$ is a finite number, the resulting coding rate κ' , after adding these extra sessions, still approaches to κ , as ϵ and δ corresponding to each $(a, b) \in \mathcal{L}(D)$ converge to zero. ■

Combining Theorem 3, Theorem 2 and the result proved by W. Gu in [19] proves the separation of source-network coding and channel coding in a wireline network with dependent sources with lossy or lossless reconstructions. In particular, this result partially extends the separation result of [6] to the case where the sources are dependent. The extension is partial since in [6] the channels can be discrete or continuous, but here we have only considered discrete channels. In the next section, we consider the case of AWGN channels.

VII. CONTINUOUS CHANNELS

While the capacity results of [8] are proven for general (discrete or continuous) alphabets, the sources and channels considered in Theorems 1 and 2 were all assumed to have finite alphabets. In this section, we prove that our results also hold for AWGN channels. In order to prove this we use the discretization method introduced in [21].

Consider a wireline network \mathcal{N} with an AWGN channel from node a to node b . Let the input and output of this channel be X and $Y = X + Z$, respectively. The coding on \mathcal{N} is performed similar to the coding described in

Section III. Assume input power constraint P and noise power N . To impose the power constraint, for a code with channel blocklength n , we require

$$\mathbb{E}[X_t^2] \leq P,$$

for $t = 1, \dots, n$. Similarly, in the N -fold stacked version of \mathcal{N} , we require

$$\frac{1}{N} \sum_{\ell=1}^N \mathbb{E}[X_t^2(\ell)] \leq P,$$

for $t = 1, \dots, n$.

Let \mathcal{N}_b be a wireline network that is identical to network \mathcal{N} except that the channel from a to b is replaced by a bit pipe of capacity $C = 0.5 \log(1 + P/N)$. Theorem 4 shows, as in the case of discrete-valued channels, that this change does not affect the set of achievable distortions, thereby generalizing Theorem 2.

Remark 3: Given a Gaussian channel with input power constraint P , usually, a code of blocklength n and rate R is defined as a code with 2^{nR} codewords $(x^n(m))_{m=1}^{2^{nR}}$, such that $\sum_{t=1}^n x_t^2(m) \leq np$, for every $m = 1, \dots, 2^{nR}$ [18], [16]. However, instead of an average power constraint on each codeword, we can put an average power constraint on each transmitted symbol and require that $\mathbb{E}[x_t^2(M)] \leq P$, for $t = 1, \dots, n$ [22]. Note that for a given code, the randomness in $\mathbb{E}[x_t^2(M)] \leq P$ is only due to the message M . This alternative definition does not affect the capacity of the channel from $C = 0.5 \log(1 + P/N)$. In this paper, we consider the latter definition because of some technical issues in the proof of the main result.

The equivalence of the capacities corresponding to the two definitions can be shown as follows. The converse of the capacity theorem stated in [18] applies to the case symbol-by-symbol power constraint as well. For the achievability, consider the code construction presented in [18] with the same encoding and decoding strategy. For each $t = 1, \dots, n$, $\mathbb{P}(\mathbb{E}[x_t^2(M)] > P) = \mathbb{P}(2^{-nR} \sum_{m=1}^{2^{nR}} x_t^2(m) > P) \leq 2^{-2^{nR}\delta(\epsilon)}$, where $\delta(\epsilon) \rightarrow 0$ as $\epsilon \rightarrow 0$. Hence, by the union bound $\mathbb{P}(\mathbb{E}[x_t^2(M)] > P, \text{ for some } t) \leq n2^{-2^{nR}\delta(\epsilon)}$. This shows that there exist a sequence of codes that both satisfy the power constraint on each coordinate and also have arbitrary small probability of error. (The analysis of the probability of error presented in [18] applies here too.)

Theorem 4: For a wireline network consisting of discrete or AWGN point-to-point channels,

$$\mathcal{D}(\kappa, \mathcal{N}_b) = \mathcal{D}(\kappa, \mathcal{N}).$$

Proof of Theorem 4: The second inclusion is immediate since the first part of the proof of Theorem 2 applies equally well for continuous channels case. To prove the first inclusion, we employ the discretization method used in [21]. Let network $\mathcal{N}^{(\mathbf{j}, \mathbf{k})}$, with $\mathbf{j} = (j_1, j_2, \dots, j_n)$ and $\mathbf{k} = (k_1, k_2, \dots, k_n)$, denote the network derived from network \mathcal{N} by replacing the AWGN channel from a to b by the structure shown in Fig. 7. The given channel relies on a pair of quantizers $Q[j]$ and $Q[k]$ parametrized by indices j and k . We allow the quantizer parameters to vary with t , setting $j = j_t$ and $k = k_t$ for each time $t \in \{1, 2, \dots, n\}$. The quantizer $Q[i]$ is defined as follows. For $i \in \{1, 2, \dots\}$, let $\Delta = 1/\sqrt{i}$, and define the quantizer $Q[i]$ with quantization levels $\mathcal{L}_i = \{-i\Delta, -(i-1)\Delta, \dots, -\Delta, 0, \Delta, \dots, (i-1)\Delta, i\Delta\}$. For any $x \in \mathbb{R}$, $Q[i]$ maps x to $[x]_i$, which is the closest

number to x in \mathcal{L}_i such that $|[x]_i| \leq x$. Note that by this definition, $\mathbb{E}[[X]_i^2] \leq \mathbb{E}[X^2]$ for any random variable X .

Lemma 1 in Appendix B shows that as j and k increase, the set of achievable distortions on $\mathcal{N}^{(j,k)}$ approaches the set of achievable distortions on the original network. More precisely,

$$\mathcal{D}_a(\kappa, \mathcal{N}) \subseteq \overline{\limsup_{j,k} \mathcal{D}(\kappa, \mathcal{N}^{(j,k)})}, \quad (37)$$

where

$$\limsup_{j,k} \mathcal{A}_{j,k} \triangleq \bigcap_{j_0, k_0} \bigcup_{\substack{j \geq j_0 \\ k \geq k_0}} \mathcal{A}_{j,k},$$

and $\overline{\mathcal{A}}$ denotes the closure of the set \mathcal{A} .

We next show that

$$\mathcal{D}(\kappa, \mathcal{N}^{(j,k)}) \subseteq \mathcal{D}(\kappa, \mathcal{N}_b). \quad (38)$$

This is sufficient to obtain the desired result since (37) and (38) together imply $\mathcal{D}_a(\kappa, \mathcal{N}) \subseteq \mathcal{D}(\kappa, \mathcal{N}_b)$ by the closure in the definition of $\mathcal{D}(\kappa, \mathcal{N}_b)$.

To prove that $\mathcal{D}(\kappa, \mathcal{N}^{(j,k)}) \subseteq \mathcal{D}(\kappa, \mathcal{N}_b)$, note that, at each time t , the structure shown in Fig. 7 behaves like a DMC with input $[X]_{j_t}$, power constraint $\mathbb{E}([X]_{j_t})^2 \leq P$ and output $[Y_{j_t}]_{k_t}$. Hence, by straightforward extension of the proof of Theorem 2,

$$\mathcal{D}(\kappa, \mathcal{N}^{(j,k)}) \subseteq \mathcal{D}(\kappa, \mathcal{N}_b^{(j,k)}),$$

where $\mathcal{N}_b^{(j,k)}$ is identical to $\mathcal{N}^{(j,k)}$ except that the channel from a to b is replaced by a bit pipe of capacity $C_{j,k}$ equal to the maximum capacity of the n DMCs. Here

$$C_{j,k} \triangleq \max_{1 \leq t \leq n} \max_{\substack{[X]_{j_t} \sim p_X \\ p_X: \mathbb{E}[X^2] \leq P}} I([X]_{j_t}; [Y_{j_t}]_{k_t}).$$

By the data processing inequality [18],

$$\begin{aligned} I([X]_{j_t}; [Y_{j_t}]_{k_t}) &\leq I([X]_{j_t}; Y_{j_t}) \\ &= h(Y_{j_t}) - h(Z). \end{aligned}$$

On the other hand, by the construction of the quantizers,

$$\begin{aligned} \mathbb{E}[Y_{j_t}^2] &= \mathbb{E}([X]_{j_t}^2) + N \\ &\leq \mathbb{E}[X^2] + N. \end{aligned}$$

Hence,

$$h(Y_{j_t}) \leq 0.5 \log(2\pi e(P + N)),$$

and as a result

$$I([X]_{j_t}; [Y_{j_t}]_{k_t}) \leq C.$$

Therefore, $\mathcal{D}(\kappa, \mathcal{N}^{(j,k)}) \subseteq \mathcal{D}_b$.

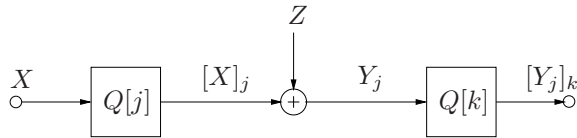


Fig. 7. Quantizing the input and output alphabets of an AWGN

VIII. CONCLUSIONS

In this paper we proved the separation of source-network and channel coding in general wireline networks of independent discrete point-to-point channels with dependent sources and arbitrary lossy or lossless reconstruction demands. We also proved that the result continues to hold when one or more channels is an AWGN channel.

APPENDIX A: PROOF OF PART II OF THEOREM 1

Let $D \in \text{int}(\mathcal{D}_s(\kappa, \underline{\mathcal{N}}))$. Then for any $\epsilon > 0$, there exist integers N , n , and L such that $L/n \geq \kappa - \epsilon$ and there exists a blocklength- n coding scheme for L source symbols on N -fold stacked network $\underline{\mathcal{N}}$ that achieves

$$\mathbb{E} \left[d_{NL}^{(a \rightarrow b)}(U^{(a), NL}, \hat{U}^{(a \rightarrow b), NL}) \right] \leq D(a, b) + \epsilon,$$

for all $a, b \in \mathcal{V}$. The same coding scheme can be used in a single-layer network as follows. Consider a single layer network where each node a observes a length- NL block of source symbols $U^{(a), NL}$ and describes the block in the next Nn time steps. Given source blocklength $L' = NL$ and channel block length $n' = Nn$, the code has rate $\kappa = L'/N' = L/N$. At each time $t \in \{1, \dots, N\}$, each node $a \in \mathcal{V}$ sends, over its outgoing edges, what it would have sent at time 1 in layer t of $\underline{\mathcal{N}}$, i.e., $\mathbf{X}_1^{(a)}(t)$, and collects, over its incoming edges, what it would have collected in layer t of $\underline{\mathcal{N}}$, i.e., $\mathbf{Y}_1^{(a)}(t)$. At times $t \in \{N+1, \dots, 2N\}$, each node a sends $\mathbf{X}_2^{(a)}(t-N)$ and collects $\mathbf{Y}_2^{(a)}(t-N)$. Here calculating $\mathbf{X}_2^{(a)}$ is possible due to the prior collection of $\mathbf{Y}_2^{(a)}$. The same strategy is used in the next $n-2$ time intervals, in interval t transmitting $\mathbf{X}_t^{(a)}$ for $t \in \{3, \dots, n\}$ and collecting $\mathbf{Y}_t^{(a)}$ for uses in future time intervals. Using this strategy, at the end of nN channel uses, each node's observation has exactly the same distribution as the collection of observations of its N copies in the stacked networks. Therefore, applying the decoding rules results in the same distortion. Hence, $D \in \mathcal{D}(\kappa, \mathcal{N})$.

APPENDIX B: LEMMA 1

Lemma 1: For any $\kappa > 0$,

$$\mathcal{D}(\kappa, \mathcal{N}) \subseteq \overline{\limsup_{j,k} \mathcal{D}(\kappa, \mathcal{N}^{(j,k)})}, \quad (\text{B-1})$$

where $\overline{\mathcal{A}}$ denotes the closure of set \mathcal{A} .

Proof: Let $D \in \mathcal{D}(\kappa, \mathcal{N})$. For any $\epsilon > 0$, and for L sufficiently large, there exist a joint source-channel code at rate κ with source blocklength L such that

$$\mathbb{E}[d_L(U^{(a),L}, \hat{U}^{(a \rightarrow b),L})] \leq D(a, b) + \epsilon, \quad (\text{B-2})$$

holds for each $(a, b) \in \mathcal{V}^2$. Let $U^L = U^{(a),L}$ and $\hat{U}^L = \hat{U}^{(a \rightarrow b),L}$ for some fixed $(a, b) \in \mathcal{V}^2$.

Conditioning the expected average distortion between U^L and \hat{U}^L on the input and output values of the AWGN channel at time $t = 1$, it follows that

$$\begin{aligned} D(a, b) + \epsilon &\geq \mathbb{E}[d_L(U^L, \hat{U}^L)] \\ &= \sum_{(x_1, y_1)} p(x_1, y_1) \mathbb{E}[d_L(U^L, \hat{U}^L) | (X_1, Y_1) = (x_1, y_1)] \\ &= \mathbb{E}[\delta^{(1)}(X_1, Y_1)] \end{aligned} \quad (\text{B-3})$$

where $\delta^{(1)}(x_1, y_1) \triangleq \mathbb{E}[d_L(U^L, \hat{U}^L) | (X_1, Y_1) = (x_1, y_1)]$.

Now assume that the same code is applied to network $\mathcal{N}^{(j_1, k_1)}$, which is identical to \mathcal{N} except that at time $t = 1$, the AWGN channel is replaced by the structure shown in Fig. 7 with parameters $j = j_1$ and $k = k_1$. The expected average distortion between U^L and \hat{U}^L in the modified network, $D^{(j_1, k_1)}(a, b)$, can be written as

$$D^{(j_1, k_1)}(a, b) = \mathbb{E}[\delta^{(1)}(X_1, \tilde{Y}_1)], \quad (\text{B-4})$$

where $\tilde{Y}_1 \triangleq [[X]_{j_1} + Z_1]_{k_1}$. Note that, conditioned on the input and output values of the AWGN channel at time $t = 1$, the two networks have identical performance.

Further, \tilde{Y}_1 converges pointwise to Y_1 almost everywhere as j and k_1 grow without bound, i.e.,

$$\lim_{k_1 \rightarrow \infty} \lim_{j_1 \rightarrow \infty} \tilde{Y}_1 = Y_1, \quad (\text{B-5})$$

almost everywhere, where $Y_1 = X + Z$.

While function $\delta^{(1)}(x_1, y_1)$ might not be continuous everywhere, by the Lusin's Theorem [23], since it is measurable, for any $\epsilon_1 > 0$, there exists a subset $\mathcal{A}_1 \subset \mathbb{R}^2$, such that $\mathbb{P}((X_1, Y_1) \in \mathcal{A}_1) < \epsilon_1$ and $\delta^{(1)}$ is

continuous on \mathcal{A}_1 . By the law of iterated expectations,

$$\begin{aligned} D^{(j_1, k_1)}(a, b) &= \mathbb{E}[\delta^{(1)}(X_1, \tilde{Y}_1) | (X_1, Y_1) \in \mathcal{A}_1] \mathbb{P}((X_1, Y_1) \in \mathcal{A}_1) \\ &\quad + \mathbb{E}[\delta^{(1)}(X_1, \tilde{Y}_1) | (X_1, Y_1) \notin \mathcal{A}_1] \mathbb{P}((X_1, Y_1) \notin \mathcal{A}_1) \\ &\leq \mathbb{E}[\delta^{(1)}(X_1, \tilde{Y}_1) | (X_1, Y_1) \in \mathcal{A}_1] + d_{\max} \epsilon_1, \end{aligned} \quad (\text{B-6})$$

and, similarly,

$$\begin{aligned} D^{(j_1, k_1)}(a, b) &\geq \mathbb{E}[\delta^{(1)}(X_1, \tilde{Y}_1) | (X_1, Y_1) \in \mathcal{A}_1] \mathbb{P}((X_1, Y_1) \in \mathcal{A}_1) \\ &\geq \mathbb{E}[\delta^{(1)}(X_1, \tilde{Y}_1) | (X_1, Y_1) \in \mathcal{A}_1] (1 - \epsilon_1). \end{aligned} \quad (\text{B-7})$$

Since $\delta(x_1, y_1)$ is continuous on \mathcal{A}_1 and is bounded, by the bounded convergence theorem, it follows from (B-6) and (B-7) that

$$\begin{aligned} \lim_{k_1 \rightarrow \infty} \lim_{j_1 \rightarrow \infty} D^{(j_1, k_1)}(a, b) &\leq \lim_{k_1 \rightarrow \infty} \lim_{j_1 \rightarrow \infty} \mathbb{E}[\delta^{(1)}(X_1, \tilde{Y}_1) | (X_1, Y_1) \in \mathcal{A}_1] + d_{\max} \epsilon_1 \\ &= \mathbb{E}[\lim_{k_1 \rightarrow \infty} \lim_{j_1 \rightarrow \infty} \delta^{(1)}(X_1, \tilde{Y}_1) | (X_1, Y_1) \in \mathcal{A}_1] + d_{\max} \epsilon_1 \\ &\leq \mathbb{E}[\delta^{(1)}(X_1, Y_1) | (X_1, Y_1) \in \mathcal{A}_1] + d_{\max} \epsilon_1, \end{aligned} \quad (\text{B-8})$$

and

$$\lim_{k_1 \rightarrow \infty} \lim_{j_1 \rightarrow \infty} D^{(j_1, k_1)}(a, b) \geq \mathbb{E}[\delta^{(1)}(X_1, Y_1) | (X_1, Y_1) \in \mathcal{A}_1] (1 - \epsilon_1). \quad (\text{B-9})$$

On the other hand,

$$\begin{aligned} \mathbb{E}[\delta^{(1)}(X_1, Y_1)] &\geq \mathbb{E}[\delta^{(1)}(X_1, Y_1) | (X_1, Y_1) \in \mathcal{A}_1] \mathbb{P}((X_1, Y_1) \in \mathcal{A}_1) \\ &\geq \mathbb{E}[\delta^{(1)}(X_1, Y_1) | (X_1, Y_1) \in \mathcal{A}_1] (1 - \epsilon_1), \end{aligned} \quad (\text{B-10})$$

and

$$\mathbb{E}[\delta^{(1)}(X_1, Y_1)] \leq \mathbb{E}[\delta^{(1)}(X_1, Y_1) | (X_1, Y_1) \in \mathcal{A}_1] + d_{\max} \epsilon_1. \quad (\text{B-11})$$

Since $\mathbb{E}[\delta^{(1)}(X_1, Y_1)] = \mathbb{E}[d_L(U^L, \hat{U}^L)]$, combining (B-8) and (B-9) with (B-10) and (B-11) yields

$$\mathbb{E}[d_L(U^L, \hat{U}^L)] - d_{\max} \epsilon_1 \leq \lim_{k_1 \rightarrow \infty} \lim_{j_1 \rightarrow \infty} D^{(j_1, k_1)}(a, b) \leq \frac{\mathbb{E}[d_L(U^L, \hat{U}^L)]}{1 - \epsilon_1} + d_{\max} \epsilon_1. \quad (\text{B-12})$$

Since ϵ_1 can be made arbitrary small, from (B-12), we have

$$\lim_{k_1 \rightarrow \infty} \lim_{j_1 \rightarrow \infty} D^{(j_1, k_1)}(a, b) = \mathbb{E}[d_L(U^L, \hat{U}^L)]. \quad (\text{B-13})$$

The prior analysis captures the expected distortion when the continuous channel is replaced by a finite alphabet channel only at time 1. To finish the proof we use induction. Assume that for times $1, 2, \dots, t-1$, the continuous

channel can be replaced a finite alphabet channel without, asymptotically, changing the expected average distortion, i.e.,

$$\begin{aligned}
& \lim_{k_1 \rightarrow \infty} \lim_{j_1 \rightarrow \infty} \dots \lim_{k_{t-1} \rightarrow \infty} \lim_{j_{t-1} \rightarrow \infty} D^{(j^{t-1}, k^{t-1})}(a, b) \\
&= \mathbb{E}[\delta^{(t-1)}(X^{t-1}, Y^{t-1})], \\
&= \mathbb{E}[d_L(U^L, \hat{U}^L)].
\end{aligned} \tag{B-14}$$

where $D^{(j^{t-1}, k^{t-1})}(a, b)$ denotes the expected average distortion between U^L and \hat{U}^L in the modified network when the parameters of the channel input and output quantizers, at times $\tau \in \{1, \dots, t-1\}$, are (j^{t-1}, k^{t-1}) , and let

$$\delta^{(t-1)}(x^{t-1}, y^{t-1}) \triangleq \mathbb{E}[d_L(U^L, \hat{U}^L) | (X^{t-1}, Y^{t-1}) = (x^{t-1}, y^{t-1})].$$

Now we need to show that if we add the quantizers at time t as well, the performance does not change.

In the original network

$$\begin{aligned}
\mathbb{E}[d_L(U^L, \hat{U}^L)] &= \mathbb{E}[\delta^{(t)}(X^t, Y^t)] \\
&= \mathbb{E} \left[\mathbb{E}[\delta^{(t)}(X^t, Y^t) | (X^t, Y^t)] \right],
\end{aligned} \tag{B-15}$$

and in the modified network,

$$\begin{aligned}
D^{(j^t, k^t)}(a, b) &= \mathbb{E}[\delta^{(t)}(\tilde{X}^t, \tilde{Y}^t)], \\
&= \mathbb{E} \left[\mathbb{E} \left[\delta^{(t)}(\tilde{X}^t, \tilde{Y}^t) \middle| (\tilde{X}^{t-1}, \tilde{Y}^{t-1}) \right] \right],
\end{aligned} \tag{B-16}$$

where for $t' \in \{1, \dots, t\}$, $\tilde{X}_{t'}$ is the channel input at time t' when the given code is applied and the Gaussian channel replaced by its quantized approximation, and $\tilde{Y}_{t'} = [[\tilde{X}_{t'}]_{j_{t'}} + Z_{t'}]_{k_{t'}}$. Note that $\tilde{X}_1 = X_1$.

While X_t and \tilde{X}_t might have different distributions due to the quantizations at times $t' = 1, \dots, t-1$, their conditional distributions given the inputs and outputs of the channel up to time $t-1$ are identical in both networks, i.e.,

$$\begin{aligned}
& \mathbb{P} \left(\tilde{X}_t < x_t \middle| (\tilde{X}^{t-1}, \tilde{Y}^{t-1}) = (x^{t-1}, y^{t-1}) \right) \\
&= \mathbb{P} \left(X_t < x_t \middle| (X^{t-1}, Y^{t-1}) = (x^{t-1}, y^{t-1}) \right).
\end{aligned} \tag{B-17}$$

Let

$$\begin{aligned}
\gamma(x^{t-1}, y^{t-1}) &\triangleq \mathbb{E}[\delta^{(t)}(X^t, Y^t) | (X^{t-1}, Y^{t-1}) = (x^{t-1}, y^{t-1})] \\
&\mathbb{E} \left[\int \delta^{(t)}(x^t, y^{t-1}, Y_t) dF(x_t | (x^{t-1}, y^{t-1})) \right]
\end{aligned} \tag{B-18}$$

and

$$\begin{aligned} \tilde{\gamma}^{(j_t, k_t)}(x^{t-1}, y^{t-1}) &\triangleq \mathbb{E}[\delta^{(t)}(\tilde{X}^t, \tilde{Y}^t) | (\tilde{X}^{t-1}, \tilde{Y}^{t-1}) = (x^{t-1}, y^{t-1})] \\ &= \mathbb{E} \left[\int \delta^{(t)}(x^t, y^{t-1}, \tilde{Y}_t) dF(x_t | (x^{t-1}, y^{t-1})) \right], \end{aligned} \quad (\text{B-19})$$

where in the last line we are using (B-17).

Using the same argument as the one used to prove (B-13), it follows that

$$\lim_{k_t \rightarrow \infty} \lim_{j_t \rightarrow \infty} \tilde{\gamma}^{(j_t, k_t)}(x^{t-1}, y^{t-1}) = \gamma(x^{t-1}, y^{t-1}). \quad (\text{B-20})$$

Hence,

$$\begin{aligned} &\lim_{k_1 \rightarrow \infty} \lim_{j_1 \rightarrow \infty} \dots \lim_{k_t \rightarrow \infty} \lim_{j_t \rightarrow \infty} D^{(j^t, k^t)}(a, b) \\ &= \lim_{k_1 \rightarrow \infty} \lim_{j_1 \rightarrow \infty} \dots \lim_{k_t \rightarrow \infty} \lim_{j_t \rightarrow \infty} \mathbb{E}[\tilde{\gamma}^{(j_t, k_t)}(\tilde{X}^{t-1}, \tilde{Y}^{t-1})] \\ &= \lim_{k_1 \rightarrow \infty} \lim_{j_1 \rightarrow \infty} \dots \lim_{k_{t-1} \rightarrow \infty} \lim_{j_{t-1} \rightarrow \infty} \mathbb{E} \left[\lim_{k_t \rightarrow \infty} \lim_{j_t \rightarrow \infty} \tilde{\gamma}^{(j_t, k_t)}(\tilde{X}^{t-1}, \tilde{Y}^{t-1}) \right] \\ &\stackrel{(a)}{=} \lim_{k_1 \rightarrow \infty} \lim_{j_1 \rightarrow \infty} \dots \lim_{k_{t-1} \rightarrow \infty} \lim_{j_{t-1} \rightarrow \infty} \mathbb{E} \left[\gamma(\tilde{X}^{t-1}, \tilde{Y}^{t-1}) \right] \\ &\stackrel{(b)}{=} \mathbb{E}[d_L(U^{(a), L}, \hat{U}^{(a \rightarrow b), L})], \end{aligned} \quad (\text{B-21})$$

where (a) follows from (B-20) plus the dominated convergence theorem, and (b) follows from our inductive hypothesis. ■

ACKNOWLEDGMENTS

This work was supported in part by Caltech's Center for the Mathematics of Information (CMI) and DARPA ITMANET grant W911NF-07-1-0029.

REFERENCES

- [1] C. E. Shannon. A mathematical theory of communication: Parts I and II. *Bell Syst. Tech. J.*, 27:379–423 and 623–656, 1948.
- [2] S. Vembu, S. Verdú, and Y. Steinberg. The source-channel separation theorem revisited. *IEEE Trans. Inform. Theory*, 41(1):44–54, January 1995.
- [3] T. Cover, A.E. Gamal, and M. Salehi. Multiple access channels with arbitrarily correlated sources. *Information Theory, IEEE Transactions on*, 26(6):648 – 657, November 1980.
- [4] M. Effros, M. Médard, T. Ho, S. Ray, D. R. Karger, and R. Koetter. A unified framework for source, channel and network coding. In *DIMACS Workshop on Network Information Theory*, Piscataway, NJ, March 2003.
- [5] A. Ramamoorthy, K. Jain, P.A. Chou, and M. Effros. Separating distributed source coding from network coding. *Information Theory, IEEE Transactions on*, 52(6):2785 – 2795, June 2006.
- [6] R. Koetter, M. Effros, and M. Médard. On the theory of network equivalence. In *IEEE Inform. Theory Workshop (ITW)*, 2009.
- [7] R. Koetter, M. Effros, and M. Médard. A theory of network equivalence, parts I and II. *arXiv:1007.1033v2*, 2010.
- [8] R. Koetter, M. Effros, and M. Médard. A theory of network equivalence - part i: Point-to-point channels. *IEEE Trans. Inform. Theory*, 57(2):972–995, 2011.

- [9] S.P. Borade. Network information flow: limits and achievability. In *Information Theory, 2002. Proceedings. 2002 IEEE International Symposium on*, 2002.
- [10] L. Song, R.W. Yeung, and N. Cai. A separation theorem for single-source network coding. *IEEE Trans. Inform. Theory*, 52(5):1861–1871, May 2006.
- [11] R. Yeung. Separation principles for multi-user communications. Unpublished work.
- [12] S. Jalali and M. Effros. On the separation of lossy source-network coding and channel coding in wireline networks. In *Proc. IEEE Int. Symp. Inform. Theory*, pages 500–504, June 2010.
- [13] Chao Tian, Jun Chen, S.N. Diggavi, and S. Shamai. Optimality and approximate optimality of source-channel separation in networks. In *Proc. IEEE Int. Symp. Inform. Theory*, pages 495–499, June 2010.
- [14] A. Orlitsky and J.R. Roche. Coding for computing. *Information Theory, IEEE Transactions on*, 47(3):903–917, March 2001.
- [15] P.W. Cuff, H.H. Permuter, and T.M. Cover. Coordination capacity. *Information Theory, IEEE Transactions on*, 56(9):4181–4206, Sept. 2010.
- [16] A. El Gamal and Y. Kim. *Network Information Theory*. Cambridge University Press, 2011.
- [17] R. Durrett. *Probability: Theory and Examples*. Duxbury Press, Belmont, CA, 1996.
- [18] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley, New York, 2nd edition, 2006.
- [19] W. Gu. *Achievable Rate Regions for Source Coding over Networks*. PhD thesis, California Institute of Technology, Pasadena, CA, 2008.
- [20] D. Slepian and J. Wolf. Noiseless coding of correlated information sources. *IEEE Trans. Inform. Theory*, 19(4):471–480, 1973.
- [21] R. J. McEliece. *The Theory of Information and Coding*. Addison-Wesley, Reading, MA, 1977.
- [22] S. Shamai and I. Bar-David. The capacity of average and peak-power-limited quadrature gaussian channels. *Information Theory, IEEE Transactions on*, 41(4):1060–1071, Jul. 1995.
- [23] A.S. Kechris and A.S. Kechris. *Classical descriptive set theory*, volume 1. Springer-Verlag New York, 1995.
- [24] B. R. Gelbaum and J. M. H. Olmsted. *Counterexamples in Analysis*. San Francisco: Holden-Day, 1964.
- [25] John M. H. Olmsted. *Real Variables*. Appleton-Century-Crofts, Inc., New York, 1956.