A Few Meta-Theorems in Network Information Theory

Yu Xiang and Young-Han Kim
Department of Electrical and Computer Engineering
University of California, San Diego
La Jolla, CA 92093, USA
Email: {yxiang,yhk}@ucsd.edu

***Abstract*—This paper reviews the relationship among several notions of capacity regions of a general discrete memoryless network under different code classes and performance criteria, such as average vs. maximal or block vs. bit error probabilities and deterministic vs. randomized codes. Applications of these meta-theorems include several structural results on capacity regions and a simple proof of the network equivalence theorem.**

I. Introduction

Consider a noisy network communication system with N nodes, where node $k \in [1 : N] := \{1, \ldots, N\}$ wishes to reliably communicate a message M_k at a rate R_k bits per transmission to a set of destination nodes $\mathcal{D}_k \subseteq [1 : N]$ over the noisy network. This system can be modeled as a multimessage discrete memoryless network (DMN) $\mathcal{N} = (\mathcal{X}_1 \times \cdots \times \mathcal{X}_N, p(y_1, \ldots, y_N|x_1, \ldots, x_N), \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_N)$ that consists of N sender–receiver alphabet pairs $(\mathcal{X}_k, \mathcal{Y}_k), k \in [1 : N]$, and a collection of probability mass functions (pmfs) $p(y_1, \ldots, y_N|x_1, \ldots, x_N)$. The network is memoryless in the sense that

$$p(y_{1i}, \ldots, y_{Ni}|x_1^i, \ldots, x_N^i, y_1^{i-1}, \ldots, y_N^{i-1}, m) = p_{Y^N|X^N}(y_{1i}, \ldots, y_{Ni}|x_{1i}, \ldots, x_{Ni}), \quad i \in [1 : n].$$

A $(2^{nR_1}, \ldots, 2^{nR_N}, n)$ code for the DMN consists of

- N message sets $[1 : 2^{nR_1}], \ldots, [1 : 2^{nR_N}]$,
- a set of encoders $\phi_{ki} : [1 : 2^{nR_k}] \times \mathcal{Y}_k^{i-1} \to \mathcal{X}_{ki}$ for $i \in [1 : n]$ and $k \in [1 : N]$, and
- a set of decoders $\psi_d : [1 : 2^{nR_k}] \times \mathcal{Y}_d^n \to \bigcup_{k:d\in\mathcal{D}_k}[1 : 2^{nR_k}] \cup \{\mathrm{e}\}$ for $d \in \bigcup_k \mathcal{D}_k$.

Assume that $(M_1, \ldots, M_N)$ is uniformly distributed over $[1 : 2^{nR_1}] \times \cdots \times [1 : 2^{nR_N}]$. The average probability of error is defined as

$$P_e^{(n)} = \mathsf{P}\{\hat{M}_{kd} \neq M_k \text{ for some } k \in [1 : N], d \in \mathcal{D}_k\}.$$

A rate tuple $(R_1, \ldots, R_N)$ is said to be achievable if there exists a sequence of $(2^{nR_1}, \ldots, 2^{nR_N}, n)$ codes such that $\lim_{n\to\infty} P_e^{(n)} = 0$. The capacity region $\mathscr{C}(\mathcal{N})$ of the DMN is the closure of the set of achievable rates.

Note that the above definition of capacity region follows the standard criteria such as *deterministic* encoders and decoders, and *average* probability of error of the entire message *block*. In the following, we recall alternative criteria in defining the capacity region (cf. [1], [2]).

Deterministic vs. randomized code. The deterministic encoders and decoders can be replaced by *randomized* encoders and decoders as $\Phi_{ki}(\cdot) = \phi_{ki}(\cdot, W_k)$ and $\Psi_d(\cdot) = \psi_d(\cdot, W_d)$, where $(W_1, \ldots, W_N)$ are independent of $(M_1, \ldots, M_N)$ and

$$p(y_{1i}, \ldots, y_{Ni}|x_1^i, \ldots, x_N^i, y_1^{i-1}, \ldots, y_N^{i-1}, m, w^N) = p_{Y^N|X^N}(y_{1i}, \ldots, y_{Ni}|x_{1i}, \ldots, x_{Ni}).$$

If $W_1, \ldots, W_N$ are independent and identically distributed (i.i.d.) $\mathrm{Unif}[0, 1]$, then the corresponding capacity region is denoted by $\mathscr{C}_{\mathrm{rand}}(\mathcal{N})$. If $W_k \equiv W \sim \mathrm{Unif}[0, 1]$, $k \in [1 : N]$, which induces more general cooperation among the nodes, then the corresponding capacity region is denoted by $\mathscr{C}_{\mathrm{cr}}(\mathcal{N})$.

Average vs. maximal probability of error. If the average probability of error is replaced by the *maximal* probability of error

$$P_{e,\max}^{(n)} = \max_{m_1,\ldots,m_N} \mathsf{P}\{\hat{M}_{kd} \neq m_k \text{ for some } k, d \mid M_1 = m_1, \ldots, M_N = m_N\},$$

then the corresponding capacity region is denoted by $\mathscr{C}_{\max}(\mathcal{N})$.

Block vs. bit error probability. Suppose that we represent each message $M_k = (S_{k1}, \ldots, S_{k,nR_k})$ as a sequence of nR_k random bits. If the block error probability is replaced by the average *bit* error probability

$$P_{e,\mathrm{bit}}^{(n)} = \max_{k,d} \frac{1}{nR_k} \sum_{\nu=1}^{nR_k} \mathsf{P}\{\hat{S}_{kd\nu} \neq S_{k\nu}\},$$

then the corresponding capacity region is denoted by $\mathscr{C}_{\mathrm{bit}}(\mathcal{N})$.

These alternatives can be combined into $12 = 3 \times 2 \times 2$ different notions of capacity region. We denote them as $\mathscr{C}_{\alpha,\beta,\gamma}(\mathcal{N})$, where

$$\alpha = \mathrm{det/rand/cr}, \quad \beta = \mathrm{avg/max}, \quad \gamma = \mathrm{blk/bit}.$$

For example, the standard capacity region with deterministic code and average block error probability can be denoted as $\mathscr{C}(\mathcal{N}) = \mathscr{C}_{\mathrm{det,avg,blk}}(\mathcal{N})$.

In this paper, we collect several meta-theorems regarding these notions of capacity regions, such as "when and how

978-1-4799-5999-0/14/$31.00 ©2014 IEEE

much randomization helps?" and "whether the bit error probability criterion is much easier to satisfy." These results are not necessarily new, but have been scattered in the literature sometimes in implicit forms. Using these results, we clarify the relationship among the twelve notions of capacity regions. As a more significant application, we revisit the network equivalence theorem by Koetter, Effros, and Médard [3] that roughly states that the capacity region of a network with orthogonal noisy links depends only on the capacities of these links, and provide a simple proof that uses one of the structural results, $\mathscr{C}(\mathcal{N}) = \mathscr{C}_{\mathrm{cr,avg,bit}}(\mathcal{N})$, the network-stacking technique [3], and the universal channel simulation lemma [4].

Throughout the paper, we mostly follow the notation in [5]. In particular, a random variable is denoted by an uppercase letter (e.g., X, Y, Z) and its realization is denoted by a lowercase letter (e.g., x, y, z). The shorthand notation X^n is used to denote the tuple of random variables $(X_1, \ldots, X_n)$, and x^n is used to denote their realizations. We measure the difference between two probability measures P and Q on $\mathcal{X}$ with pmfs $p(x)$ and $q(x)$ by the total variation distance

$$\begin{aligned} d_{\mathrm{TV}}(\mathsf{P}, \mathsf{Q}) &= d_{\mathrm{TV}}(p(x), q(x)) \\ &= \max_{\mathcal{A} \subseteq \mathcal{X}} |\,\mathsf{P}(\mathcal{A}) - \mathsf{Q}(\mathcal{A})| \\ &= \frac{1}{2} \sum_{x \in \mathcal{X}} |p(x) - q(x)| = \frac{1}{2} \|p(x) - q(x)\|_1. \end{aligned}$$

II. Meta-Theorems on the Twelve Capacity Regions

In this section, we state several meta-theorems on the capacity regions defined in Section I and ultimately establish a general relationship among all twelve of them. The following is straightforward.

Lemma 1:

$$\mathscr{C}_{\mathrm{det,avg},\gamma}(\mathcal{N}) \subseteq \mathscr{C}_{\mathrm{rand,avg},\gamma}(\mathcal{N}) \subseteq \mathscr{C}_{\mathrm{cr,avg},\gamma}(\mathcal{N}), \quad (1)$$

$$\mathscr{C}_{\alpha,\mathrm{max},\gamma}(\mathcal{N}) \subseteq \mathscr{C}_{\alpha,\mathrm{avg},\gamma}, \quad (2)$$

$$\mathscr{C}_{\alpha,\beta,\mathrm{blk}} \subseteq \mathscr{C}_{\alpha,\beta,\mathrm{bit}} \quad (3)$$

for all $\alpha = \mathrm{det/rand/cr}$, $\beta = \mathrm{avg/max}$, and $\gamma = \mathrm{blk/bit}$.

Since the (common) randomness is independent of the messages and the channel, it can be easily shown (cf. [5, Prob. 3.6]) that its availability does not increase the capacity region under the average error probability criterion.

Lemma 2 (Randomization does not help on average):

$$\mathscr{C}_{\mathrm{det,avg},\gamma}(\mathcal{N}) = \mathscr{C}_{\mathrm{rand,avg},\gamma}(\mathcal{N}) = \mathscr{C}_{\mathrm{cr,avg},\gamma}(\mathcal{N}). \quad (4)$$

Dueck [6] showed that the capacity region under the maximal error probability criterion can be strictly smaller than that under the average error probability criterion via the counterexample of the binary erasure multiple access channel, that is,

$$\mathscr{C}_{\mathrm{det,max},\gamma}(\mathcal{N}) \subsetneq \mathscr{C}_{\mathrm{det,avg},\gamma}(\mathcal{N}). \quad (5)$$

Noting that the two capacity regions coincide at each of the N axes (for example, $R_2 = \cdots = R_N = 0$) and using time sharing, we can establish the following.

Lemma 3 (Cost of worst-case error):

$$\begin{aligned} \mathscr{C}_{\mathrm{det,avg},\gamma}(\mathcal{N}) &\subseteq N \mathscr{C}_{\mathrm{det,max},\gamma}(\mathcal{N}) \\ &= \{NR^n \colon R^n \in \mathscr{C}_{\mathrm{det,max},\gamma}(\mathcal{N})\}. \end{aligned} \quad (6)$$

It is clear that common randomness can be exploited to construct a good code under the maximal error probability criterion from a good code under the average error probability criterion; each pair of encoder and decoder can dither its message using the common randomness (that is, send $M + W \mod 2^{nR}$, where W is uniformly distributed and independent of M). One important implication of Lemma 3 is that local randomness can achieve the same goal by first transmitting the dither W with a good code under the maximal error probability criterion and then using it as the common randomness [1, Problem 14.5]. Since the same common randomness can be used repeatedly, the rate loss is asymptotically negligible. This argument establishes the following.

Lemma 4 (Local randomness transforms worst to average):

$$\mathscr{C}_{\mathrm{rand,max},\gamma}(\mathcal{N}) = \mathscr{C}_{\mathrm{rand,avg},\gamma}(\mathcal{N}). \quad (7)$$

Note that Lemmas 2 and 4 imply that

$$\mathscr{C}_{\mathrm{rand,max},\gamma}(\mathcal{N}) = \mathscr{C}_{\mathrm{cr,max},\gamma}(\mathcal{N}), \quad (8)$$

since $\mathscr{C}_{\mathrm{rand,avg},\gamma}(\mathcal{N}) \subseteq \mathscr{C}_{\mathrm{cr,max},\gamma}(\mathcal{N}) \subseteq \mathscr{C}_{\mathrm{cr,avg},\gamma}(\mathcal{N})$.

Now we turn our attention to the bit vs. block error probability criteria. By concatenating a good (inner) code under the bit error probability criterion with a capacity-achieving (outer) code for the binary symmetric channel with exponentially small block error probability (cf. [7]), we can establish the following (the proof of which will be given elsewhere).

Lemma 5 (Bits are no cleaner on average than as a whole):

$$\mathscr{C}_{\mathrm{det},\beta,\mathrm{bit}}(\mathcal{N}) = \mathscr{C}_{\mathrm{det},\beta,\mathrm{blk}}(\mathcal{N}). \quad (9)$$

Combined with Lemmas 2 and 4, this implies more generally that

$$\mathscr{C}_{\alpha,\beta,\mathrm{bit}}(\mathcal{N}) = \mathscr{C}_{\alpha,\beta,\mathrm{blk}}(\mathcal{N}). \quad (10)$$

We can summarize the relationship among the twelve capacity regions as follows (see also Fig. 1):

Theorem 1: If $\alpha \neq \mathrm{det}$,

$$\mathscr{C}_{\alpha,\beta,\gamma}(\mathcal{N}) = \mathscr{C}_{\alpha,\beta,\gamma}(\mathcal{N}). \quad (11)$$

Moreover, in general,

$$\mathscr{C}_{\mathrm{det,max},\gamma}(\mathcal{N}) \subsetneq \mathscr{C}_{\mathrm{det,avg},\gamma}(\mathcal{N}) \subseteq N \mathscr{C}_{\mathrm{det,max},\gamma}(\mathcal{N}). \quad (12)$$

Remark 1: While our multimessage DMN model does not include broadcast, namely, multiple messages communicated by a single source to different destination nodes, a similar structural result can be easily established. Note that when there is only a single source node (with multiple messages to be multicast and broadcast), then all twelve capacity regions are identical. This follows by a simple modification of the technique by Willems [8] (see also [1, Problem 14.13] and [5, Problem 8.11]) for single-hop broadcast channels.

$$\begin{array}{ccccccccc}
\mathscr{C}_{\mathrm{det,max,blk}}(\mathcal{N}) & \overset{(5)}{\subsetneq} & \mathscr{C}_{\mathrm{det,avg,blk}}(\mathcal{N}) & \overset{(4)}{=} & \mathscr{C}_{\mathrm{cr,avg,blk}}(\mathcal{N}) & \overset{(4)}{=} & \mathscr{C}_{\mathrm{rand,avg,blk}}(\mathcal{N}) & \overset{(7)}{=} & \mathscr{C}_{\mathrm{rand,max,blk}}(\mathcal{N}) & \overset{(8)}{=} & \mathscr{C}_{\mathrm{cr,max,blk}}(\mathcal{N}) \\
\| \text{ (9)} & & \| \text{ (9)} & & \| \text{ (10)} & & \| \text{ (10)} & & \| \text{ (10)} & & \| \text{ (10)} \\
\mathscr{C}_{\mathrm{det,max,bit}}(\mathcal{N}) & \overset{(5)}{\subsetneq} & \mathscr{C}_{\mathrm{det,avg,bit}}(\mathcal{N}) & \overset{(4)}{=} & \mathscr{C}_{\mathrm{cr,max,bit}}(\mathcal{N}) & \overset{(4)}{=} & \mathscr{C}_{\mathrm{rand,avg,bit}}(\mathcal{N}) & \overset{(7)}{=} & \mathscr{C}_{\mathrm{rand,max,bit}}(\mathcal{N}) & \overset{(8)}{=} & \mathscr{C}_{\mathrm{cr,max,bit}}(\mathcal{N})
\end{array}$$

Fig. 1. The relationship among the twelve capacity regions. Here the numbers on top of binary relations refer to the corresponding equation numbers in Section II.

III. Strong Channel Simulation and a Proof of the Network Equivalence Theorem

As an application of the meta-theorems in Section II and their corollaries, we provide a simple proof of the network equivalence theorem [3]. This theorem by Koetter, Effros, and Médard, which states that the capacity region (defined *operationally* as in Section I) of a network of orthogonal discrete memoryless channels (DMCs) of nonzero capacities stays the same when the DMCs are replaced by noiseless links with matching capacities, is significant in that it holds even when the capacity region itself does not have a computable characterization (cf. earlier results on equivalence of multicast networks [9], [10]). One can interpret the network equivalence theorem as a manifestation of the optimality of the separation between channel coding and network coding, and this view has spurred even stronger separation theorems for source, channel, and network coding; see, for example, [11], [12].

For simplicity of the presentation, we consider the problem of *unicasting* $M = M_1$ from node 1 to node N (i.e., $R_2 = \cdots = R_N = 0$, $\mathcal{D}_1 = N$) and focus on the corresponding capacity $C(\mathcal{N})$, namely, the maximum achievable rate of M. The DMN of our interest has the form

$$p(y_1, y_2, (y_3, y), y_4, \ldots, y_N | x_1, (x_2, x), x_3, x_4, \ldots, x_N) = p(y^N|x^N)p(y|x), \quad (13)$$

where $p(y|x)$ is an orthogonal DMC from node 2 to node 3; see Fig. 2. Here (X_2, X) is the channel input at node 2 and (Y_3, Y) is the channel output at node 3. Therefore, the encoder at node 2 is

$$(x_{2i}, x_i)(y_2^{i-1}), \quad i \in [1:n],$$

and the encoder at node 3 is

$$x_{3i}(y_3^{i-1}, y^{i-1}), \quad i \in [1:n].$$

The operations at other nodes are as before (with $M_4 = \cdots = M_N = \emptyset$). We assume throughout this section that a DMN follows the structure in (13).

We are now ready to state a version of the network equivalence theorem [3], which captures the essence of the result by showing the effect of the (capacity of the) orthogonal DMC on the network capacity.

Theorem 2: Let $\mathcal{N}_1 = q_0(y^N|x^N)q_1(y|x)$ and $\mathcal{N}_2 = q_0(y^N|x^N)q_2(\tilde{y}|\tilde{x})$ be two DMNs, where $q_1(y|x)$ and $q_2(\tilde{y}|\tilde{x})$ are DMCs (on potentially different alphabet pairs) with capacities $C_1 < C_2$. Then,

$$C(\mathcal{N}_1) \le C(\mathcal{N}_2).$$

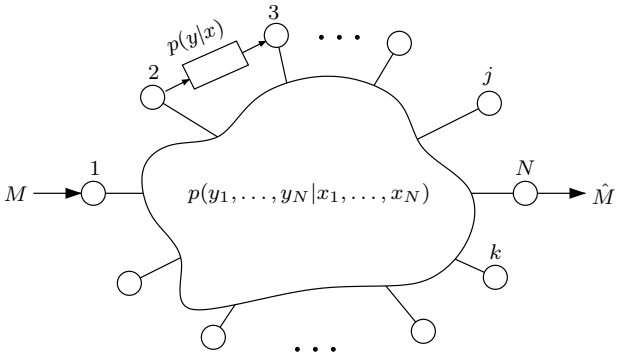

Fig. 2. DMN with an orthogonal DMC $p(y|x)$ from node 2 to node 3.

Roughly speaking, Theorem 2 and its generalization to multiple messages state that the network capacity region depends on the constituent DMC only through its capacity C. (To make this statement precise, we need some continuity of the network capacity region in C, which holds, for example, when $C > 0$ [3] or when the network has special structures [9], [10], [13].)

The existing proofs [3], [11], [12] of the network equivalence theorem and its generalizations are based on joint typicality of input–output sequences in the network or on arguments similar to empirical channel simulation [14] of the DMC q_1 from the DMC q_2. Throughout these proofs, the network-stacking technique [3], which combines multiple independent copies of the network in a carefully orchestrated manner, plays a crucial role.

Our proof is no exception and hinges heavily on the network-stacking technique. We leverage, however, the following fact from Theorem 1:

$$C(\mathcal{N}) = C_{\mathrm{cr,avg,bit}}(\mathcal{N}). \quad (14)$$

This identity allows for the use of *strong* channel simulation of q_1 from q_2, which leads to a simple, alternative proof of Theorem 2.

Before we describe the detailed proof, we recall key results on channel simulation.

A. Channel Simulation

Consider a pair of DMCs $(\mathcal{X}, q_1(y|x), \mathcal{Y})$ and $(\tilde{\mathcal{X}}, q_2(\tilde{y}|\tilde{x}), \tilde{\mathcal{Y}})$ with capacities C_1 and C_2, respectively. Suppose that a pair of sender and receiver, who share common randomness $W \sim \mathrm{Unif}[0,1]$, is connected through the DMC $q_2(\tilde{y}|\tilde{x})$ and wishes to simulate the behavior of the DMC $q_1(y|x)$ for l uses of the channel; see Fig. 3. More

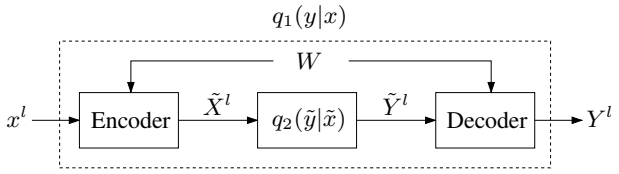

Fig. 3. Simulating $q_1(y|x)$ from $q_2(\tilde{y}|\tilde{x})$.

formally, we define a length l channel simulation code (f, g) by

- $f : \mathcal{X}^l \times [0,1] \to \tilde{\mathcal{X}}^l$ that assigns $\tilde{x}^l$ to each (x^l, w) and
- $g : \tilde{\mathcal{Y}}^l \times [0,1] \to \mathcal{Y}^l$ that assigns y^l to each $(\tilde{y}^l, w)$.

The code (f, g) induces the conditional pmf

$$p(y^l|x^l) = \int_0^1 \Big(\prod_{j=1}^{l} q_2(\tilde{y}_j | f_j(x^l, w)) \Big) p(y^l|\tilde{y}^l, w)\, dw,$$

where f_j denotes the j-th coordinate of $f(x^l, w)$ and $p(y^l|\tilde{y}^l, w) = 1$ if $y^l = g(\tilde{y}^l, w)$ and 0 otherwise. The fidelity of simulation is measured by

$$\max_{x^l} d_{\mathrm{TV}}\big(p(y^l|x^l), q_1(y^l|x^l)\big), \tag{15}$$

where $q_1(y^l|x^l) = \prod_{j=1}^{l} q_1(y_j|x_j)$.

We state the asymptotic behavior of channel simulation, which is a slight generalization of a recent result by Bennett et al. [4].

Lemma 6 (Universal channel simulation): If $C_1 < C_2$, then there exists a sequence of length-l channel simulation codes (f, g) such that

$$\lim_{l\to\infty} \max_{x^l} d_{\mathrm{TV}}\big(p(y^l|x^l), q_1(y^l|x^l)\big) = 0.$$

The proof of this lemma is somewhat involved and will be presented elsewhere.

We remark on two other fidelity criteria for channel simulation that are widely used in the literature. Let $X^l \sim p(x^l) = \prod_{j=1}^{l} p_X(x_j)$. First, *empirical* channel simulation [14] aims to achieve

$$d_{\mathrm{TV}}\big(\pi(x, y|X^l, Y^l), p(x)q_1(y|x)\big) \to 0 \quad \text{in probability,}$$

where $\pi(x, y|X^l, Y^l) := |\{j : (X_j, Y_j) = (x, y)\}| \,/\, l$ denotes the joint type of (X^l, Y^l). Second, *strong* channel simulation [14]–[17] aims to achieve

$$\lim_{l\to\infty} d_{\mathrm{TV}}\big(p(x^l)p(y^l|x^l), p(x^l)q_1(y^l|x^l)\big) = 0.$$

It is easy to verify that universal simulation implies strong simulation and the latter, in turn, implies empirical simulation.

B. Proof of Theorem 2

In light of (14), it suffices to establish that

$$C(\mathcal{N}_1) \le C_{\mathrm{cr,avg,bit}}(\mathcal{N}_2).$$

We show this by constructing a $(2^{nlR}, nl)$ randomized code for $\mathcal{N}_2$ with average bit error probability $\le 2\epsilon$ from l copies of a $(2^{nR}, n)$ code $(\phi, \psi) = (\phi_1^n, \ldots, \phi_N^n, \psi_N)$ for $\mathcal{N}_1$ with average (block) error probability $\le \epsilon$ and n copies of a length-l code (f, g) for simulating q_1 from q_2 with maximum total variation distance $\le \epsilon/n$ (cf. (15)).

As illustrated in Fig. 4, we apply the length-l simulation code (f, g) to the DMC q_2 n times over transmission times $(1, \ldots, l)$, $(l+1, \ldots, 2l)$, $\ldots$, $((n-1)l+1, \ldots, nl)$. This "inner code" induces the channel

$$p(y^{nl}|x^{nl}) = \prod_{i=1}^{n} p(y_{(i-1)l+1}^{il} | x_{(i-1)l+1}^{il})$$

from the physical channel $\prod_{t=1}^{nl} q_2(\tilde{y}_t|\tilde{x}_t)$. We then apply the $(2^{nR}, n)$ code (ϕ, ψ) to the channel $q_0(y^N|x^N)$, as well as the induced channel $p(y^{nl}|x^{nl})$, l times over transmission times $(1, l+1, \ldots, (n-1)l+1), (2, l+2, \ldots, (n-1)l+2), \ldots, (l, 2l, \ldots, nl)$ along the horizontal direction in Fig. 4. In particular, the j-th component of this "outer code" is used to communicate 2^{nR} bits $S^{nR}(j)$ with channel inputs $X^n(j)$ and

$$X_*^n(j) = (X_1^n(j), X_2^n(j), \ldots, X_N^n(j)),$$

and outputs $Y_n(j)$ and

$$Y_*^n(j) = (Y_1^n(j), Y_2^n(j), \ldots, Y_N^n(j)).$$

Note that the overall joint pmf of the message bits and channel variables is of the form

$$\prod_{j=1}^{l} \Bigg[p(s^{nR}(j)) p(y_*^n(j)|x_*^n(j)) \cdot \Big(\prod_{i=1}^{n} p(x_{*i}(j), x_i(j)|y_*^{i-1}(j), y^{i-1}(j)) \Big) \cdot p(\hat{s}^{nR}(j)|y_N^n(j)) \Bigg] \cdot \prod_{i=1}^{n} p(y_{(i-1)l+1}^{il} | x_{(i-1)l+1}^{il}),$$

where the factors on the second and third lines follow the $(2^{nR}, n)$ code (ϕ, ψ) and the factors on the last line follow the length-l simulation code (f, g). We denote this (true)

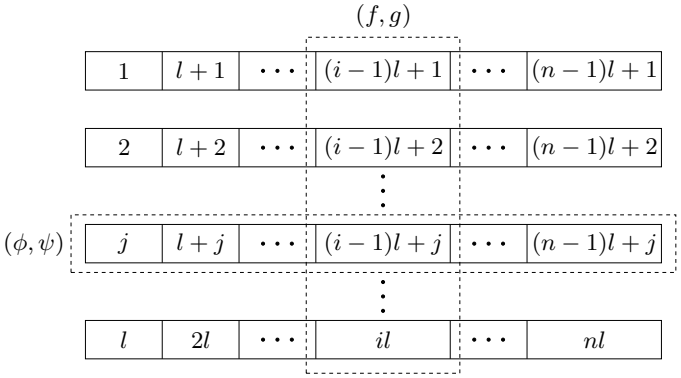

Fig. 4. A network-stacking technique of constructing a $(2^{nlR}, nl)$ code for $\mathcal{N}_2$ by horizontally applying a $(2^{nR}, n)$ code (ϕ, ψ) for $\mathcal{N}_1$ and vertically applying a length-l channel simulation code (f, g). The numbers in the boxes denote the transmission time indices.

probability law by P. Now, if the same code is used instead for the network $\mathcal{N}_1$, then the overall joint pmf is identical, except for the last line replaced by

$$\prod_{i=1}^{n} q_1(y_{(i-1)l+1}^{il}|x_{(i-1)l+1}^{il}) := \prod_{j=1}^{l}\prod_{i=1}^{n} q_1(y_{(i-1)l+j}|x_{(i-1)l+j}).$$

We denote this (target) probability law by Q_1. Since the induced channel simulates the channel q_1 in the sense that

$$d_{\mathrm{TV}}\big(p(y_{(i-1)l+1}^{il}|x_{(i-1)l+1}^{il}), q_1(y_{(i-1)l+1}^{il}|x_{(i-1)l+1}^{il})\big) \le \frac{\epsilon}{n}$$

for every $x_{(i-1)l+1}^{il} \in \mathcal{X}^l$ and $i \in [1:n]$, it follows by the properties of the total variation distance that the marginal distributions of the induced channel satisfy

$$d_{\mathrm{TV}}\big(p(y_{(i-1)l+j}|x_{(i-1)l+j}), q_1(y_{(i-1)l+j}|x_{(i-1)l+j})\big) \le \frac{\epsilon}{n}$$

for every $x_{(i-1)l+j} \in \mathcal{X}$, $j \in [1:l]$, and $i \in [1:n]$, and consequently, their product (across horizontal time indices) satisfies

$$d_{\mathrm{TV}}\big(p(y^n(j)|x^n(j)), q_1(y^n(j)|x^n(j))\big) \le \epsilon \qquad (16)$$

for every $x^n(j) \in \mathcal{X}^n$ and $j \in [1:l]$.

We are now ready to bound the average bit error probability of the $(2^{nlR}, nl)$ code for the DMN $\mathcal{N}_2$, namely,

$$P_{e,\mathrm{bit}}^{(n)}(\mathcal{N}_2) = \frac{1}{l}\sum_{j=1}^{l}\frac{1}{nR}\sum_{\nu=1}^{nR}\mathsf{P}\{\hat{S}_\nu(j) \neq S_\nu(j)\}.$$

We show that $\mathsf{P}\{\hat{S}_\nu(j) \neq S_\nu(j)\} \le 2\epsilon$ for every ν and j. By the definition of total variation distance, we have

$$\mathsf{P}\{\hat{S}_\nu(j) \neq S_\nu(j)\} \le \mathsf{Q}\{\hat{S}_\nu(j) \neq S_\nu(j)\} + d_{\mathrm{TV}}(\mathsf{P}, \mathsf{Q}_1 \mid \hat{S}_\nu(j), S_\nu(j)),$$

where the second term denotes the total variation distance between the two marginal distributions of $(\hat{S}_\nu(j), S_\nu(j))$. The first term can be upper bounded by ϵ from our assumption on the $(2^{nR}, n)$ code (f, g) for $\mathcal{N}_1$ (i.e., under Q_1) with average block (and consequently bit) error probability $\le \epsilon$. The second term can be upper bounded by the total variation distance $d_{\mathrm{TV}}(\mathsf{P}, \mathsf{Q}_1)$ between the distributions on all the random variables for the j-th code component, namely, $S^{nR}(j)$, $\hat{S}^{nR}(j)$, $X_*^n(j)$, $Y_*^n(j)$, $X^n(j)$, and $Y^n(j)$. The two distributions are identical except for $p(y^n(j)|x^n(j))$ and $q_1(y^n(j)|x^n(j))$ and their distance is upper bounded by ϵ due to (16). To see this, let $\mathbf{Z} = (S^{nR}(j), \hat{S}^{nR}(j), X_*^n(j), Y_*^n(j))$, $\mathbf{X} = X^n(j)$, and $\mathbf{Y} = Y^n(j)$, and note that

$$\begin{aligned} d_{\mathrm{TV}}(\mathsf{P}, \mathsf{Q}_1) &= d_{\mathrm{TV}}\big(p(\mathbf{z}, \mathbf{x}, \mathbf{y}), q_1(\mathbf{z}, \mathbf{x}, \mathbf{y})\big) \\ &= \sum_{\mathbf{z},\mathbf{x}} p(\mathbf{z}, \mathbf{x}|\mathbf{y}) d_{\mathrm{TV}}\big(p(\mathbf{y}|\mathbf{x}), q_1(\mathbf{y}|\mathbf{x})\big) \\ &\le \epsilon \sum_{\mathbf{z},\mathbf{x}} p(\mathbf{z}, \mathbf{x}|\mathbf{y}) \\ &\le \epsilon, \end{aligned}$$

regardless of the specific factorization structure of $p(\mathbf{z}, \mathbf{x}|\mathbf{y})$. This completes the proof of Theorem 2.

Remark 2: We can obtain an alternative proof by using *strong* channel simulation and following essentially identical steps to the current proof. Bounding the total variation distances, however, is less transparent in that approach.

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation under Grant CCF-1320895.

REFERENCES

[1] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 2nd ed. Cambridge: Cambridge University Press, 2011.

[2] N. Cai, "The maximum error probability criterion, random encoder, and feedback, in multiple input channels," *Entropy*, vol. 16, no. 3, pp. 1211–1242, 2014.

[3] R. Koetter, M. Effros, and M. Médard, "A theory of network equivalence–part I: Point-to-point channels," *IEEE Trans. Inf. Theory*, vol. 57, no. 2, pp. 972–995, Feb. 2011.

[4] C. Bennett, I. Devetak, A. Harrow, P. Shor, and A. Winter, "Quantum reverse Shannon theorem," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2926–2959, May 2014.

[5] A. El Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge: Cambridge University Press, 2011.

[6] G. Dueck, "Maximal error capacity regions are smaller than average error capacity regions for multi-user channels," *Probl. Control Inf. Theory*, vol. 7, no. 1, pp. 11–19, 1978.

[7] G. D. Forney, Jr., *Concatenated codes*. Cambridge, MA: MIT Press, 1966.

[8] F. M. J. Willems, "The maximal-error and average-error capacity region of the broadcast channel are identical: A direct proof," *Probl. Control Inf. Theory*, vol. 19, no. 4, pp. 339–347, 1990.

[9] S. P. Borade, "Network information flow: Limits and achievability," in *Proc. IEEE Internat. Symp. Inf. Theory*, Lausanne, Switzerland, July 2002, p. 139.

[10] L. S. Song, R. W. Yeung, and N. Cai, "A separation theorem for single-source network coding," *IEEE Trans. Inf. Theory*, vol. 52, no. 5, pp. 1861–1871, May 2006.

[11] S. Jalali and M. Effros, "On the separation of lossy source-network coding and channel coding in wireline networks," in *Proc. IEEE Internat. Symp. Inf. Theory*, Austin, USA, June 2010, pp. 596–600.

[12] C. Tian, J. Chen, S. Diggavi, and S. Shamai, "Optimality and approximate optimality of source-channel separation in networks," *IEEE Trans. Inf. Theory*, vol. 60, no. 2, pp. 904–918, Feb. 2014.

[13] S. Jalali, M. Effros, and T. Ho, "On the impact of a single edge on the network coding capacity," in *Proc. UCSD Inf. Theory Appl. Workshop*, 2011, pp. 1–5.

[14] P. Cuff, H. H. Permuter, and T. M. Cover, "Coordination capacity," *IEEE Trans. Inf. Theory*, vol. 56, no. 9, pp. 4181–206, Sep. 2010.

[15] C. Bennett, P. Shor, J. Smolin, and A. Thapliyal, "Entanglement-assisted capacity of a quantum channel and the reverse Shannon theorem," *IEEE Trans. Inf. Theory*, vol. 48, no. 10, pp. 2637–2655, Oct. 2002.

[16] A. Winter, "Compression of sources of probability distributions and density operators," 2002. [Online]. Available: http://arXiv:quant-ph/0208131

[17] P. Cuff, "Distributed channel synthesis," *IEEE Trans. Inf. Theory*, vol. 59, no. 11, pp. 7071–7096, Nov. 2013.