

On Capacity of Line Networks

Urs Niesen, *Student Member, IEEE*, Christina Fragouli, *Member, IEEE*, and Daniela Tuninetti, *Member, IEEE*

Abstract—We consider communication through a cascade of discrete memoryless channels (DMCs). The source and destination node of this cascade are allowed to use coding schemes of arbitrary complexity, but the intermediate relay nodes are restricted to process only blocks of a fixed length. We investigate how the processing at the relays must be chosen in order to maximize the capacity of the cascade, that is, the maximum achievable end-to-end rate between the source and the destination. For infinite cascades with fixed intermediate processing length at the relays, we prove that this intermediate processing can be chosen to be identical without loss of optimality, and that the capacity of the cascade coincides with the rate of the best zero-error code of length equal to the block length of the intermediate processing. We further show that for fixed and identical intermediate processing at all relays, convergence of capacity as the length of the cascade goes to infinity is exponentially fast. Finally, we characterize how the block length of the intermediate processing must scale with the length of the cascade to guarantee a constant end-to-end rate. We prove that it is sufficient that the block length scales logarithmically with the network length in order to achieve any rate above the zero-error capacity. We show that in many cases of interest logarithmic growth is also necessary.

Index Terms—Capacity, network coding, zero-error capacity.

I. INTRODUCTION

COMMUNICATION systems today are organized in large-scale networks, the Internet being the most conspicuous example, where information needs to traverse multiple hops to reach a destination. Each of the hops may introduce errors that become more and more pronounced as the size of the network grows. Two main approaches are used for error correction: automatic repeat request (ARQ) schemes and packet-level forward error correction (FEC) schemes. The schemes employed today are end-to-end: error correction is only performed at the source and destination node (by retransmission of packets in

the ARQ case, and by coding and decoding in the FEC case), while intermediate nodes are only allowed to replicate and forward packets. This end-to-end approach can lead to a significant waste of resources, that is becoming increasingly less acceptable as multimedia applications become more popular and bandwidth demanding.

From a theoretical point of view, it is well known that decoding and re-encoding the information sent by the source node (without imposing computational restrictions) at all intermediate nodes achieves the “min-cut capacity,” as rigorously described in [1]. Such a scheme imposes heavy computational requirements on the intermediate nodes of the network, especially since these nodes typically need to accommodate a large number of traffic connections. Moreover, it incurs large delay, which is prohibitive for real-time applications.

Recently, it was demonstrated that even for lossless links, allowing intermediate nodes to process information can increase the achievable rate in a multicasting scenario [2], [3]. The proposed approach, termed “network coding” requires intermediate nodes to perform linear combinations over a finite field. The complexity of the computations is proportional to the size of the finite field, which in turn is bounded as a function of the number of receivers [4]–[6]. The interesting observation is that allowing intermediate nodes to perform *finite complexity* processing may not only increase the achievable end-to-end rate, but actually achieve the min-cut capacity of the network. Moreover, the emergence of network coding helped to realize that intermediate node processing is plausible and compelling for new protocols, designed, for example, for overlay networks.

Motivated by these observations, we investigate in this paper what benefits finite complexity processing at intermediate nodes may offer. We restrict our attention to unicast communication, i.e., a single source–destination pair. In fact, today in the Internet, almost all traffic, including multicasting, is implemented via multiple unicast sessions.

We consider a communication network where a source node transmits information to a destination node along a path that comprises L consecutive links of the network. We assume that each link corresponds to a discrete memoryless channel (DMC). In other words, we model the communication path between the source and the destination as a line network consisting of L cascaded identical DMCs. This model captures both physical layer and application layer communication.

To measure complexity, we allow each intermediate node to process blocks of N symbols, and use N as our complexity measure. This definition of complexity allows to bound not only processing complexity, but also delay and memory requirements at intermediate nodes. Moreover, it is well suited to an environment where information is transmitted in packets. We allow the source and the destination to possibly code and decode over an unbounded number of length- N blocks.

Manuscript received December 22, 2005; revised May 1, 2007. This work was supported in part by the Swiss National Science Foundation under award PP002–110483. The material in this paper was presented in part at the IEEE International Symposium on Information Theory and Applications, Parma, Italy October 2004; at the IEEE International Symposium on Information Theory, Adelaide, Australia, September 2005; at the First Information Theory and Applications Workshop, San Diego, CA January 2006; at the 43rd Allerton Conference on Communication, Control and Computing, Monticello, IL, October 2005; and at the IEEE International Symposium on Information Theory, Seattle, WA, July 2006.

U. Niesen is with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: uniesen@mit.edu).

C. Fragouli is with the School of Computer and Communication Sciences, ALGO, EPFL, Lausanne, CH 1015, Switzerland (e-mail: christina.fragouli@epfl.ch).

D. Tuninetti is with the Electrical and Computer Engineering Department, University of Illinois at Chicago, UIC, Chicago, IL 60607-7053 USA (e-mail: daniela@ece.uic.edu).

Communicated by T. J. Richardson, Associate Editor for Coding Theory.
Digital Object Identifier 10.1109/TIT.2007.907513

This assumption reflects the strong motivation of the source and the destination to devote resources toward their own communication.

We are interested in the maximum information-theoretic rate at which the source can reliably transmit data to the destination as a function of the block length N and the network length L . The main contributions of this work are as follows. We show that, if the network length increases ($L \rightarrow \infty$) but the block length N is fixed, the optimal processing is identical at each relay and corresponds to a zero-error code. The capacity of the cascade is the rate of this zero-error code. The zero-error capacity is the maximum rate at which we can communicate over a channel with zero probability of error [7]. An intuitive interpretation of this result is that, as $L \rightarrow \infty$, the zero-error capacity is the only part of the transmitted information rate that we may hope to preserve. This limiting result, apart from its theoretical interest, applies to large networks where a packet needs to traverse a large, but not infinite, number of relays and the relays use the same fixed processing. For example, in wireless *ad hoc* networks with n nodes, the average path length L scales as \sqrt{n} , and it is a standard assumption that the nodes are simple identical devices [8]. In the case of identical processing at each relay, we show that the rate of convergence to the limiting end-to-end rate is exponential in the number of cascaded channels. We also examine how fast the block length N needs to grow with the network length L in order to achieve a constant fraction of the min-cut (as opposed to the zero-error) capacity as $L \rightarrow \infty$. We show that logarithmic growth is sufficient to achieve any fraction of the min-cut capacity above the zero-error capacity and is also necessary in many cases of interest.

Cascades of DMCs *without processing* at the intermediate nodes have been considered also in [9]–[11]. FEC schemes, that employ intermediate processing, but designed assuming $N \rightarrow \infty$ are proposed in [12], [13].

The work in [9] gives an expression for the capacity of a cascade of L identical channels without intermediate processing in terms of the eigendecomposition of the channel transition matrix, which is assumed diagonalizable and nonsingular. It is implicit in [9, eq. (10)] that in the limit of long cascades, only terms associated with eigenvalues of modulus one determine the limiting end-to-end rate. Here, we consider cascades of L identical channels with general transition matrices, not necessarily diagonalizable and nonsingular, and with intermediate processing. For the case without intermediate processing, we make explicit the dependency of the limiting capacity on the eigenvalues of modulus one, and we characterize the asymptotic rate of convergence to the limiting capacity in term of the second largest eigenvalue modulus.

The work in [10] considers the capacity and the error probability of a cascade of identical binary channels, not necessarily symmetric, with “delayless operation of the intermediate station,” i.e., intermediate processing of length $N = 1$. It identifies which binary channel gives the lowest error probability in a cascade. In the appendix of [10], it is also shown that the number of Z-channels which can be cascaded while still guaranteeing some constant (but small) end-to-end rate is inversely proportional to the probability of error over one link. The techniques used in [10] depend crucially on the properties of a Z-channel

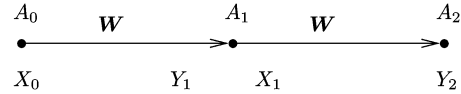


Fig. 1. A line network with two channels and one relay ($L = 2$).

and are only valid for repetition coding at the relays, that is, the approach does neither directly carry over to other channels nor extend for rates bigger than $1/N$. Here, we develop a theory for general intermediate processing and general DMCs.

The work in [11] considers the optimal ordering of different binary channels, not necessarily symmetric, such that the capacity of the overall cascade is maximized. The optimal ordering is found for the special cases of binary channel “when all channel matrices have positive determinant, and when relabeling of inputs and outputs is allowed between channels in a cascade,” i.e., with processing of length $N = 1$. Theorem 9 in [11] gives the rate of convergence to the limiting capacity of cascades of nonbinary channels with a single eigenvalue of modulus one in terms of the second largest eigenvalue modulus. Here, we develop a theory for general intermediate processing, not necessarily $N = 1$, and for general DMCs not necessarily for DMCs with a single eigenvalue of modulus one.

In this paper, we restrict our attention to cascades of DMCs. Our approach and employed tools are not directly applicable to cascades of *continuous* channels with intermediate processing. Indeed, optimizing the capacity for cascades of continuous channels depends not only on the length of the cascade and of the intermediate processing, but also on constraints, such as power constraints, that can significantly impact the problem’s optimal solution.

This paper is organized as follows. Section II formally introduces the network model and briefly summarizes the main results of the paper. Section III calculates the capacity of an infinite cascade of identical channels without intermediate processing. Section IV identifies the optimal finite length intermediate processing for an infinite cascade. Section V derives upper and lower bounds on the capacity. Section VI determines how the length of the processing must scale with the length of the cascade in order to achieve a fraction of the min-cut capacity above the zero-error capacity. Finally, Section VII concludes the paper and briefly discusses open problems and future work directions.

II. NETWORK MODEL AND MAIN RESULTS

A. Network Model, Notation, and Basic Concepts

We consider line networks with $L - 1$ relays as depicted in Fig. 1. The source A_0 sends information to the destination A_L via relays $\{A_i\}_{i=1}^{L-1}$. Each link corresponds to the same DMC with finite input alphabet \mathcal{X} , finite output alphabet \mathcal{Y} , and arbitrary transition probability matrix \mathbf{V} . We impose that all the DMCs in the cascade are identical.

We use the following notational conventions. We denote by $\mathcal{S}_{n,m}$ the set of stochastic matrices of dimension $n \times m$. We use boldface to indicate probability mass functions and transition probabilities, represented as vectors and matrices, respectively. All logarithms are with respect to the natural basis, with the exception of plots, where capacities are in bits per channel use.

We restrict the relays $\{A_i\}_{i=1}^{L-1}$ to perform operations from blocks of *exactly* N symbols in \mathcal{Y} to blocks of *exactly* N symbols in \mathcal{X} in a memoryless way across blocks. Using N times the channel \mathbf{V} between A_i and A_{i+1} , amounts to connecting A_i and A_{i+1} through an equivalent DMC with input alphabet \mathcal{X}^N , output alphabet \mathcal{Y}^N , and transition probability matrix

$$\mathbf{W} \triangleq \mathbf{V}^{\otimes N}$$

where \otimes denotes the Kronecker product. We denote by $Y_i \in \mathcal{Y}^N$ the output of channel i (observed by relay A_i) and by $X_i \in \mathcal{X}^N$ the input of channel $i + 1$ (sent by relay A_i). X_i is then a (not necessarily deterministic) function of Y_i . This function can be described by a transition probability matrix $\mathbf{M}_i \in \mathcal{S}_{|\mathcal{Y}|^N, |\mathcal{X}|^N}$ specifying for each realization x of X_i and y of Y_i the probability $\mathbb{P}(X_i = x | Y_i = y)$.

We allow the source A_0 and the destination A_L to perform coding and decoding of arbitrary complexity, across an arbitrary number of symbols in \mathcal{X}^N and \mathcal{Y}^N .

We are interested in identifying the set of processing $\{\mathbf{M}_i\}_{i=1}^{L-1}$ performed at the relays that maximizes the achievable rate between the source and the destination. This is exactly the capacity of the overall channel

$$\mathbf{W} \prod_{i=1}^{L-1} (\mathbf{M}_i \mathbf{W}) \quad (1)$$

that includes the intermediate processing $\{\mathbf{M}_i\}_{i=1}^{L-1}$ as part of the channel transition probability matrix. Our goal is to determine

$$C_{N,L}(\mathbf{V}) \triangleq \max_{\{\mathbf{M}_i\}_{i=1}^{L-1}} \frac{1}{N} C \left(\mathbf{W} \prod_{i=1}^{L-1} (\mathbf{M}_i \mathbf{W}) \right),$$

the capacity of the overall channel (1) normalized by N , the number of uses of the underlying channel \mathbf{V} . Here and in the following:

$$C(\mathbf{Q}) \triangleq \max_{\mathbf{p}} I(\mathbf{p}, \mathbf{Q})$$

where $I(\mathbf{p}, \mathbf{Q})$ is the mutual information between the input X and the output Y when $X \sim \mathbf{p}$ and $Y|X \sim \mathbf{Q}$.

Intuitively, the processing at the relays can be understood as a decoder followed by an encoder for an inner code of length N used over the channel \mathbf{V} . More precisely, we can show [14] that for every relay A_i , the optimal processing \mathbf{M}_i corresponds to a deterministic function and that it implements a decode–re-encode operation for some code of rate¹ $\frac{1}{N} \log \text{rank}(\mathbf{M}_i)$.

In this paper, we will make connections between $C_{N,L}(\mathbf{V})$ and the zero-error capacity of the underlying channel \mathbf{V} . Recall that the zero-error capacity is defined as the maximum rate at which communication is possible with zero probability of error. The notion of zero-error capacity was introduced in [7] (see [15] for further details). It can be computed as follows. For a channel with transition matrix \mathbf{V} , we call two input letters k and ℓ adjacent if there exists an output letter j such that

¹A matrix is said to be binary if all its entries are in $\{0, 1\}$. A stochastic matrix \mathbf{M}_i corresponds to a deterministic processing if it is binary. For such a processing, $\text{rank}(\mathbf{M}_i)$ is the number of inputs of the channel $i + 1$ in the cascade that are actually used.

$[\mathbf{V}]_{k,j} > 0$ and $[\mathbf{V}]_{\ell,j} > 0$. We then construct a graph $G(\mathbf{V})$ corresponding to the stochastic matrix \mathbf{V} having as vertex set the possible inputs of \mathbf{V} and in which two vertices are connected if the corresponding input letters are adjacent. Denote by $M_0(\mathbf{V})$ the largest number of vertices in $G(\mathbf{V})$ no two of which are connected by an edge (or, equivalently, the largest number of input letters of \mathbf{V} no two of which are adjacent). In graph theory, $M_0(\mathbf{V})$ is called the *independence number* of $G(\mathbf{V})$. The zero-error capacity of \mathbf{V} is then given by

$$C_0(\mathbf{V}) \triangleq \lim_{N \rightarrow \infty} \frac{1}{N} \log M_0(\mathbf{V}^{\otimes N}). \quad (2)$$

It can be shown that $C_0(\mathbf{V}) \geq \frac{1}{N} \log M_0(\mathbf{V}^{\otimes N})$ for all $N \in \mathbb{N}$, and hence, that the limit in (2) is equal to the supremum over all N [15].

Clearly, for any DMC with transition probability matrix \mathbf{V} , any intermediate processing of length N , and any network length L , we have

$$\frac{1}{N} \log M_0(\mathbf{V}^{\otimes N}) \leq C_{N,L}(\mathbf{V}) \leq C(\mathbf{V}) \quad (3)$$

where the lower bound is achievable by using the same zero-error code of length N at each node in the network and the upper bound is the network min-cut capacity [1, Theorem 14.10.1].

B. Main Results

Our main result states that the capacity of an infinite cascade of identical DMCs with transition probability matrix \mathbf{V} , and with intermediate nodes restricted to process only blocks of N symbols, cannot exceed the zero-error capacity of the underlying channel. More precisely, it states that the lower bound in (3) is tight as $L \rightarrow \infty$.

Theorem II.1: The capacity of an infinite cascade of identical DMCs with channel matrix \mathbf{V} and optimal intermediate processing of finite length N is

$$\lim_{L \rightarrow \infty} C_{N,L}(\mathbf{V}) = \frac{1}{N} \log M_0(\mathbf{V}^{\otimes N}). \quad (4)$$

This theorem is proved in Section IV. It tells us that in the limit as $L \rightarrow \infty$, the optimal intermediate processing is identical at each relay and corresponds to the best, in the sense of highest rate, zero-error code of length N for the channel \mathbf{V} . Hence, the capacity of the infinite cascade equals the rate of this zero-error code and is always upper-bounded by the zero-error capacity of \mathbf{V} . Notice that, by (2), any rate strictly below the zero-error capacity is achievable with finite length processing.

The next result illustrates the behavior of long cascades of identical channels, arising, for example, when the intermediate nodes perform the same processing independent of L .

Theorem II.2: For any square stochastic matrix \mathbf{Q}

$$\lim_{L \rightarrow \infty} C(\mathbf{Q}^L) = \log D(\mathbf{Q})$$

where $D(\mathbf{Q})$ is the number of eigenvalues of modulus (magnitude) one of \mathbf{Q} .

This theorem is proven in Section III. It tells us that $\lim_{L \rightarrow \infty} C(\mathbf{Q}^L)$ exists and can be easily computed as the

logarithm on the number of eigenvalues of modulus one of \mathbf{Q} . In Section III, we also show that convergence to the limiting capacity in Theorem II.2 is exponentially fast in L , and we provide tight bounds on the exponent as a function of the second largest eigenvalue modulus of \mathbf{Q} .

For finite L and $N \rightarrow \infty$, the relays can use a capacity-achieving code and communicate reliably as long as the rate of this code is below the capacity of the channel \mathbf{V} . That no other coding strategy can do better than this is clear from either the min-cut bound [1, Theorem 14.10.1], or directly from the data processing inequality. Hence, the capacity of the cascaded channel with infinite complexity processing at the relays is

$$\lim_{N \rightarrow \infty} C_{N,L}(\mathbf{V}) = C(\mathbf{V}), \quad (5)$$

i.e., the upper bound in (3) is tight. From Theorem II.1, we have that for $L \rightarrow \infty$ and finite N , the lower bound in (3) is tight. The limits in (5) and (4) might differ quite substantially. It is thus natural to ask what happens if both N and L are allowed to grow. Our last result tells us how fast N needs to grow with L in order to achieve rates above the zero-error capacity.

Theorem II.3: For any $\alpha \in [0, 1]$, define

$$N^*(L, \alpha) \triangleq \inf\{N \in \mathbb{N} : C_{N,L}(\mathbf{V}) \geq R(\alpha)\}$$

where

$$R(\alpha) \triangleq (1 - \alpha)C_0(\mathbf{V}) + \alpha C(\mathbf{V}).$$

For every channel matrix \mathbf{V} there exists $\beta \in [0, 1]$ such that² for $\alpha \in [0, 1]$

$$N^*(L, \alpha) = \mathcal{O}(\log L)$$

and for $\alpha \in (\beta, 1]$

$$N^*(L, \alpha) = \Omega(\log L)$$

as $L \rightarrow \infty$.

This theorem is proved in Section VI. The derivation of these results is founded on upper and lower bounds on $C_{N,L}(\mathbf{V})$ (derived in Section V) that are valid for all values of N and L . These bounds have merit on their own, as they increase our understanding of achievable rates for finite values of N and L .

From Theorem II.1, we know that, for any network length L , any rate below the zero-error capacity can be achieved with a processing of finite length N . Theorem II.3, on the other hand, tells us that for rates above the zero-error capacity, N needs to increase at most logarithmically with the length of the network L . Moreover, in many cases of interest, logarithmic growth is also necessary. This is the case in the following example that illustrates the use of Theorems II.1, II.2, and II.3.

Example II.1. Line Network of Binary Symmetric Channels: Consider a cascade of L BSC(p) (binary symmetric channels with crossover probability p), where $p \in (0, 1/2]$. The capacity of a BSC(p) is

$$C(\mathbf{V}) = \log(2) - \mathcal{H}(p) \triangleq C(p)$$

²We use Knuth's notation: $f(n) = \mathcal{O}(g(n))$ means that there exists a constant c and integer n_0 such that $f(n) \leq cg(n)$ for $n > n_0$; $f(n) = \Omega(g(n))$ denotes that $g(n) = \mathcal{O}(f(n))$.

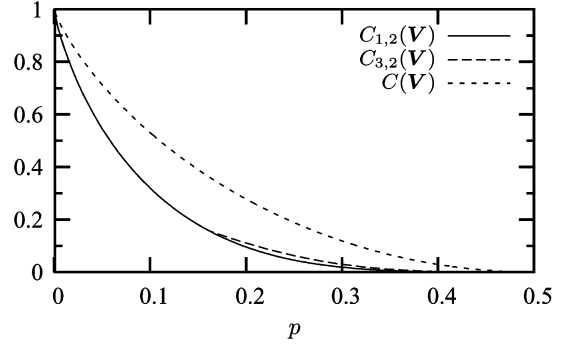


Fig. 2. Capacity of $L = 2$ cascaded BSC(p) with simple forwarding (optimal for $N \in \{1, 2\}$) $C_{1,2}(\mathbf{V})$, optimal intermediate processing of block length three $C_{3,2}(\mathbf{V})$, and optimal infinite length processing at the relay $C(\mathbf{V})$.

where $\mathcal{H}(p) \triangleq -p \log p - (1-p) \log(1-p)$ is the binary entropy function. The zero-error capacity is $C_0(\mathbf{V}) = 0$.

Theorems II.1 and II.2 yield $\lim_{L \rightarrow \infty} C_{N,L}(\mathbf{V}) = 0$, that is, as $L \rightarrow \infty$ finite-length processing does not offer any benefits. In other words, whether we use finite-length processing or no processing at all, we cannot reliably transmit information from the source to the destination at any positive rate. Moreover, if all intermediate nodes perform the same processing independent of L then this limit is reached exponentially fast in L .

Now assume that we need to transmit information through this network at a strictly positive rate, our next question is how N should scale with L in order to achieve this. Theorem II.3 tells us that logarithmic growth of N with L is *sufficient* to achieve any positive fraction of the min-cut capacity, i.e., $N^*(L, \alpha) = \mathcal{O}(\log L)$ for all $\alpha \in [0, 1)$. Furthermore, as we will show in Example VI.2, for a network of binary symmetric channels, the constant β in Theorem II.3 is equal to zero, and hence logarithmic growth of N with L is also *necessary* to achieve any positive fraction of the min-cut capacity, i.e., $N^*(L, \alpha) = \Omega(\log L)$ for all $\alpha \in (0, 1]$.

Note that, for finite L , finite complexity processing at the intermediate nodes can benefit the overall end-to-end achievable rate. To see this, consider the case where the intermediate nodes simply *forward* the incoming bits to the next node without further processing. For a cascade of BSC(p), forwarding is clearly the optimal processing for $N = 1$. In fact, since a cascade of L BSC(p) is itself a binary symmetric channel with parameter $\frac{1-(1-2p)^L}{2}$, the end-to-end achievable rate for $N = 1$ equals

$$C_{1,L}(\mathbf{V}) = C(\mathbf{V}^L) = C\left(\frac{1-(1-2p)^L}{2}\right).$$

Forwarding is also the optimal processing for $L = 2$ and $N = 2$, that is, $C_{2,2}(\mathbf{V}) = C_{1,2}(\mathbf{V})$. Fig. 2 illustrates, for $L = 2$ cascaded BSC(p), the forwarding capacity $C_{1,2}(\mathbf{V})$ and the min-cut capacity $C(\mathbf{V})$. The same figure also shows the achievable end-to-end rate when the intermediate node decodes and re-encodes a repetition code of length $N = 3$. By exhaustive search, for $L = 2$ and $N = 3$, the optimal processing can be found to be either forwarding (small p) or repetition coding (large p). \diamond

III. CAPACITY OF A CASCADE OF IDENTICAL CHANNELS

In Section IV, we will prove that *identical* processing at the relays is optimal for an infinitely long cascade of channels. In this section, we hence focus on the capacity of an infinite cascade of a DMC with input and output alphabets of the same cardinality and transition matrix \mathbf{Q} , that is, we compute

$$\lim_{L \rightarrow \infty} C(\mathbf{Q}^L). \quad (6)$$

In other words, the channel \mathbf{Q} is cascaded with itself without any intermediate processing, or \mathbf{Q} can be the result of applying the same intermediate processing at each relay, that is, $\mathbf{Q} = \mathbf{M}\mathbf{W}$, for some \mathbf{M} independent of L .

In Section III-A, we briefly review the canonical decomposition of nonnegative matrices, and we compute the limit and rate of convergence of \mathbf{Q}^L as $L \rightarrow \infty$. We use these results in Section III-B to characterize the limiting capacity of an arbitrary channel cascaded L times with itself, and in Section III-C, to characterize the rate of convergence to the limiting value.

A. Canonical Form of Stochastic Matrices

Our exposition closely follows [16]. Let \mathbf{Q} be an $m \times m$ square stochastic matrix and denote by $\mathcal{J} \triangleq \{1, \dots, m\}$ the set of its (row and column) indices. We say that the index i leads to index j , and write $i \rightarrow j$, if $[\mathbf{Q}^k]_{i,j} > 0$ for some $k \geq 1$. If $i \rightarrow j$ and $j \rightarrow i$, we say that i and j communicate. An index i is called *essential* if $i \rightarrow j$ implies $j \rightarrow i$. Otherwise i is called *inessential*. This partitions the set of indices \mathcal{J} into the set of essential indices \mathcal{E} and inessential indices \mathcal{I} . The set of essential indices \mathcal{E} can furthermore be partitioned into communicating classes $\mathcal{C} = \{\mathcal{E}_i\}_{i=1}^{|\mathcal{C}|}$, such that all indices communicating with each other are in the same class. The canonical form of a matrix \mathbf{Q} is obtained by relabeling its indices in such a way that all indices of the same essential communicating class are consecutive, and every inessential index is greater than any essential index. Formally, this corresponds to pre- and post-multiplying \mathbf{Q} by some permutation matrix $\mathbf{\Pi}$. This results in a matrix of the canonical form

$$\mathbf{\Pi}\mathbf{Q}\mathbf{\Pi}^T = \begin{pmatrix} \mathbf{P}_1 & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_2 & \dots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{P}_{|\mathcal{C}|} & \mathbf{0} \\ \mathbf{R}_1 & \mathbf{R}_2 & \dots & \mathbf{R}_{|\mathcal{C}|} & \mathbf{S} \end{pmatrix}. \quad (7)$$

The square matrix \mathbf{P}_i in (7) contains the transition probabilities within the i th essential communicating class, the square matrix \mathbf{S} contains the transition probabilities between the inessential indices \mathcal{I} , and the (not necessarily square) matrix \mathbf{R}_i contains the transition probabilities from the inessential indices to the i th essential communicating class. The submatrices $\{\mathbf{P}_i\}_{i=1}^{|\mathcal{C}|}$ are by construction *irreducible*.

The *period* of an index i is defined as the greatest common divisor of those k for which $[\mathbf{Q}^k]_{i,i} > 0$. All indices in the same communicating class have the same period, which is referred to as the period of the class. Denote by d_i the period of the submatrix \mathbf{P}_i . If $d_i = 1$, then \mathbf{P}_i is called *primitive*, i.e., it is irreducible and aperiodic. If $d_i > 1$, then \mathbf{P}_i can be written in a

canonical form (again by permuting indices) such that, for any integer ℓ , $\mathbf{P}_i^{d_i \ell}$ is of the form

$$\mathbf{P}_i^{d_i \ell} = \begin{pmatrix} \mathbf{P}_{i,1}^\ell & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_{i,2}^\ell & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{P}_{i,d_i}^\ell \end{pmatrix} \quad (8)$$

where the square matrices $\{\mathbf{P}_{i,j}\}_{j=1}^{d_i}$ on the main diagonal are primitive, i.e., irreducible and aperiodic.

The following lemma gives the limiting expression of \mathbf{Q}^L when $L \rightarrow \infty$ for certain \mathbf{Q} . As we shall see, the class of \mathbf{Q} covered by the theorem is general enough for the purposes of computing the capacity in (6).

Lemma III.1: Let \mathbf{Q} be a square stochastic matrix in the canonical form (7) with $|\mathcal{C}|$ primitive diagonal submatrices $\{\mathbf{P}_i\}_{i=1}^{|\mathcal{C}|}$ (i.e., \mathbf{Q} has $|\mathcal{C}|$ aperiodic essential communicating classes). Then

$$\mathbf{Q}^\infty \triangleq \lim_{L \rightarrow \infty} \mathbf{Q}^L = \begin{pmatrix} \mathbf{1}\boldsymbol{\pi}_1 & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{1}\boldsymbol{\pi}_2 & \dots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{1}\boldsymbol{\pi}_{|\mathcal{C}|} & \mathbf{0} \\ \mathbf{a}_1\boldsymbol{\pi}_1 & \mathbf{a}_2\boldsymbol{\pi}_2 & \dots & \mathbf{a}_{|\mathcal{C}|}\boldsymbol{\pi}_{|\mathcal{C}|} & \mathbf{0} \end{pmatrix} \quad (9)$$

where the row vector $\boldsymbol{\pi}_i$ is the unique stationary distribution of \mathbf{P}_i , i.e., $\boldsymbol{\pi}_i = \boldsymbol{\pi}_i \mathbf{P}_i$, and the column vector \mathbf{a}_i equals $\mathbf{a}_i \triangleq (\mathbf{I} - \mathbf{S})^{-1} \mathbf{R}_i \mathbf{1}$ (\mathbf{I} indicates the identity matrix and $\mathbf{1}$ the column vector of all ones).

Proof: The proof follows from [16, Theorems 4.1, 4.2, 4.3, and 4.7] \square

The speed of convergence of \mathbf{Q}^L to the limiting expression in (9) is exponential in L , with the exponent related to $\{|\lambda_i(\mathbf{Q})|\}_{i=1}^s$, $s \leq m$, the ordered (written without repetition) eigenvalues of \mathbf{Q} such that

$$|\lambda_1(\mathbf{Q})| > |\lambda_2(\mathbf{Q})| > \dots > |\lambda_s(\mathbf{Q})|.$$

Note that for a stochastic matrix \mathbf{Q} , $|\lambda_1(\mathbf{Q})| = 1$. Moreover, a primitive stochastic matrix has exactly one eigenvalue of modulus one. This eigenvalue, referred to as *Perron–Frobenius root*, has an algebraic and geometric multiplicity of one [16, Theorem 1.1]. The next lemma provides detailed information about the speed at which convergence in Lemma III.1 takes place.

Lemma III.2: Let \mathbf{Q} be a square stochastic matrix in the canonical form (7) with $|\mathcal{C}|$ aperiodic essential communicating classes. The entries of \mathbf{Q}^L converge to the entries of \mathbf{Q}^∞ in (9) exponentially fast³ in L , with exponent not smaller than $-\log |\lambda_2(\mathbf{Q})|$.

Proof: From [16, Theorems 1.2 and 4.2], the entries of \mathbf{P}_i^L converge to the entries of $\mathbf{1}\boldsymbol{\pi}_i$ exponentially fast, all with exponent $-\log |\lambda_2(\mathbf{P}_i)|$.

From [16, Theorem 4.3] and [17, Corollary 5.6.14], the entries of \mathbf{S}^L converge to zero exponentially fast, with exponent

³ $f(L)$ converges exponentially to a with exponent not smaller than b if $\lim_{L \rightarrow \infty} f(L) = a$ and $\liminf_{L \rightarrow \infty} -\frac{1}{L} \log |f(L) - a| \geq b$. If $\lim_{L \rightarrow \infty} -\frac{1}{L} \log |f(L) - a| = b$ then $f(L)$ converges exponentially fast with exponent b .

not smaller than $-\log |\lambda_1(\mathbf{S})|$. Note that $|\lambda_1(\mathbf{S})| < 1$ and thus the exponent is positive.

From the proof of [16, Theorem 4.7], the entries of $\mathbf{R}_i^{(L)} = \sum_{j=1}^{L-1} \mathbf{S}^j \mathbf{R}_i \mathbf{P}_i^{L-1-j}$ (the position of the matrix $\mathbf{R}_i^{(L)}$ in \mathbf{Q}^L is the same as the position of the matrix \mathbf{R}_i in \mathbf{Q}) converge to the entries of $\mathbf{a}_i \boldsymbol{\pi}_i$ exponentially fast, with exponent not smaller than $-\log \max\{|\lambda_2(\mathbf{P}_i)|, |\lambda_1(\mathbf{S})|\}$.

Hence, all entries of \mathbf{Q} converge to their respective limit with exponent not smaller than

$$-\log \max\{|\lambda_1(\mathbf{S})|, |\lambda_2(\mathbf{P}_1)|, \dots, |\lambda_2(\mathbf{P}_{|C|})|\}. \quad (10)$$

Since the eigenvalues of the block lower triangular matrix \mathbf{Q} are the union of the eigenvalues of its diagonal blocks [18, Exercise 4, p. 64], and since $|\lambda_1(\mathbf{Q})| = 1$ because \mathbf{Q} is a stochastic matrix, (10) equals $-\log |\lambda_2(\mathbf{Q})|$, thus yielding the desired result. \square

An immediate consequence of Lemma III.1 is as follows.

Corollary III.3: Let \mathbf{Q} be a square stochastic matrix with $|C|$ aperiodic essential communicating classes. Then

$$|C| = \text{rank} \left(\lim_{L \rightarrow \infty} \mathbf{Q}^L \right) = D(\mathbf{Q})$$

where $D(\mathbf{Q})$ is the multiplicity (algebraic and also geometric) of $|\lambda_1(\mathbf{Q})| = 1$.

Proof: That the rank of $\lim_{L \rightarrow \infty} \mathbf{Q}^L$ is $|C|$ follows immediately from inspection of (9).

That $|C|$ is also the multiplicity of $|\lambda_1(\mathbf{Q})| = 1$ is a consequence of the following facts. First, the eigenvalues of a block lower triangular matrix are the union of the eigenvalues of its diagonal blocks [18, Exercise 4, p. 64]. Second, each stochastic matrix \mathbf{P}_i on the main diagonal of \mathbf{Q} is primitive and thus has exactly one eigenvalue of maximum modulus ($|\lambda_1(\mathbf{P}_i)| = 1$). Third, all the eigenvalues of \mathbf{S} are in modulus strictly less than one [16, Theorem 4.3]. Hence, the eigenvalue of \mathbf{Q} of maximum modulus ($|\lambda_1(\mathbf{Q})| = 1$) has multiplicity $D(\mathbf{Q})$ given by the number of primitive stochastic matrices \mathbf{P}_i on the main diagonal of \mathbf{Q} , which is $|C|$ by definition. \square

The following example illustrates these definitions and results.

Example III.1: Let $p \in (0, 1)$, $t \in (0, 1)$. Consider the DMC

$$\mathbf{Q} = \begin{pmatrix} 1-p & 0 & 0 & p \\ 0 & 1 & 0 & 0 \\ 0 & t & 1-t & 0 \\ p & 0 & 0 & 1-p \end{pmatrix}$$

which has essential indices $\mathcal{E} = \{1, 2, 4\}$, inessential indices $\mathcal{I} = \{3\}$, and two essential communicating classes $C = \{\{1, 4\}, \{2\}\}$. The canonical form, which we shall denote again by \mathbf{Q} for ease of notation, is

$$\mathbf{Q} = \begin{pmatrix} 1-p & p & 0 & 0 \\ p & 1-p & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & t & 1-t \end{pmatrix} \quad (11)$$

with

$$\mathbf{P}_1 = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}, \quad \mathbf{P}_2 = (1)$$

and

$$\mathbf{R}_1 = (0 \ 0), \quad \mathbf{R}_2 = (t), \quad \mathbf{S} = (1-t).$$

Both matrices \mathbf{P}_1 and \mathbf{P}_2 have period 1. Hence, the matrix \mathbf{Q} falls into the category covered by Lemmas III.1 and III.2 as well as Corollary III.3. The eigenvalues of \mathbf{Q} take the values $\{1, 1-2p, 1-t\}$, and $|\lambda_1(\mathbf{Q})| = 1$ has multiplicity $D(\mathbf{Q}) = 2$.

From Lemma III.1

$$\mathbf{Q}^\infty = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \quad (12)$$

and, from Lemma III.2, the speed of convergence is exponential with exponent not smaller than

$$-\log \max\{1-2p, 1-t\}. \quad (13)$$

From Corollary III.3, the rank of \mathbf{Q}^∞ is $D(\mathbf{Q}) = 2$.

In this specific example, we could also have computed directly the limiting value in (12) and the exponent in (13) from the following explicit expression for \mathbf{Q}^L :

$$\mathbf{Q}^L = \mathbf{Q}^\infty + \begin{pmatrix} +x/2 & -x/2 & 0 & 0 \\ -x/2 & +x/2 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -y & +y \end{pmatrix} \quad (14)$$

with $x = (1-2p)^L$ and $y = (1-t)^L$. \diamond

B. Capacity of an Infinite Cascade of Identical Channels

We now use the results from Section III-A to find the capacity of an infinite cascade of an arbitrary channel \mathbf{Q} without intermediate processing.

Theorem III.4: For any square stochastic matrix \mathbf{Q}

$$\lim_{L \rightarrow \infty} C(\mathbf{Q}^L) = \log D(\mathbf{Q}) \quad (15)$$

where $D(\mathbf{Q})$ is the multiplicity of eigenvalues of modulus one of \mathbf{Q} .

Proof: For any square stochastic matrix \mathbf{Q} , the limit $\lim_{L \rightarrow \infty} C(\mathbf{Q}^L)$ exists since the sequence $C(\mathbf{Q}^L)$ is nonincreasing in L (by the data processing inequality) and bounded below (by nonnegativity of mutual information). The existence of the limit implies that for any $d \in \mathbb{N}$

$$\lim_{L \rightarrow \infty} C(\mathbf{Q}^L) = \lim_{L \rightarrow \infty} C(\mathbf{Q}^{dL}). \quad (16)$$

By continuity of capacity in the channel transition probability matrix (shown in Appendix I, Lemma I.1) and if $\lim_{L \rightarrow \infty} \mathbf{Q}^{dL}$ exists (which, for appropriately chosen d , will be shown to hold in the next paragraph)

$$\lim_{L \rightarrow \infty} C(\mathbf{Q}^{dL}) = C \left(\lim_{L \rightarrow \infty} \mathbf{Q}^{dL} \right). \quad (17)$$

Without loss of generality, assume that \mathbf{Q} is in the canonical form (7). The notation is the same as in Section III-A. Let d_i be the period of the irreducible matrix \mathbf{P}_i and denote by d the least common multiple of the $\{d_i\}$. By (16), we can limit our attention to the powers of \mathbf{Q}^d . From (8), we know that \mathbf{P}_i^d is

block diagonal, since d is a multiple of d_i , and it has exactly d_i primitive square matrices on the main diagonal. Hence, \mathbf{Q}^d has the form of (7) with $\sum_i d_i$ aperiodic essential communicating classes. Hence, $\lim_{L \rightarrow \infty} \mathbf{Q}^{dL}$ is given by (9) in Lemma III.1, where the number of aperiodic essential communicating classes is $\sum_i d_i$.

Since capacity is upper-bounded by the logarithm of the rank of the channel transition probability matrix [19] and from Corollary III.3, we have that

$$\begin{aligned} C\left(\lim_{L \rightarrow \infty} \mathbf{Q}^{dL}\right) &\leq \log \text{rank}\left(\lim_{L \rightarrow \infty} \mathbf{Q}^{dL}\right) \\ &= \log \sum_i d_i \\ &= \log D(\mathbf{Q}^d) \\ &= \log D(\mathbf{Q}) \end{aligned} \quad (18)$$

where the last equality in (18) follows since the number of eigenvalues of modulus one of \mathbf{Q} is the same as the number of eigenvalues of modulus one of \mathbf{Q}^k for any integer k . It is easy to see that equality in (18) is achievable by using one input per essential communicating class of \mathbf{Q}^d with uniform probability. This shows that (15) holds, concluding the proof. \square

C. Convergence to Asymptotic Value

We showed in the last subsection that the capacity of an infinite cascade of identical DMCs can be easily computed in terms of the number of eigenvalues of modulus one of the channel transition probability matrix. In this subsection, we show that convergence to this limiting expression is exponential in the length of the cascade. This implies that for long, but still finite, cascades of identical channels, the limiting result derived in the previous section is meaningful.

We define the exponential rate at which capacity decays as

$$E_L(\mathbf{Q}) \triangleq -\frac{1}{L} \log (C(\mathbf{Q}^L) - \log D(\mathbf{Q})).$$

The following theorem asymptotically bounds $E_L(\mathbf{Q})$ in terms of the eigenvalues of the channel matrix \mathbf{Q} .

Theorem III.5: Let \mathbf{Q} be a stochastic matrix, and define $\hat{\mathbf{Q}}$ as the stochastic matrix obtained by deleting all inessential indices from \mathbf{Q} . Then

$$\begin{aligned} -\log |\lambda_2(\mathbf{Q})| &\leq \liminf_{L \rightarrow \infty} E_L(\mathbf{Q}) \leq \limsup_{L \rightarrow \infty} E_L(\mathbf{Q}) \\ &\leq -2 \log |\lambda_2(\hat{\mathbf{Q}})| \end{aligned} \quad (19)$$

where $|\lambda_2(\mathbf{Q})|$ denotes the second largest eigenvalue modulus of the channel matrix \mathbf{Q} .

Moreover, if $\mathbf{Q} = \hat{\mathbf{Q}}$, then $\lim_{L \rightarrow \infty} E_L(\mathbf{Q})$ exists and is equal to the upper bound in (19).

Proof: The proof, due to its length and technicality, is reported in Appendix II. \square

Interestingly, the speed of convergence of $C(\mathbf{Q}^L)$ in Theorem III.5 is not necessarily the same as the speed of convergence of \mathbf{Q}^L , which was derived in Lemma III.2 and was found to be equal to $-\log |\lambda_2(\mathbf{Q})|$.

We illustrate the use of Theorem III.5 by further developing Example III.1. This example shows that both the upper and the lower bounds in Theorem III.5 can be tight.

Example III.2: For the channel in Example III.1, we have $|\lambda_2(\mathbf{Q})| = \max\{|1 - 2p|, 1 - t\}$ and $|\lambda_2(\hat{\mathbf{Q}})| = |1 - 2p|$. The matrices \mathbf{Q} , \mathbf{Q}^L , and \mathbf{Q}^∞ are given in (11), (14), and (12), respectively. Theorem III.5 says that the convergence of $C(\mathbf{Q}^L)$ to $C(\mathbf{Q}^\infty) = \log(2)$ is exponentially fast in L with exponent

$$\begin{aligned} &-\log \max\{|1 - 2p|, 1 - t\} \\ &\leq \liminf_{L \rightarrow \infty} E_L(\mathbf{Q}) \leq \limsup_{L \rightarrow \infty} E_L(\mathbf{Q}) \\ &\leq -\log ((1 - 2p)^2). \end{aligned} \quad (20)$$

We now show that with the right choice of the parameters p and t , both the upper and the lower bound in (20) can be achieved. To do that, we will directly derive the speed of convergence as follows. The capacity of \mathbf{Q}^L in (14) can be computed for every L as the channel is the sum of a binary symmetric channel with parameter $p_L \triangleq \frac{1 - (1 - 2p)^L}{2}$, with capacity

$$C_{\text{BSC}}(p_L) = \log(2) - \mathcal{H}(p_L) \rightarrow \frac{(1 - 2p_L)^2}{2}$$

as $L \rightarrow \infty$, and a Z-channel with parameter $t_L \triangleq 1 - (1 - t)^L$, with capacity

$$C_Z(t_L) = \log\left(1 + \exp\left(-\frac{\mathcal{H}(t_L)}{1 - t_L}\right)\right) \rightarrow \frac{1 - t_L}{e}$$

as $L \rightarrow \infty$. Hence, for large L

$$\begin{aligned} C(\mathbf{Q}^L) &= \log(\exp(C_{\text{BSC}}(p_L)) + \exp(C_Z(t_L))) \\ &\approx \log\left(2 + \frac{(1 - 2p)^{2L}}{2} + \frac{(1 - t)^L}{e}\right) \\ &\approx \log 2 + \max\{\beta_1(1 - 2p)^{2L}, \beta_2(1 - t)^L\} \end{aligned}$$

for some constants β_1 and β_2 independent of L . We conclude that

$$\lim_{L \rightarrow \infty} E_L(\mathbf{Q}) = -\log \max\{(1 - 2p)^2, 1 - t\}.$$

Thus, if $(1 - t) \leq (1 - 2p)^2$, the upper bound in (20) is tight; if $(1 - 2p)^2 < (1 - t) < |1 - 2p|$ neither the upper nor the lower bound is tight; if $(1 - 2p)^2 \leq |1 - 2p| \leq (1 - t)$, the lower bound is tight. \diamond

IV. CAPACITY OF AN INFINITE CASCADE OF CHANNELS WITH INTERMEDIATE PROCESSING

In this section, we characterize the optimal finite-length intermediate processing for an infinite cascade and establish connections with the zero-error capacity. We start by showing that, similar to the usual capacity, the zero-error capacity obeys a sort of data processing inequality. This result is then used to show that the optimal finite-length processing for an infinite cascade is a zero-error code.

Proposition IV.1: Consider a cascade of L channels $\{\mathbf{Q}_i\}_{i=1}^L$. Then for any $j \in \{1, \dots, L\}$

$$M_0\left(\left(\prod_{i=1}^L \mathbf{Q}_i\right)^{\otimes n}\right) \leq M_0(\mathbf{Q}_j^{\otimes n}). \quad (21)$$

Remark: Equation (21) implies that

$$C_0 \left(\prod_{i=1}^L \mathbf{Q}_i \right) \leq \min_{j \in \{1, \dots, L\}} C_0(\mathbf{Q}_j).$$

Proof: By definition, if $M_0(\mathbf{Q})$ is the maximum number of inputs of the channel \mathbf{Q} that can be distinguished from the output with zero error, then for every $m \leq M_0(\mathbf{Q})$ (and for no $m > M_0(\mathbf{Q})$) there exist two binary stochastic matrices, an encoder \mathbf{M}_E and a decoder \mathbf{M}_D , such that $\mathbf{M}_E \mathbf{Q} \mathbf{M}_D$ is the identity matrix of dimension m . Conversely, if there exist two binary stochastic matrices \mathbf{M}_E and \mathbf{M}_D such that $\mathbf{M}_E \mathbf{Q} \mathbf{M}_D$ is the identity matrix of dimension m , then $m \leq M_0(\mathbf{Q})$.

Let $(\mathbf{M}_E, \mathbf{M}_D)$ be the optimal encoder and decoder for the matrix $\left(\prod_{i=1}^L \mathbf{Q}_i \right)^{\otimes n}$. By the properties of the Kronecker product [18, Theorem 4.2.10], we have

$$\begin{aligned} \mathbf{I}_m &= \mathbf{M}_E \left(\prod_{i=1}^L \mathbf{Q}_i \right)^{\otimes n} \mathbf{M}_D \\ &= \mathbf{M}_E \prod_{i=1}^L (\mathbf{Q}_i^{\otimes n}) \mathbf{M}_D \\ &= \mathbf{M}_{E,j} \mathbf{Q}_j^{\otimes n} \mathbf{M}_{D,j} \end{aligned}$$

where \mathbf{I}_m is the identity matrix of dimension

$$m \triangleq M_0 \left(\left(\prod_{i=1}^L \mathbf{Q}_i \right)^{\otimes n} \right)$$

and

$$\begin{aligned} \mathbf{M}_{E,j} &\triangleq \mathbf{M}_E \prod_{i=1}^{j-1} \mathbf{Q}_i^{\otimes n} \\ \mathbf{M}_{D,j} &\triangleq \left(\prod_{i=j+1}^L \mathbf{Q}_i^{\otimes n} \right) \mathbf{M}_D. \end{aligned}$$

Hence, there exists at least one pair of *randomized* zero-error encoder and decoder (i.e., $\mathbf{M}_{E,j}$ and $\mathbf{M}_{D,j}$) for the channel $\mathbf{Q}_j^{\otimes n}$ yielding the same rate as $\prod_{i=1}^L \mathbf{Q}_i^{\otimes n}$.

It remains to show that we can find a *nonrandomized* (i.e., binary stochastic) zero-error encoder and decoder for $\mathbf{Q}_j^{\otimes n}$. Note that the set of all stochastic matrices of fixed dimension is closed, bounded, and convex. An easy argument shows that its extreme points are binary stochastic matrices. Hence, every pair of stochastic matrices $(\mathbf{M}_E, \mathbf{M}_D)$ may be written as a convex combination of binary stochastic matrices [20, Corollary 18.5.1]

$$\begin{aligned} \mathbf{M}_{E,j} &= \sum_{i=1}^B \lambda_i \mathbf{B}_i \\ \mathbf{M}_{D,j} &= \sum_{i=1}^B \tilde{\lambda}_i \mathbf{B}_i \end{aligned}$$

where $\{\mathbf{B}_i\}_{i=1}^B$ is the set of all binary stochastic matrices of appropriate dimension for $\mathbf{M}_{E,j}$ and $\mathbf{M}_{D,j}$. Then

$$\mathbf{I}_m = \mathbf{M}_{E,j} \mathbf{Q}_j^{\otimes n} \mathbf{M}_{D,j} = \sum_{i=1}^B \sum_{l=1}^B \lambda_i \tilde{\lambda}_l \mathbf{B}_i \mathbf{Q}_j^{\otimes n} \mathbf{B}_l.$$

Each term $\mathbf{B}_i \mathbf{Q}_j^{\otimes n} \mathbf{B}_l$ is a stochastic matrix, and hence, in particular, nonnegative with at least one positive entry in each row. Hence, their sum can only be equal to \mathbf{I}_m if $\mathbf{B}_i \mathbf{Q}_j^{\otimes n} \mathbf{B}_l = \mathbf{I}_m$ whenever $\lambda_i \tilde{\lambda}_l > 0$. Since at least one such $\lambda_i \tilde{\lambda}_l$ must exist, we can choose \mathbf{B}_i and \mathbf{B}_l as nonrandomized zero-error encoder and decoder for the channel $\mathbf{Q}_j^{\otimes n}$, proving the result. \square

The next theorem shows that for an infinite cascade of identical DMCs, identical processing at the relays is optimal. This theorem is crucial as it allows us to optimize over only one intermediate processing instead of having to optimize over an infinite sequence of processing $\{\mathbf{M}_i\}_{i=1}^{\infty}$.

Theorem IV.2: For a cascade of L identical DMCs, identical processing at the relays is optimal as $L \rightarrow \infty$, i.e.,

$$\begin{aligned} \lim_{L \rightarrow \infty} \max_{\{\mathbf{M}_i\}_{i=1}^{L-1}} \frac{1}{N} C \left(\mathbf{W} \prod_{i=1}^{L-1} (\mathbf{M}_i \mathbf{W}) \right) \\ = \lim_{L \rightarrow \infty} \max_{\mathbf{M}} \frac{1}{N} C \left((\mathbf{M} \mathbf{W})^L \right). \end{aligned}$$

Moreover, the maximization on the right-hand side can be restricted to be over all binary stochastic matrices.

Proof: Recall that $\mathbf{W} \triangleq \mathbf{V}^{\otimes N}$ is the channel transition probability matrix of the equivalent DMC between any pair of relays, and that $\{\mathbf{M}_i\}_{i=1}^{L-1}$ is the processing at the relays. With $\mathbf{Q}_i \triangleq \mathbf{M}_i \mathbf{W} \in \mathcal{S}_{\mathcal{Y}^N, \mathcal{Y}^N}$ for $i = 1, \dots, L-1$, the overall channel between the source and the destination node is

$$\mathbf{W} \prod_{i=1}^{L-1} (\mathbf{M}_i \mathbf{W}) = \mathbf{W} \prod_{i=1}^{L-1} \mathbf{Q}_i.$$

An interval chain σ of length ℓ is defined to be a sequence of intervals $\{\sigma_i\}_{i=1}^{\ell}$, where the σ_i are each integer intervals $\{l_i, l_i + 1, \dots, r_i\}$ and have the property that $l_i = r_{i-1} + 1$ for all $i \in \{2, \dots, \ell\}$. Consider the product $\prod_{i=1}^{L-1} \mathbf{Q}_i$ and define $\mathbf{Q}_{\sigma_j} \triangleq \prod_{i=l_j}^{r_j} \mathbf{Q}_i$ for any integer interval $\sigma_j \subseteq \{1, \dots, L-1\}$. We will use [21, Lemma 2.4], a result originally due to Erdős and Szekeres, to show that, as $L \rightarrow \infty$, there exists an interval chain σ of arbitrary length ℓ such that all \mathbf{Q}_{σ_j} are almost identical. More precisely, for every L there exists an $\ell = \ell(L)$ satisfying $\lim_{L \rightarrow \infty} \ell(L) = \infty$ and such that

$$C \left(\mathbf{W} \prod_{i=1}^{L-1} \mathbf{Q}_i \right) = C \left(\mathbf{P} \mathbf{Q}_{\sigma_1}^{\ell(L)} \tilde{\mathbf{P}} \right) + \varepsilon(L) \quad (22)$$

for some stochastic matrices \mathbf{P} and $\tilde{\mathbf{P}}$ and with $\varepsilon(L) \rightarrow 0$ as $L \rightarrow \infty$.

For a fixed $k \in \mathbb{N}$, construct $\hat{\mathbf{Q}}$ from \mathbf{Q} by quantizing every component of \mathbf{Q} to the closest of the points $\{j/k\}_{j=0}^k$. The set of all possible quantized matrices (which are, in general, not stochastic) has cardinality $K \triangleq (k+1)^{n^2}$, with $n \triangleq |\mathcal{Y}|^N$. By [21, Lemma 2.4], we have that if $L-1 \geq \ell^K$ then there exists an interval chain σ of length ℓ such that $\hat{\mathbf{Q}}_{\sigma_1} = \hat{\mathbf{Q}}_{\sigma_j}$ for all $j \in \{1, \dots, \ell\}$. Note that $\hat{\mathbf{Q}}_{\sigma_j}$ is defined as the quantized version of \mathbf{Q}_{σ_j} and hence, $\hat{\mathbf{Q}}_{\sigma_j}$ and \mathbf{Q}_{σ_j} differ componentwise by at most $1/k$.

We now show, by induction over ℓ , that this implies that the product $\prod_{j=1}^{\ell} \mathbf{Q}_{\sigma_j}$ and $\mathbf{Q}_{\sigma_1}^{\ell}$ differ componentwise at most by

$\frac{1}{k}(4n)^{\ell-1}$. The result trivially holds for $\ell = 1$. Assume then it holds for $\ell - 1$, and let $\mathbf{A} \triangleq \prod_{i=1}^{\ell-1} \mathbf{Q}_{\sigma_i}$, $\mathbf{B} \triangleq \mathbf{Q}_{\sigma_1}^{\ell-1}$. Consider an arbitrary component of $\mathbf{A}\mathbf{Q}_{\sigma_\ell}$ and $\mathbf{B}\mathbf{Q}_{\sigma_1}$, say $(1, 1)$. We have

$$\begin{aligned} & \left| \sum_{i=1}^n a(i|1)q_{\sigma_\ell}(1|i) - \sum_{i=1}^n b(i|1)q_{\sigma_1}(1|i) \right| \\ & \leq \sum_{i=1}^n |a(i|1)q_{\sigma_\ell}(1|i) - b(i|1)q_{\sigma_1}(1|i)| \\ & \leq \sum_{i=1}^n \max_{a,q \in [0,1]} \left| \left(a + \frac{1}{k}(4n)^{\ell-2} \right) \left(q + \frac{1}{k} \right) \right. \\ & \quad \left. - \left(a - \frac{1}{k}(4n)^{\ell-2} \right) \left(q - \frac{1}{k} \right) \right| \\ & = \sum_{i=1}^n \max_{a,q \in [0,1]} \left(\frac{2}{k}(4n)^{\ell-2}q + \frac{2}{k}a \right) \\ & \leq \frac{1}{k}(4n)^{\ell-1} \end{aligned}$$

concluding the induction step.

Let $\log^i(x)$ denote the i -fold logarithm of x . Choosing

$$\begin{aligned} k(L) & \triangleq \left\lceil \left(\frac{\log(L-1)}{\log^4(L)} \right)^{1/n^2} - 1 \right\rceil \\ \ell(L) & \triangleq \left\lceil (L-1)^{1/(k(L)+1)n^2} \right\rceil \end{aligned}$$

we have $\lim_{L \rightarrow \infty} \frac{1}{k(L)}(4n)^{\ell(L)-1} = 0$, and at the same time $L-1 \geq \ell(L)^{K(L)}$ for all L and $\lim_{L \rightarrow \infty} \ell(L) = \infty$. By Appendix I Lemma I.1, $C(\cdot)$ is uniformly continuous, and hence

$$\lim_{L \rightarrow \infty} \left| C\left(\mathbf{P}\mathbf{Q}_{\sigma_1}^{\ell(L)}\tilde{\mathbf{P}}\right) - C\left(\mathbf{P}\left(\prod_{i=1}^{\ell(L)} \mathbf{Q}_{\sigma_i}\right)\tilde{\mathbf{P}}\right) \right| = 0$$

proving (22), where

$$\begin{aligned} \mathbf{P} & = \mathbf{W} \prod_{j=1}^{l_1-1} \mathbf{Q}_j \\ \tilde{\mathbf{P}} & = \prod_{j=r_\ell+1}^{L-1} \mathbf{Q}_j. \end{aligned}$$

By the data processing inequality, the capacity in (22) can be upper-bounded by

$$C\left(\mathbf{W}\prod_{i=1}^{L-1} \mathbf{Q}_i\right) \leq C\left(\mathbf{Q}_{\sigma_1}^{\ell(L)}\right) + \varepsilon(L).$$

But any stochastic matrix \mathbf{Q}_{σ_1} resulting from this procedure can be written as the product $\mathbf{M}\mathbf{W}$ for some stochastic matrix \mathbf{M} and, hence, $\mathbf{Q}_{\sigma_1}^{\ell}$ can be constructed from a cascade of ℓ channels \mathbf{W} by using the same processing at each relay. Thus

$$\begin{aligned} & \lim_{L \rightarrow \infty} \max_{\{\mathbf{M}_i\}_{i=1}^{L-1}} \frac{1}{N} C\left(\mathbf{W}\prod_{i=1}^{L-1} (\mathbf{M}_i\mathbf{W})\right) \\ & \leq \lim_{L \rightarrow \infty} \max_{\mathbf{M}} \frac{1}{N} C\left((\mathbf{M}\mathbf{W})^{\ell(L)}\right) \\ & = \lim_{L \rightarrow \infty} \max_{\mathbf{M}} \frac{1}{N} C\left((\mathbf{M}\mathbf{W})^L\right). \end{aligned}$$

It remains to show that \mathbf{M} can be restricted to be a binary stochastic matrix without loss of optimality. As argued before (see the proof of Proposition IV.1), every matrix \mathbf{M} may be written as a convex combination of binary stochastic matrices $\mathbf{M} = \sum_{i=1}^B \lambda_i \mathbf{B}_i$, where $\{\mathbf{B}_i\}_{i=1}^B$ is the set of all binary stochastic matrices of appropriate dimension. The maximization over \mathbf{M} can then equivalently be written as a maximization over all valid weights $\{\lambda_i\}_{i=1}^B$. By convexity of mutual information in the channel argument

$$\begin{aligned} & \max_{\{\lambda_i\}} C\left(\left(\sum_{i=1}^B \lambda_i \mathbf{B}_i\right) \mathbf{W}\right)^L \\ & \leq \max_{\{\lambda_i\}} \sum_{i_1=1}^B \dots \sum_{i_L=1}^B \lambda_{i_1} \dots \lambda_{i_L} C\left(\prod_{j=1}^L (\mathbf{B}_{i_j} \mathbf{W})\right). \end{aligned} \quad (23)$$

We can interpret the weights $\{\lambda_i\}_{i=1}^B$ as a probability distribution on $\{\mathbf{B}_i\}_{i=1}^B$. Then the argument of the right-hand side maximization in (23) is the expected value of capacity resulting from choosing L intermediate processing from $\{\mathbf{B}_i\}_{i=1}^B$ independently at random and according to the probability distribution $\{\lambda_i\}_{i=1}^B$. We will break this argument into two terms, with one term containing all the events where a particular \mathbf{B}_i appears at least ρ times in succession, and the other term containing the remaining events.

Let i^* denote the index corresponding to the maximum value of λ_i , i.e., $i^* = \arg \max_i \lambda_i$. Clearly, $\lambda_{i^*} \geq \frac{1}{B}$. Thus, for the corresponding \mathbf{B}_{i^*} the probability that it occurs ρ times in succession during the experiment is at least $\frac{1}{B^\rho}$. Define $\mathcal{A}^\rho(L)$ to be the set of L -tuples of indices $\{i_j\}_{j=1}^L$ such that the value λ_{i^*} (and correspondingly \mathbf{B}_{i^*}) appears at least ρ times in succession in (23)

$$\begin{aligned} \mathcal{A}^\rho(L) & \triangleq \left\{ \{i_j\}_{j=1}^L \in \{1, \dots, B\}^L : \exists 0 \leq j_0 \leq L - \rho - 1 \right. \\ & \quad \left. \text{s.t. } i_{j_0+1} = \dots = i_{j_0+\rho} = i^* \right\}. \end{aligned}$$

For a fixed ρ, L , and assuming $L = m\rho$ for some $m \in \mathbb{N}$, we have

$$\begin{aligned} \mathcal{A}^\rho(L) & \supset \left\{ \{i_j\}_{j=1}^L \in \{1, \dots, B\}^L : \exists 0 \leq m_0 \leq L/\rho - 1 \right. \\ & \quad \left. \text{s.t. } i_{\rho m_0+1} = \dots = i_{\rho(m_0+1)} = i^* \right\} \end{aligned}$$

and therefore

$$\sum_{\{i_j\} \notin \mathcal{A}^\rho(L)} \lambda_{i_1} \dots \lambda_{i_L} \leq \left(1 - \frac{1}{B^\rho}\right)^m \leq \exp(-m/B^\rho).$$

With this, we can further upper-bound (23) by

$$\begin{aligned} & \max_{\{\lambda_i\}} \sum_{\{i_j\} \in \mathcal{A}^\rho(L)} \lambda_{i_1} \dots \lambda_{i_L} C\left(\prod_{j=1}^L (\mathbf{B}_{i_j} \mathbf{W})\right) \\ & \quad + \exp(-m/B^\rho) N \log |\mathcal{X}| \\ & \leq \max_{i \in \{1, \dots, B\}} C\left((\mathbf{B}_i \mathbf{W})^\rho\right) \\ & \quad + \exp(-m/B^\rho) N \log |\mathcal{X}|, \end{aligned} \quad (24)$$

where we have used the data processing inequality. Choosing $\rho(L) \triangleq \lfloor \log^2(L) \rfloor$ and $m(L) \triangleq \lfloor L/\rho(L) \rfloor$, the second term on the right-hand side in (24) goes to zero as $L \rightarrow \infty$ while $\lim_{L \rightarrow \infty} \rho(L) = \infty$, proving the result. \square

With this last result, we are now in position to find the optimal intermediate processing of block length N for an infinite cascade of identical channels and to compute the resulting capacity of the cascade. Theorem IV.3 shows that the optimal intermediate processing at the relays corresponds to using the best zero-error code of block length N for the channel \mathbf{V} . The resulting capacity of the cascade equals the rate of this zero-error code.

Theorem IV.3: The capacity of an infinite cascade of identical DMCs with channel matrix \mathbf{V} and optimal intermediate processing of finite length N is

$$\lim_{L \rightarrow \infty} C_{N,L}(\mathbf{V}) = \frac{1}{N} \log M_0(\mathbf{V}^{\otimes N}).$$

Proof: Decompose the intermediate processing $\mathbf{M} = \mathbf{M}_D \mathbf{M}_E$, for some stochastic matrices $\mathbf{M}_D, \mathbf{M}_E$. The main steps of the proof are as follows:

$$\begin{aligned} \lim_{L \rightarrow \infty} C_{N,L}(\mathbf{V}) &= \lim_{L \rightarrow \infty} \max_{\mathbf{M}_E, \mathbf{M}_D} \frac{1}{N} C((\mathbf{M}_E \mathbf{W} \mathbf{M}_D)^L) \end{aligned} \quad (25a)$$

$$= \max_{\mathbf{M}_E, \mathbf{M}_D} \lim_{L \rightarrow \infty} \frac{1}{N} C((\mathbf{M}_E \mathbf{W} \mathbf{M}_D)^L) \quad (25b)$$

$$= \max_{\mathbf{M}_E, \mathbf{M}_D} \frac{1}{N} \log D(\mathbf{M}_E \mathbf{W} \mathbf{M}_D) \quad (25c)$$

$$= \max_{\mathbf{M}_E, \mathbf{M}_D} \frac{1}{N} \log M_0 \left(\lim_{L \rightarrow \infty} (\mathbf{M}_E \mathbf{W} \mathbf{M}_D)^{dL} \right) \quad (25d)$$

$$= \frac{1}{N} \log M_0(\mathbf{W}). \quad (25e)$$

Equality in (25a) follows from Theorem IV.2. The next step is to show that the limit and the maximization operation in (25a) can be exchanged, thus giving (25b). We postpone the proof of this technical step to a later stage. Equality in (25c) follows from Theorem III.4, where $D(\mathbf{M}_E \mathbf{W} \mathbf{M}_D)$ is the number of eigenvalues of modulus one of $\mathbf{M}_E \mathbf{W} \mathbf{M}_D$. Call d the least common multiple of the periods of the essential communicating classes of $\mathbf{M}_E \mathbf{W} \mathbf{M}_D$. From Lemma III.1, the limiting channel $\lim_{L \rightarrow \infty} (\mathbf{M}_E \mathbf{W} \mathbf{M}_D)^{dL}$ has $D(\mathbf{M}_E \mathbf{W} \mathbf{M}_D)$ nonadjacent inputs, and hence (25c) is the rate of the best zero-error code of length one for the limiting channel $\lim_{L \rightarrow \infty} (\mathbf{M}_E \mathbf{W} \mathbf{M}_D)^{dL}$, thus giving (25d). Finally, equality in (25e) follows by applying Proposition IV.1, which states that

$$M_0 \left(\lim_{L \rightarrow \infty} (\mathbf{M}_E \mathbf{W} \mathbf{M}_D)^{dL} \right) \leq M_0(\mathbf{W})$$

with equality if $(\mathbf{M}_E, \mathbf{M}_D)$ is an optimal zero-error code for $\mathbf{W} = \mathbf{V}^{\otimes N}$, as in this case $(\mathbf{M}_E \mathbf{W} \mathbf{M}_D)^\ell = \mathbf{I}^\ell = \mathbf{I}$ for any integer ℓ .

In order to complete the proof, we need to show that the limit and the maximization operation in (25a) can be interchanged. Clearly

$$\begin{aligned} \lim_{L \rightarrow \infty} \max_{\mathbf{M}_E, \mathbf{M}_D} \frac{1}{N} C((\mathbf{M}_E \mathbf{W} \mathbf{M}_D)^L) \\ \geq \max_{\mathbf{M}_E, \mathbf{M}_D} \lim_{L \rightarrow \infty} \frac{1}{N} C((\mathbf{M}_E \mathbf{W} \mathbf{M}_D)^L). \end{aligned}$$

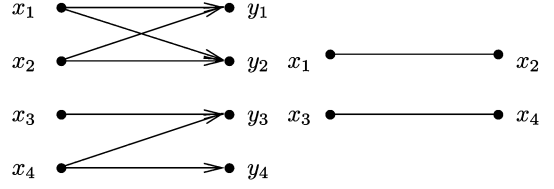


Fig. 3. The channel \mathbf{V} in Example IV.1 and its graph $G(\mathbf{V})$.

On the other hand, let $\{(\mathbf{M}_E^*(L), \mathbf{M}_D^*(L))\}_{L=1}^\infty$ be a sequence of optimal processings. By Theorem IV.2, every pair $(\mathbf{M}_E^*(L), \mathbf{M}_D^*(L))$ can be assumed to be binary matrices. Since there are only finitely many binary matrices, there exists an infinite subsequence $\{L_k\}_{k=1}^\infty$ such that all $(\mathbf{M}_E^*(L_k), \mathbf{M}_D^*(L_k))$ are identical (equal to $(\mathbf{M}_E^*, \mathbf{M}_D^*)$, say). Hence

$$\begin{aligned} \lim_{L \rightarrow \infty} \max_{\mathbf{M}_E, \mathbf{M}_D} \frac{1}{N} C((\mathbf{M}_E \mathbf{W} \mathbf{M}_D)^L) \\ = \lim_{k \rightarrow \infty} \frac{1}{N} C((\mathbf{M}_E^*(L_k) \mathbf{W} \mathbf{M}_D^*(L_k))^{L_k}) \\ = \lim_{k \rightarrow \infty} \frac{1}{N} C((\mathbf{M}_E^* \mathbf{W} \mathbf{M}_D^*)^{L_k}) \\ \leq \max_{\mathbf{M}_E, \mathbf{M}_D} \lim_{k \rightarrow \infty} \frac{1}{N} C((\mathbf{M}_E \mathbf{W} \mathbf{M}_D)^{L_k}) \end{aligned}$$

proving (25b). \square

Theorem IV.3 states that the zero-error capacity is the only part of the transmitted information rate that can be preserved as the length of the cascade $L \rightarrow \infty$. Another interpretation of this result can be obtained by considering the concept of common information [22]. The common information between the pair of random variables (X, Y) is defined as the entropy of the function $\xi = f(X) = g(Y)$ of maximum range. In other words, ξ can be computed from either X or Y with probability one. If X has distribution \mathbf{p} and Y has conditional distribution \mathbf{Q} , we denote the common information between them as

$$J(\mathbf{p}, \mathbf{Q}) \triangleq H(\xi).$$

As is shown in [23], [24], the zero-error capacity can be defined through common information in a similar manner as ordinary capacity can be defined through mutual information. More precisely

$$\log M_0(\mathbf{V}^{\otimes N}) = \max_{\mathbf{p}} J(\mathbf{p}, \mathbf{V}^{\otimes N}),$$

and thus Theorem IV.3 can equivalently be written as

$$\lim_{L \rightarrow \infty} C_{N,L}(\mathbf{V}) = \max_{\mathbf{p}} \frac{1}{N} J(\mathbf{p}, \mathbf{V}^{\otimes N}).$$

Hence, as $L \rightarrow \infty$, the only part of mutual information we can preserve between the input and the output of the cascade is exactly the common information between them.

The following examples illustrate the use of Theorems IV.2 and IV.3.

Example IV.1: Consider $\mathbf{V} = \mathbf{Q}$, where \mathbf{Q} is given in (11) in Example III.1. The corresponding transition matrix \mathbf{V} and graph $G(\mathbf{V})$ are depicted in Fig. 3.

For this channel

$$C_0(\mathbf{V}) = M_0(\mathbf{V}) = \log 2$$

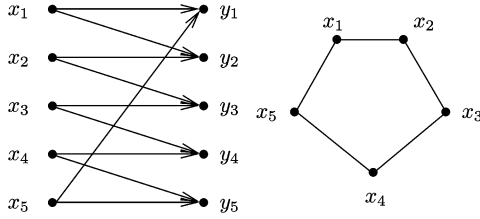


Fig. 4. The “pentagon” channel \mathbf{V} and its graph $G(\mathbf{V})$.

that is, the zero-error capacity is achieved by a zero-error code of block length one (this code for example might use x_1 and x_3 that are nonadjacent). Theorem IV.3 states that for an infinite cascade of these channels, for any finite N

$$\lim_{L \rightarrow \infty} C_{N,L}(\mathbf{V}) = \lim_{L \rightarrow \infty} C_{1,L}(\mathbf{V}) = \log 2.$$

In other words, if N is restricted to be finite, $N = 1$ is optimal.

The limiting capacity can be achieved by using at all intermediate nodes the decoder

$$\mathbf{M}_D = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}^T$$

followed by the encoder

$$\mathbf{M}_E = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

Notice that, for this specific example, the product $\mathbf{M}_D \mathbf{M}_E$ is actually equivalent to forwarding. In fact, a decoder cannot distinguish between x_1 and x_2 , and between x_3 and x_4 . Nonetheless, x_1 or x_2 are never mistaken for x_3 or x_4 . Moreover, in the limit for large L , \mathbf{M}_E can be used at the source and \mathbf{M}_D can be used at the destination without loss of optimality. Hence, for long line networks of channels \mathbf{V} , the intermediate nodes need not to perform any processing at all, the source conveys the bits with value zero by sending, for example, x_1 or x_2 , and the bits with value one by sending x_3 or x_4 . The destination decodes a bit to be zero if either y_1 or y_2 is received, and one otherwise. This scheme does not incur any delay, does not require any intermediate processing, and is optimal provided the length of intermediate processing is restricted to be finite while $L \rightarrow \infty$. \diamond

The simplest nontrivial DMC for which the zero-error capacity is achieved by a zero-error code of length larger than one is the so-called “pentagon” channel [7], which we analyze in more detail in the next example.

Example IV.2. The Pentagon Channel: Consider the “pentagon” channel whose transition matrix \mathbf{V} , for $p \in (0, 1)$, is

$$\mathbf{V} = \begin{pmatrix} 1-p & p & 0 & 0 & 0 \\ 0 & 1-p & p & 0 & 0 \\ 0 & 0 & 1-p & p & 0 \\ 0 & 0 & 0 & 1-p & p \\ p & 0 & 0 & 0 & 1-p \end{pmatrix}.$$

The corresponding graph $G(\mathbf{V})$ is depicted in Fig. 4.

For this channel, we have $M_0(\mathbf{V}) = 2$ (for example, x_1 and x_3 are nonadjacent) and $M_0(\mathbf{V}^{\otimes 2}) = 5$ (for example, (x_1, x_1) ,

(x_2, x_3) , (x_3, x_5) , (x_4, x_2) , and (x_5, x_4) are nonadjacent). It was conjectured in [7] and shown in [25] that for this channel the zero-error capacity is

$$C_0(\mathbf{V}) = \frac{1}{2} M_0(\mathbf{V}^{\otimes 2}) = \frac{1}{2} \log 5$$

that is, a zero-error code of block length two is optimal.

Theorem IV.3 states that for an infinite cascade of “pentagon” channels

$$\lim_{L \rightarrow \infty} C_{1,L}(\mathbf{V}) = \log 2,$$

and

$$\lim_{L \rightarrow \infty} C_{2,L}(\mathbf{V}) = \frac{1}{2} \log 5.$$

Moreover, for any other finite N , we can never achieve more than $\frac{1}{2} \log 5$. Hence, for an infinite cascade of “pentagon” channels, an intermediate processing of length $N = 2$ is optimal if N is restricted to be finite. The optimal limiting capacity can be achieved by using at all intermediate nodes and at the source a zero-error encoder of length 2, and at all intermediate nodes and at the destination a zero-error decoder of length 2. Notice that in this example, if the intermediate nodes simply forward the incoming data (i.e., they perform a *suboptimal* processing of length $N = 1$) then the limiting capacity is $\log D(\mathbf{V}) = \log 1 = 0$ (by Theorem III.4), and this limit is approached exponentially fast with exponent

$$-2 \log |\lambda_2(\mathbf{V})| = -\log(1 - 2p(1-p) \sin^2(2\pi/5))$$

(by Theorem III.5). In other words, intermediate processing is *necessary* if a nonvanishing rate is to be achieved in a long line network of pentagon channels, but even an optimal one-symbol processing suffices to achieve a strictly positive end-to-end rate. \diamond

V. BOUNDS ON CAPACITY

We derive an upper and a lower bound on $C_{N,L}(\mathbf{V})$, the capacity of the cascade with optimal intermediate processing at the relays, that apply for all values of N and L .

A. Upper Bound

In this subsection, we derive an upper bound for $C_{N,L}(\mathbf{V})$ expressed as a convex combination of the min-cut capacity and of a term reminiscent of the zero-error capacity. The basic idea is to decompose the channel transition matrix $\mathbf{V}^{\otimes N}$ into a convex combination of two stochastic matrices, one of which has rank as close as possible to $M_0(\mathbf{V}^{\otimes N})$. We also discuss efficient algorithms to determine such a decomposition.

Theorem V.1: For any stochastic matrix \mathbf{V} and any integer N , if there exist two stochastic matrices \mathbf{A}_N and \mathbf{B}_N , and $\delta_N \in (0, 1]$ such that

$$\mathbf{V}^{\otimes N} = \delta_N \mathbf{A}_N + (1 - \delta_N) \mathbf{B}_N \tag{26}$$

then

$$C_{N,L}(\mathbf{V}) \leq (1 - (1 - \delta_N)^{L-1}) \frac{\log \text{rank}(\mathbf{A}_N)}{N} + (1 - \delta_N)^{L-1} C(\mathbf{V}). \tag{27}$$

Remark: If $\log \text{rank}(\mathbf{A}_N)/N < C(\mathbf{V})$ then, for L large enough, the bound in (27) is strictly better than the min-cut upper bound $C_{N,L}(\mathbf{V}) \leq C(\mathbf{V})$.

Proof: Assume (26) holds, then

$$\begin{aligned} C_{N,L}(\mathbf{V}) &= \frac{1}{N} C \left(\mathbf{W} \prod_{i=1}^{L-1} (\mathbf{M}_i \mathbf{W}) \right) \\ &= \frac{1}{N} C \left((\delta_N \mathbf{A}_N + (1 - \delta_N) \mathbf{B}_N) \prod_{i=1}^{L-1} (\mathbf{M}_i \mathbf{W}) \right) \\ &\leq \delta_N \frac{1}{N} C \left(\mathbf{A}_N \prod_{i=1}^{L-1} (\mathbf{M}_i \mathbf{W}) \right) \\ &\quad + (1 - \delta_N) \frac{1}{N} C \left(\mathbf{B}_N \prod_{i=1}^{L-1} (\mathbf{M}_i \mathbf{W}) \right) \quad (28a) \end{aligned}$$

$$\leq \delta_N \frac{C(\mathbf{A}_N)}{N} + (1 - \delta_N) \frac{1}{N} C \left(\mathbf{W} \prod_{i=2}^{L-1} (\mathbf{M}_i \mathbf{W}) \right) \quad (28b)$$

where (28a) follows from the convexity of mutual information in the channel matrix, and (28b) from the data processing inequality. By repeating the same argument, we get

$$C_{N,L}(\mathbf{V}) \leq (1 - (1 - \delta_N)^{L-1}) \frac{C(\mathbf{A}_N)}{N} + (1 - \delta_N)^{L-1} C(\mathbf{V}). \quad (29)$$

We can further upper-bound $C(\mathbf{A}_N)$ in (29) with the logarithm of the rank of \mathbf{A}_N [19] to yield (27). \square

The following example illustrates the use of Theorem V.1.

Example V.1: Consider again $\mathbf{V} = \mathbf{Q}$, where \mathbf{Q} is given by (11) in Example III.1. Then, we can choose

$$\delta_1 \triangleq \min\{2p, 2(1-p), t\}$$

and $\mathbf{A}_1 \triangleq \mathbf{Q}^\infty$, where \mathbf{Q}^∞ is given by (12). In this case, $\log \text{rank}(\mathbf{A}_1) = \log(2) = C_0(\mathbf{V})$. Setting $\mathbf{A}_N \triangleq \mathbf{A}_1^{\otimes N}$, Theorem V.1 yields

$$C_{N,L}(\mathbf{V}) \leq (1 - (1 - \delta_1^N)^{L-1}) C_0(\mathbf{V}) + (1 - \delta_1^N)^{L-1} C(\mathbf{V}). \quad (30)$$

For this channel $\lim_{L \rightarrow \infty} C_{N,L}(\mathbf{V}) = C_0(\mathbf{V})$ for any finite N , as already pointed out in Example IV.1. Moreover, we see from (30) that the decay of $C_{N,L}(\mathbf{V})$ to $C_0(\mathbf{V})$ is exponential in L with exponent lower-bounded by

$$\liminf_{L \rightarrow \infty} -\frac{1}{L} \log(C_{N,L}(\mathbf{V}) - C_0(\mathbf{V})) \geq -\log(1 - \delta_1^N). \quad \diamond$$

In order to obtain the best bound for any given N , \mathbf{A}_N should be chosen to have the smallest rank possible. A possible choice is to take $\mathbf{A}_N \triangleq \mathbf{A}_1^{\otimes N}$ and $\delta_N \triangleq \delta_1^N$ as in Example V.1. With this the bound in (27) reduces to

$$C_{N,L}(\mathbf{V}) \leq \left(1 - (1 - \delta_1^N)^{L-1}\right) \log \text{rank}(\mathbf{A}_1) + (1 - \delta_1^N)^{L-1} C(\mathbf{V})$$

TABLE I
MAXIMUM INDEPENDENT SET PROBLEM, SET COVER PROBLEM, AND THEIR LP RELAXATIONS

Set Cover Problem	LP Relaxation (Primal)
$\min \sum_i s_i$ $\sum_{i:u \in \mathcal{S}_i} s_i \geq 1, \forall u \in \mathcal{U}$ $s_i \in \{0, 1\}, \forall i$	$\min \sum_i s_i$ $\sum_{i:u \in \mathcal{S}_i} s_i \geq 1, \forall u \in \mathcal{U}$ $0 \leq s_i \leq 1, \forall i$
Max Independent Set Problem	LP Relaxation (Dual)
$\max \sum_j t_j$ $\sum_{j \in \mathcal{S}_i} t_j \leq 1 \forall \mathcal{S}_i \in \mathcal{S}$ $t_j \in \{0, 1\}, \forall j$	$\max \sum_j t_j$ $\sum_{j \in \mathcal{S}_i} t_j \leq 1 \forall \mathcal{S}_i \in \mathcal{S}$ $0 \leq t_j \leq 1, \forall j$

since $\text{rank}(\mathbf{A}_1^{\otimes N}) = (\text{rank}(\mathbf{A}_1))^N$. However, the choice $\mathbf{A}_N = \mathbf{A}_1^{\otimes N}$ does not give the best possible bound for $C_{N,L}(\mathbf{V})$ in (27) in general, as we show in Example V.2.

Example V.2: Consider again the ‘‘pentagon’’ channel introduced in Example IV.2. For $N = 1$ we can find a matrix \mathbf{A}_1 with $\text{rank}(\mathbf{A}_1) = 3$. However, for $N = 2$ we can find a matrix \mathbf{A}_2 with $\text{rank}(\mathbf{A}_2) = 8 < \text{rank}(\mathbf{A}_1)^2 = 9$. \diamond

Note that for any matrix \mathbf{A}_N such that (26) holds, we have

$$\frac{1}{N} \log M_0(\mathbf{V}^{\otimes N}) = \lim_{L \rightarrow \infty} C_{N,L}(\mathbf{V}) \leq \frac{1}{N} \log \text{rank}(\mathbf{A}_N).$$

If for some N we find $\text{rank}(\mathbf{A}_N) = M_0(\mathbf{V}^{\otimes N})$ (like⁽³¹⁾ in Example V.1), then Theorem V.1 implies that the decay of $C_{N,L}(\mathbf{V})$ to $\lim_{L \rightarrow \infty} C_{N,L}(\mathbf{V})$ is exponentially fast in L . We have already seen in Section III-C that if we impose the constraint that all $\{\mathbf{M}_i\}_{i=1}^{L-1}$ are identical and independent of L then the limiting capacity is approached exponentially fast in L , and we gave tight upper and lower bounds on the exponent. If $\text{rank}(\mathbf{A}_N) = M_0(\mathbf{V}^{\otimes N})$ for some N , then exponential decay also applies to nonidentical processing. The lower bound on the exponent that can be derived from Theorem V.1, namely, $-\log(1 - \delta_N)$, is, however, not tight in general.

The problem of finding the matrix \mathbf{A}_N with minimum rank is equivalent to the *Set Cover Problem* described as follows [26]. Given a universe \mathcal{U} of n elements, a collection $\mathcal{S} = \{\mathcal{S}_1 \dots \mathcal{S}_m\}$ of subsets of \mathcal{U} , and a cost function for each subset in \mathcal{S} , find a minimum cost subcollection of \mathcal{S} that covers all the elements in \mathcal{U} . This problem can be formulated as an integer program. Assign a variable s_i for each set $\mathcal{S}_i \in \mathcal{S}$, where $s_i = 1$ if set \mathcal{S}_i takes part in the subcollection and $s_i = 0$ otherwise. The constraint is that every element in \mathcal{U} must belong to at least one of the picked sets \mathcal{S}_i . The set cover problem and its linear program (LP) relaxation (primal) are provided in Table I, for the special case where the cost of all sets is one, which is the case of interest here. A variety of approximation algorithms are available in the literature for the set cover problem [26]. These algorithms run in polynomial time in n and provide approximations with gap at most $\log n$ from the optimal solution.

In our case, the universe \mathcal{U} is the set of $n = |\mathcal{X}|^N$ inputs of the channel $\mathbf{V}^{\otimes N}$. The set of $m = |\mathcal{Y}|^N$ outputs defines \mathcal{S} , in that the subset \mathcal{S}_i contains the inputs that result with nonzero probability in output i , for $i = 1, \dots, m$. The next theorem shows that this set cover problem is equivalent to the problem of finding the matrix \mathbf{A}_N of minimum rank.

Proposition V.2: The minimum rank matrix \mathbf{A}_N has rank $\text{OPT}_{\mathbf{A}_N}$ if and only if $\text{OPT}_{\mathbf{A}_N}$ is the minimum cost of the previously described set cover problem.

Proof: Consider a solution of the set cover problem with cost OPT_{A_N} . This provides us with a set of outputs that cover all inputs. We can construct a matrix \mathbf{A}_N that has rank OPT_{A_N} as follows. Take the binary matrix $\lceil \mathbf{V}^{\otimes N} \rceil$, where the ceiling operation is component-wise. To construct matrix \mathbf{A}_N , replace with the all-zero columns those columns of $\lceil \mathbf{V}^{\otimes N} \rceil$ that correspond to outputs where $s_i = 0$ in the set cover problem. Then normalize all rows so that they sum to one. The matrix \mathbf{A}_N contains at most OPT_{A_N} linearly independent nonzero columns and $m - \text{OPT}_{A_N}$ all-zero columns, and thus, $\text{rank}(\mathbf{A}_N) \leq \text{OPT}_{A_N}$. Moreover, \mathbf{A}_N is a valid stochastic matrix because its nonzero columns “cover” all inputs, i.e., each row has at least one nonzero element, and all rows sum to one. Choose δ_N to be the largest number such that the matrix

$$\mathbf{B}_N = \frac{1}{1 - \delta_N} (\mathbf{V}^{\otimes N} - \delta_N \mathbf{A}_N)$$

is a valid stochastic matrix. It is easy to see that this δ_N is at least as large as the minimum of the entries of $\mathbf{V}^{\otimes N}$ that correspond to a nonzero entry in \mathbf{A}_N . We conclude that $\min_{\mathbf{A}_N} \text{rank}(\mathbf{A}_N) \leq \text{OPT}_{A_N}$.

For the reverse direction, note first that by nonnegativity of \mathbf{A}_N , we can assume without loss of generality that the columns of \mathbf{A}_N are linearly independent (linearly dependent columns do not affect the rank, and correspond to sets with already covered elements). Set $s_i = 1$ if the i th column of \mathbf{A}_N is nonzero and $s_i = 0$ otherwise. Since \mathbf{A}_N is a stochastic matrix, this defines a valid cover of \mathcal{U} with cost equal to $\text{rank}(\mathbf{A}_N)$. Therefore, $\min_{\mathbf{A}_N} \text{rank}(\mathbf{A}_N) \geq \text{OPT}_{A_N}$. \square

By using one of the available approximation algorithms, we can calculate in polynomial time in $n = |\mathcal{X}|^N$ a matrix \mathbf{A}_N that has rank bigger than the minimum by a factor of at most $\log n = N \log |\mathcal{X}|$. As we are really interested in $\frac{1}{N} \log \text{rank}(\mathbf{A}_N)$ this implies that the loss we incur by using these approximation algorithms to find an \mathbf{A}_N matrix goes to zero as $\frac{1}{N} \log(N)$ as $N \rightarrow \infty$.

The problem of finding the matrix \mathbf{A}_N with minimum rank is closely related to computing $M_0(\mathbf{V}^{\otimes N})$. To see this, consider the (strong) LP dual of the set cover LP relaxation described in Table I. The dual LP is the LP relaxation of the *Maximum Independent Set Problem*. The maximum independent set problem takes as input a graph adjacency matrix and calculates the graph independence number. We already pointed out in Section II-A that $M_0(\mathbf{V}^{\otimes N})$ is the independence number of the graph $G(\mathbf{V}^{\otimes N})$. This maximum independent set problem can be formulated as an integer program as follows. Assign a variable t_j for each vertex of the graph, $t_j = 1$ if the vertex takes part in the independent set and $t_j = 0$ otherwise. The constraint is that no two picked vertices are connected with an edge.

For our purposes, we use the adjacency matrix corresponding to the graph $G(\mathbf{V}^{\otimes N})$. The solution of the maximum independent set problem for $G(\mathbf{V}^{\otimes N})$ directly leads to the construction of an optimal zero-error code of length N for the channel \mathbf{V} . Denote by OPT_{M_0} this optimal solution. Obviously, OPT_{A_N} is lower-bounded by OPT_{M_0} . In fact, OPT_{A_N} is the minimum number of outputs such that all inputs are covered. Since all inputs are covered, this implies that any $\text{OPT}_{A_N} + 1$ inputs

have at least one output in common. Thus, $\text{OPT}_{M_0} \leq \text{OPT}_{A_N}$, yielding another proof of (31).

The integrality gap between the set cover problem and its LP relaxation is subexponential in N [26]. Hence, if the integrality gap between the maximum independent set problem and its LP relaxation is also subexponential in N , then there exists a matrix \mathbf{A}_N such that $\frac{1}{N} \log \text{rank}(\mathbf{A}_N) \approx \frac{1}{N} \log M_0(\mathbf{V}^{\otimes N})$, i.e., the inequality in (31) becomes an equality.

B. Lower Bound

We next derive a lower bound on $C_{N,L}(\mathbf{V})$. To do so, we choose a particular (possibly suboptimal) communication scheme and find a lower bound on the rate achievable with this scheme. Assume we use an inner encoder $\mathbf{M}_E \in \mathcal{S}_{M \times |\mathcal{X}|^N}$ (for some $M \in \mathbb{N}$) at the source A_0 and a corresponding maximum-likelihood decoder $\mathbf{M}_D \in \mathcal{S}_{|\mathcal{Y}|^N \times M}$ at the relay A_1 . At A_1 the message is then re-encoded using again \mathbf{M}_E and transmitted over the second channel. We continue in the same manner at every A_i , and the destination A_L uses an inner decoder \mathbf{M}_D . This corresponds to using intermediate processing $\mathbf{M}_i = \mathbf{M}_D \mathbf{M}_E$ for all i . The rate r of this inner code is determined by M , through

$$r \triangleq \frac{1}{N} \log(M).$$

Note that this scheme constructs an overall channel between the source and the destination node

$$(\mathbf{M}_E \mathbf{W} \mathbf{M}_D)^L = \mathbf{Q}^L$$

with $\mathbf{Q} \triangleq \mathbf{M}_E \mathbf{W} \mathbf{M}_D \in \mathcal{S}_{M,M}$. The source A_0 and the destination A_L then use an outer code over the channel \mathbf{Q}^L .

Using random coding arguments, we know that there exist good codes (defined by the tuple $(\mathbf{M}_E, \mathbf{M}_D)$) in the sense that the average probability of decoding error is bounded by

$$P_e(r) \leq \exp(-N E_{\mathcal{R}}(r)) \triangleq \delta(r)$$

where $E_{\mathcal{R}}(\cdot)$ is the random coding error exponent for the channel \mathbf{V} as a function of the rate r [27, Theorem 5.6.2]. We use such a good code as our inner code. With this, the trace of \mathbf{Q} is lower-bounded by $M(1 - \delta(r))$, but unfortunately this result gives no information about the off-diagonal entries of \mathbf{Q} . To get a lower bound on the achievable rate, we construct the worst cascade \mathbf{Q}^L such that \mathbf{Q} satisfies the trace constraint. This worst cascade is found in Appendix III Lemma III.1. The resulting lower bound for $C_{N,L}(\mathbf{V})$ is given in the next theorem.

Theorem V.3:

$$C_{N,L}(\mathbf{V}) \geq \sup_{r \in [0, C(\mathbf{V})]} r (1 - \delta(r))^L - \frac{1}{N} \quad (32)$$

with equality as $N \rightarrow \infty$, i.e.,

$$\lim_{N \rightarrow \infty} \sup_{r \in [0, C(\mathbf{V})]} r (1 - \delta(r))^L - \frac{1}{N} = C(\mathbf{V}). \quad (33)$$

Proof: Let $\mathcal{W}_1(a)$ be the set of stochastic matrices \mathbf{Q}^L such that $\text{tr}(\mathbf{Q}) \geq Ma$, i.e.,

$$\mathcal{W}_1(a) \triangleq \left\{ \widetilde{\mathbf{W}} = \mathbf{Q}^L : \mathbf{Q} \in \mathcal{S}_{M,M}, \text{tr}(\mathbf{Q}) \geq Ma \right\}$$

and $\mathcal{W}_2(a)$ be the set of stochastic matrices \mathbf{Q} such that $\text{tr}(\mathbf{Q}) \geq Ma^L$, i.e.,

$$\mathcal{W}_2(a) \triangleq \left\{ \tilde{\mathbf{W}} \in \mathcal{S}_{M,M} : \text{tr}(\tilde{\mathbf{W}}) \geq Ma^L \right\}.$$

For any $\mathbf{Q} \in \mathcal{S}_{M,M}$ with $\text{tr}(\mathbf{Q}) \geq Ma$, we have by Jensen's inequality

$$\text{tr}(\mathbf{Q}^L) = \sum_{i_1, \dots, i_{L-1}, j} q_{ji_1} q_{i_1 i_2} \dots q_{i_{L-1} j} \geq \sum_j q_{jj}^L \geq Ma^L$$

where $q_{ij} = [\mathbf{Q}]_{i,j}$. Hence, $\mathcal{W}_1(a) \subset \mathcal{W}_2(a)$ and therefore

$$\begin{aligned} C_{N,L}(V) &\geq \frac{1}{N} \inf_{\tilde{\mathbf{W}} \in \mathcal{W}_1(1-P_e(r))} \max_{\mathbf{p}} I(\mathbf{p}, \tilde{\mathbf{W}}) \\ &\geq \frac{1}{N} \min_{\tilde{\mathbf{W}} \in \mathcal{W}_2(1-P_e(r))} \max_{\mathbf{p}} I(\mathbf{p}, \tilde{\mathbf{W}}) \\ &\geq \frac{1}{N} \min_{\tilde{\mathbf{W}} \in \mathcal{W}_2(1-P_e(r))} I(\mathbf{p}, \tilde{\mathbf{W}}) \end{aligned} \quad (34)$$

for any \mathbf{p} . In particular, let \mathbf{p} be the uniform distribution on $\{1, \dots, M\}$. The minimization in (34) can be solved using Appendix III Lemma III.1 and the example following it. If $(1 - P_e(r))^L \leq 1/M$, the worst channel is $\frac{1}{M} \mathbf{1}\mathbf{1}^T$, and capacity is (trivially) lower-bounded by 0. Else, the worst channel has diagonal entries equal to $(1 - P_e(r))^L$ and all other entries equal to $(1 - (1 - P_e(r))^L)/(M - 1)$. In both cases, the minimizer is a symmetric channel, for which the uniform distribution is capacity achieving and the capacity is

$$\log(M) - H(\tilde{\mathbf{w}}_1)$$

where $\tilde{\mathbf{w}}_1$ is the first row of the minimizing channel matrix $\tilde{\mathbf{W}}$ [1, Theorem 8.2.1]. Thus, the lower bound in (34) can be calculated as

$$\begin{aligned} \frac{1}{N} \left(\log(M) - (1 - (1 - P_e(r))^L) \log(M - 1) \right. \\ \left. - \mathcal{H}((1 - P_e(r))^L) \right) \geq r(1 - P_e(r))^L - \frac{1}{N} \end{aligned} \quad (35)$$

where the above bound is valid for all values of the inner code rate r such that

$$(1 - P_e(r))^L \geq 1/M.$$

If $(1 - P_e(r))^L \leq 1/M$ then (35) becomes

$$\frac{\log(M)}{N} (1 - P_e(r))^L - \frac{1}{N} \leq \frac{1}{N} \left(\frac{\log(M)}{M} - 1 \right) \leq 0.$$

Hence, (35) is also (trivially) valid in this range, and we can maximize over $r \in [0, C(\mathbf{V})]$, thus, showing that

$$C_{N,L}(\mathbf{V}) \geq \sup_{r \in [0, C(\mathbf{V})]} r(1 - P_e(r))^L - \frac{1}{N} \quad (36)$$

$$\geq \sup_{r \in [0, C(\mathbf{V})]} r(1 - \delta(r))^L - \frac{1}{N}. \quad (37)$$

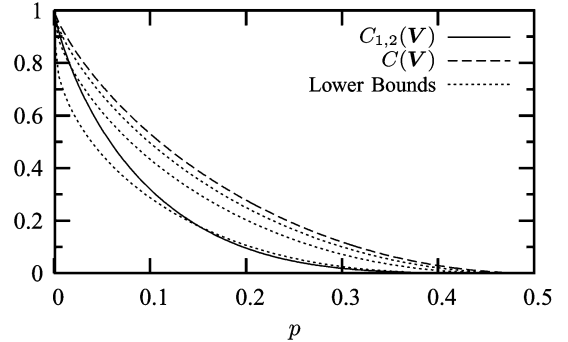


Fig. 5. Lower bounds on capacity of the cascade of two BSC(p) for different values of block length $N \in \{10^2, 10^3, 10^4\}$ (from bottom to top in the figure). Also shown are the min-cut upper bound $C(\mathbf{V})$ and $C_{1,2}(\mathbf{V})$. Note that the lower bound derived in this section is not very good for small values of N . Indeed, for $N = 100$ the forwarding lower bound yields better results for some values of p . The bound becomes, however, tight as $N \rightarrow \infty$ and converges to the min-cut capacity.

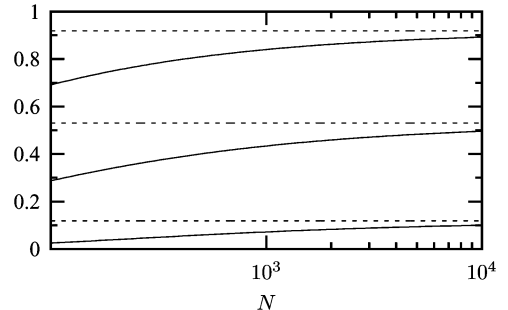


Fig. 6. Lower bounds on capacity of the cascade of two BSC(p) for different values of crossover probability $p \in \{0.01, 0.1, 0.3\}$ (from top to bottom in the figure, solid lines) as a function of the block length N of intermediate processing. Also shown are the corresponding limiting expressions $C(\mathbf{V})$ (dashed lines).

We now show that the bound is tight as $N \rightarrow \infty$. As $E_{\mathcal{R}}(r)$ is strictly positive for all $r < C(\mathbf{V})$ [27, Theorem 5.6.4], we have for every $\varepsilon > 0$ and with $r = C(\mathbf{V}) - \varepsilon$

$$\begin{aligned} \lim_{N \rightarrow \infty} (C(\mathbf{V}) - \varepsilon) (1 - \exp(-NE_{\mathcal{R}}(C(\mathbf{V}) - \varepsilon)))^L - \frac{1}{N} \\ = C(\mathbf{V}) - \varepsilon. \end{aligned}$$

As ε is arbitrary, (33) follows.

Example V.3: Consider a cascade of $L = 2$ BSC(p). Fig. 5 compares for this channel the lower bound derived in this section with $C_{1,2}(\mathbf{V})$ and the min-cut capacity $C(\mathbf{V})$ (achievable when $N \rightarrow \infty$). It can be seen that, while the lower bound is not very good for small values of N , it is quite good for larger values of N and tight as $N \rightarrow \infty$.

How fast the convergence of the lower bound to the upper bound $C(\mathbf{V})$ takes place, is depicted in Fig. 6 as a function of N and for various values of crossover probability p . It can be seen that the lower bound is already quite close to the limiting expression for $N \geq 10^4$. \diamond

VI. SCALING LAWS

In this section, we investigate how the block length N needs to scale with the network length L in order to achieve a constant fraction of the min-cut capacity (as opposed to the zero-error

capacity). We will show that logarithmic growth of N with L is sufficient, and in many cases also necessary.

From (2), any rate strictly below the zero-error capacity can be achieved with finite block length processing. Here we are interested in rates that indeed need an infinite block length. Hence, for any $\alpha \in [0, 1]$ we define

$$N^*(L, \alpha) \triangleq \inf\{N \in \mathbb{N} : C_{N,L}(\mathbf{V}) \geq R(\alpha)\}$$

where

$$R(\alpha) \triangleq (1 - \alpha)C_0(\mathbf{V}) + \alpha C(\mathbf{V}).$$

The next theorem gives an upper bound on $N^*(L, \alpha)$, establishing that logarithmic growth of N with L is sufficient to achieve $R(\alpha)$.

Theorem VI.1: For all $\alpha \in [0, 1]$

$$N^*(L, \alpha) \leq \inf_{\varepsilon > 0} \left(1 + \max \left\{ \frac{1}{\varepsilon}, \inf_{r \in (R(\alpha) + \varepsilon, C(\mathbf{V}))} \frac{1}{E_{\mathcal{R}}(r)} \log \left(\frac{L}{1 - \frac{R(\alpha) + \varepsilon}{r}} \right) \right\} \right). \quad (38)$$

Proof: Theorem V.3 asserts that for any $r \in [0, C(\mathbf{V})]$

$$\begin{aligned} C_{N,L}(\mathbf{V}) &\geq r(1 - \exp(-NE_{\mathcal{R}}(r)))^L - \frac{1}{N} \\ &\geq r(1 - L \exp(-NE_{\mathcal{R}}(r))) - \frac{1}{N}, \end{aligned}$$

by using the fact that $(1 - x)^L \geq 1 - Lx$ for all $x \in [0, 1]$. Thus

$$\begin{aligned} N^*(L, \alpha) &\leq \inf \left\{ N \in \mathbb{N} : r(1 - L \exp(-NE_{\mathcal{R}}(r))) - \frac{1}{N} \geq R(\alpha) \right\} \\ &\leq \inf \left\{ N \in \mathbb{N} : N \geq \frac{1}{\varepsilon}, \right. \\ &\quad \left. r(1 - L \exp(-NE_{\mathcal{R}}(r))) - \varepsilon \geq R(\alpha) \right\} \\ &\leq \left[\max \left\{ \frac{1}{\varepsilon}, \frac{1}{E_{\mathcal{R}}(r)} \log \left(\frac{L}{1 - \frac{R(\alpha) + \varepsilon}{r}} \right) \right\} \right] \\ &\leq 1 + \max \left\{ \frac{1}{\varepsilon}, \frac{1}{E_{\mathcal{R}}(r)} \log \left(\frac{L}{1 - \frac{R(\alpha) + \varepsilon}{r}} \right) \right\}. \end{aligned}$$

Since this is true for all $r \in (R(\alpha) + \varepsilon, C(\mathbf{V}))$, we can minimize over r to get the tightest bound. \square

Notice that by solving the bound of Theorem V.3 (see (36)) for L , we get also that for fixed N to guarantee a rate R we can cascade up to

$$L \leq \sup_{r \in [R + \frac{1}{N}, C(\mathbf{V})]} \frac{1}{P_e(r)} \left(1 - \frac{R + \frac{1}{N}}{r} \right)$$

channels. That is, for a fixed r , R , and N , the number of channels cascaded scales at least like $1/P_e(r)$. This last result is interesting as it recovers, as a special case, the same $1/P_e(r)$

scaling derived in [10] for cascades of binary Z-channels with repetition coding at the relays.

The next theorem establishes that logarithmic growth of N with L is necessary to achieve $R(\alpha)$ for all $\alpha > \beta_m$ where β_m is a nonnegative constant.

Theorem VI.2: Let $m \in \mathbb{N}$, $\mathbf{A}_m, \mathbf{B}_m \in \mathcal{S}_{|\mathcal{X}|^m, |\mathcal{Y}|^m}$, and $\delta_m \in (0, 1]$ such that

$$\mathbf{V}^{\otimes m} = \delta_m \mathbf{A}_m + (1 - \delta_m) \mathbf{B}_m$$

and $\frac{1}{m} \log \text{rank}(\mathbf{A}_m) \leq C(\mathbf{V})$. Then

$$N^*(L, \alpha) \geq \frac{\log(L - 1) - \log \log \frac{1}{\alpha - \beta_m}}{\log \frac{1}{\delta_m}}$$

for all

$$\alpha > \beta_m \triangleq \frac{\frac{1}{m} \log \text{rank}(\mathbf{A}_m) - C_0(\mathbf{V})}{C(\mathbf{V}) - C_0(\mathbf{V})}. \quad (39)$$

Proof: From Theorem V.1 (for $N = km$ for some $k \in \mathbb{N}$)

$$\begin{aligned} C_{N,L}(\mathbf{V}) &\leq (1 - \delta_m^{N/m})^{L-1} \left(C(\mathbf{V}) - \frac{1}{m} \log \text{rank}(\mathbf{A}_m) \right) \\ &\quad + \frac{1}{m} \log \text{rank}(\mathbf{A}_m) \\ &\leq \exp\left(- (L - 1) \delta_m^{N/m}\right) (C(\mathbf{V}) - C_0(\mathbf{V})) \\ &\quad + \frac{1}{m} \log \text{rank}(\mathbf{A}_m). \end{aligned}$$

Now as $C_{N,L}(\mathbf{V}) \leq C_{mN,L}(\mathbf{V})$, we have

$$\begin{aligned} N^*(L, \alpha) &= \inf \left\{ \tilde{N} \in \mathbb{N} : C_{\tilde{N},L}(\mathbf{V}) \geq R(\alpha) \right\} \\ &\geq \frac{1}{m} \inf \left\{ N \in \mathbb{N} : C_{N,L}(\mathbf{V}) \geq R(\alpha) \right\} \\ &\geq \frac{1}{m} \inf \left\{ N \in \mathbb{R} : \exp\left(- (L - 1) \delta_m^{N/m}\right) \geq \alpha - \beta_m \right\} \\ &= \frac{\log(L - 1) - \log \log \frac{1}{\alpha - \beta_m}}{\log \frac{1}{\delta_m}}. \quad \square \end{aligned}$$

For every $m \in \mathbb{N}$, let \mathbf{A}_m^* be the matrix of minimal rank among all decompositions of $\mathbf{V}^{\otimes m}$ in Theorem VI.2. Define

$$\beta \triangleq \min \left\{ 1, \liminf_{m \rightarrow \infty} \frac{\frac{1}{m} \log \text{rank}(\mathbf{A}_m^*) - C_0(\mathbf{V})}{C(\mathbf{V}) - C_0(\mathbf{V})} \right\}.$$

We obtain the following corollary of Theorems VI.1 and VI.2.

Corollary VI.3: As $L \rightarrow \infty$, for $\alpha \in [0, 1]$

$$N^*(L, \alpha) = \mathcal{O}(\log L)$$

and for $\alpha \in (\beta, 1]$

$$N^*(L, \alpha) = \Omega(\log L).$$

Example VI.1: Consider again the ‘‘pentagon’’ channel introduced in Example IV.2. Theorem VI.1 asserts that logarithmic growth of N with L is sufficient to achieve rate $R(\alpha)$ for any $\alpha \in [0, 1]$. In Example V.2, we showed that a decomposition

with $\log \text{rank}(\mathbf{A}_2) = 8$ exists. Here $C(\mathbf{V}) = \log 5 - \mathcal{H}(p)$, $C_0(\mathbf{V}) = \frac{1}{2} \log 5$, and hence

$$\beta_2 = \frac{\frac{1}{2} \log \text{rank}(\mathbf{A}_2) - C_0(\mathbf{V})}{C(\mathbf{V}) - C_0(\mathbf{V})} = \frac{\log 8/5}{\log 5 - 2\mathcal{H}(p)}.$$

For example, for $p = 0.01$, we get $\beta_2 = 0.314$, showing that logarithmic growth of N with L is necessary to achieve any $R(\alpha)$ with $0.314 < \alpha \leq 1$. With $m = 2$, Theorem VI.2 does not in this case allow to state that logarithmic growth is necessary in the range $\alpha \leq 0.314$. \diamond

Example VI.2. Scaling Laws for Line Networks of Binary Symmetric Channels: Consider a cascade of L BSC(p) as in Example II.1. We know that any finite-length processing performed at every node in the network will result in a zero end-to-end rate as $L \rightarrow \infty$.

From Theorem VI.1, we see that logarithmic growth of N with L is *sufficient* to achieve any fraction $\alpha \in [0, 1)$ of the min-cut capacity. We now show that logarithmic growth is also *necessary* to achieve any positive fraction $\alpha \in (0, 1]$ of the min-cut capacity. Decompose \mathbf{V} as

$$\mathbf{V} = p\mathbf{1}\mathbf{1}^T + (1 - 2p)\mathbf{B}$$

for some stochastic matrix $\mathbf{B} \in \mathcal{S}_{2,2}$, and where $\mathbf{1}$ is the all-one column vector of length 2. Applying Theorem VI.2 with $m = 1$, we get that

$$N^*(L, \alpha) \geq \frac{\log(L-1) - \log \log \frac{1}{\alpha}}{\log \frac{1}{\delta}}$$

with $\delta \triangleq 2p$. Hence, logarithmic growth of N with L is necessary for all positive rates. \diamond

VII. DISCUSSION AND CONCLUSION

In this paper, we investigated communication through a cascade of L channels, where intermediate nodes can perform processing of finite complexity (measured by the processing length N).

When N is fixed, L goes to infinity, and all relays use the same processing independent of L , we showed that the limiting capacity can be easily computed as the logarithm of the number of eigenvalues of maximum modulus of the equivalent channel transition probability matrix that comprises the intermediate processing as part of the channel. We also showed how the rate of decay to the limiting capacity is related to the second largest eigenvalue modulus of the equivalent channel transition probability matrix.

When N is fixed and L goes to infinity, we showed that the optimal finite complexity processing is identical at each relay and corresponds to an optimal zero-error code of block length N for the underlying channel. The resulting limiting capacity is then the rate of this zero-error code and can never exceed the zero-error capacity of the underlying channel.

We also derived bounds on the capacity of finite length cascades and used them to show that logarithmic growth of N with L is sufficient to achieve any constant fraction of the min-cut capacity above the zero-error capacity. Moreover, we showed

that for rates above some threshold, logarithmic growth is also necessary. We conjecture that logarithmic growth is in fact necessary to achieve any fraction of the min-cut capacity above the zero-error capacity. To prove our conjecture, a tighter upper bound on capacity is needed.

The fact that for a fixed intermediate processing the decay to the limiting capacity is exponential contrasts the logarithmic scaling law found for the optimal N as a function of L and emphasizes the importance of a well-chosen intermediate processing at the relays.

In this work, we did not address the problem of identifying the optimal processing for any finite pair (N, L) , which is very interesting and combinatorial in nature. We have also restricted attention to cascades of identical channels. The problem of optimal communication over a nonhomogeneous cascade of channels is in general as difficult as the problem of finding the optimal processing for finite (N, L) discussed above. To see this, consider a cascade consisting of L identical channels, followed by an infinite number of noiseless channels. Finding the optimal processing reduces to finding the optimal processing for the initial L channels.

We view this work as a first step toward a more comprehensive understanding of how we should efficiently use limited network resources to achieve reliable communication. Extending this work for more general (other than line) networks, and more general traffic configurations and resource constraints is part of ongoing work [28], [29].

APPENDIX I

UNIFORM CONTINUITY OF CAPACITY

Lemma I.1: $C(\cdot)$ is uniformly continuous.

Proof: Mutual information $I : \mathcal{S}_{1,|\mathcal{X}|} \times \mathcal{S}_{|\mathcal{X}|,|\mathcal{Y}|} \rightarrow \mathbb{R}$ is continuous, and, as its domain is compact, also uniformly continuous [30, Theorem 4.19].

Consider now

$$|\max_{\mathbf{p}} I(\mathbf{p}, \mathbf{Q}) - \max_{\tilde{\mathbf{p}}} I(\tilde{\mathbf{p}}, \tilde{\mathbf{Q}})|.$$

By uniform continuity of mutual information, for any $\varepsilon > 0$ there exists $\delta > 0$ such that for all \mathbf{p} and all $\|\mathbf{Q} - \tilde{\mathbf{Q}}\|_\infty \leq \delta$ we have

$$|I(\mathbf{p}, \mathbf{Q}) - I(\mathbf{p}, \tilde{\mathbf{Q}})| \leq \varepsilon.$$

Assume without loss of generality that

$$\max_{\mathbf{p}} I(\mathbf{p}, \mathbf{Q}) \geq \max_{\tilde{\mathbf{p}}} I(\tilde{\mathbf{p}}, \tilde{\mathbf{Q}}).$$

Then

$$\begin{aligned} & |\max_{\mathbf{p}} I(\mathbf{p}, \mathbf{Q}) - \max_{\tilde{\mathbf{p}}} I(\tilde{\mathbf{p}}, \tilde{\mathbf{Q}})| \\ &= \max_{\mathbf{p}} \left(I(\mathbf{p}, \tilde{\mathbf{Q}}) + I(\mathbf{p}, \mathbf{Q}) - I(\mathbf{p}, \tilde{\mathbf{Q}}) \right) - \max_{\tilde{\mathbf{p}}} I(\tilde{\mathbf{p}}, \tilde{\mathbf{Q}}) \\ &\leq \varepsilon, \end{aligned}$$

thus proving the theorem. \square

APPENDIX II
PROOF OF THEOREM III.5

The proof of Theorem III.5 is broken up into two lemmas. The first lemma proves the lower bound in Theorem III.5. For both of these lemmas we will use some common initial steps that we will now review.

Assume without loss of generality that \mathbf{Q} is in the canonical form (7) with diagonal irreducible submatrices $\{\mathbf{P}_i\}$ with periods $\{d_i\}$. Recall that $D(\mathbf{Q}) = \sum_i d_i$ is the number of eigenvalues of modulus one of the matrix \mathbf{Q} . For simplicity of notation, we will use $D \triangleq D(\mathbf{Q})$ in the following. Call d the least common multiple of the $\{d_i\}$, and let

$$\tilde{\mathbf{Q}} \triangleq \mathbf{Q}^d. \tag{40}$$

By construction, $\tilde{\mathbf{Q}}$ has D primitive (irreducible and aperiodic) submatrices $\{\tilde{\mathbf{P}}_i\}_{i=1}^D$, with corresponding stationary distributions $\{\tilde{\pi}_i\}_{i=1}^D$, and thus $D(\tilde{\mathbf{Q}}) = D(\mathbf{Q}) = D$. From Lemma III.1

$$\tilde{\mathbf{Q}}^\ell = \tilde{\mathbf{Q}}^\infty + \mathbf{E}^{(\ell)}. \tag{41}$$

From Lemma III.2, the components of $\mathbf{E}^{(\ell)}$ converge to zero exponentially fast in ℓ with exponent not smaller than

$$-\log \left| \lambda_2(\mathbf{Q}^d) \right| = -d \log |\lambda_2(\mathbf{Q})|. \tag{42}$$

We will also use that, as $C(\mathbf{Q}^L)$ is decreasing in L , we have

$$C(\tilde{\mathbf{Q}}^{\lceil L/d \rceil}) \leq C(\mathbf{Q}^L) \leq C(\tilde{\mathbf{Q}}^{\lfloor L/d \rfloor}) \tag{43}$$

and that, from Theorem III.4

$$\lim_{L \rightarrow \infty} C(\mathbf{Q}^L) = C(\tilde{\mathbf{Q}}^\infty) = \log D. \tag{44}$$

Lemma II.1: For any square stochastic matrix

$$\liminf_{L \rightarrow \infty} E_L(\mathbf{Q}) \geq -\log |\lambda_2(\mathbf{Q})|$$

where $E_L(\mathbf{Q}) \triangleq -\frac{1}{L} \log (C(\mathbf{Q}^L) - \log D)$.

Proof: We start from the initial steps (40)–(44), and attempt to bound the quantity

$$\log \left(C(\tilde{\mathbf{Q}}^\ell) - \log D \right).$$

To do so, we will bound $C(\tilde{\mathbf{Q}}^\ell)$ using an upper bound on capacity from [31], that states

$$C(\mathbf{P}) \leq \log \sum_y \max_x p(y|x).$$

Let $\mathbf{P} = \tilde{\mathbf{Q}}^\ell$ and use (41). It is easily seen that the maximum entry of every column of $\tilde{\mathbf{Q}}^\infty$ lies in the corresponding diagonal block. Moreover, the exponents of the entries in $\mathbf{E}^{(\ell)}$ are all lower-bounded by (42). Thus, for some subexponential⁴ $b(\ell) \geq 0$

$$C(\tilde{\mathbf{Q}}^\ell) \leq \log (D + b(\ell) \exp (d\ell \log |\lambda_2(\mathbf{Q})|))$$

⁴We call a function $b(\ell)$ subexponential if $\lim_{\ell \rightarrow \infty} \frac{1}{\ell} \log b(\ell) = 0$.

where we used the fact that the stationary distribution of each diagonal block of $\tilde{\mathbf{Q}}$ sums to one. Then

$$\begin{aligned} & \log \left(C(\tilde{\mathbf{Q}}^\ell) - \log D \right) \\ & \leq \log \left(\log (D + b(\ell) \exp (d\ell \log |\lambda_2(\mathbf{Q})|)) - \log D \right) \\ & = \log \left(\log \left(1 + \tilde{b}(\ell) \exp (d\ell \log |\lambda_2(\mathbf{Q})|) \right) \right) \\ & \leq \log \left(\tilde{b}(\ell) \exp (d\ell \log |\lambda_2(\mathbf{Q})|) \right) \end{aligned}$$

where $\tilde{b}(\ell) \triangleq b(\ell)/D$ and we used that $\log(1+x) \leq x$. Thus, we obtain

$$\begin{aligned} & \liminf_{L \rightarrow \infty} -\frac{1}{L} \log (C(\mathbf{Q}^L) - \log D) \\ & \geq \liminf_{L \rightarrow \infty} -\frac{1}{L} \log \left(C(\tilde{\mathbf{Q}}^{\lfloor L/d \rfloor}) - \log D \right) \\ & \geq \liminf_{L \rightarrow \infty} -\frac{1}{L} \log \tilde{b}(\lfloor L/d \rfloor) - \frac{d \lfloor L/d \rfloor}{L} \log |\lambda_2(\mathbf{Q})| \\ & = -\log |\lambda_2(\mathbf{Q})|. \quad \square \end{aligned}$$

As a next step, we restrict attention to stochastic matrices \mathbf{Q} without inessential indices, for which we can find the exponent $\lim_{L \rightarrow \infty} E_L(\mathbf{Q})$ exactly.

Lemma II.2: If \mathbf{Q} is a stochastic matrix without inessential indices then

$$\lim_{L \rightarrow \infty} E_L(\mathbf{Q}) = -2 \log |\lambda_2(\mathbf{Q})|$$

where $E_L(\mathbf{Q}) \triangleq -\frac{1}{L} \log (C(\mathbf{Q}^L) - \log D)$.

Proof: Again start from the initial steps (40)–(44), and define the probability distribution

$$\tilde{\pi} \triangleq (\tilde{\pi}_1, \dots, \tilde{\pi}_D)/D. \tag{45}$$

We first show that

$$\liminf_{L \rightarrow \infty} E_L(\mathbf{Q}) \geq -2 \log |\lambda_2(\mathbf{Q})|$$

by upper-bounding the capacity $C(\tilde{\mathbf{Q}}^\ell)$. Define the set

$$\mathcal{A}(\beta) \triangleq \left\{ \hat{\mathbf{Q}} \in \mathcal{S}_{m,m} : \|\hat{\mathbf{Q}} - \tilde{\mathbf{Q}}^\infty\|_\infty \leq \beta, \tilde{\mathbf{Q}}^\infty \gg \hat{\mathbf{Q}} \right\},$$

where m is the dimension of $\tilde{\mathbf{Q}}$, and where $\mathbf{A} \gg \mathbf{B}$ denotes that $[\mathbf{B}]_{i,j} = 0$ whenever $[\mathbf{A}]_{i,j} = 0$. From (41), $\tilde{\mathbf{Q}}^\ell \in \mathcal{A}(b(\ell)|\lambda_2(\mathbf{Q})|^{d\ell})$ for some subexponential $b(\ell)$. In particular, for

$$\beta \triangleq b(\ell)|\lambda_2(\mathbf{Q})|^{d\ell},$$

$C(\tilde{\mathbf{Q}}^\ell) \leq \sup_{\hat{\mathbf{Q}} \in \mathcal{A}(\beta)} C(\hat{\mathbf{Q}})$. We can now use the minimax formula for capacity in [32, Problem 2.3.1] and substitute $\tilde{\pi}$ for \mathbf{p}_Y to obtain

$$\begin{aligned} C(\tilde{\mathbf{Q}}^\ell) & \leq \sup_{\hat{\mathbf{Q}} \in \mathcal{A}(\beta)} C(\hat{\mathbf{Q}}) \\ & = \sup_{\hat{\mathbf{Q}} \in \mathcal{A}(\beta)} \min_{\mathbf{p}_Y} \max_{x \in \mathcal{X}} D(\hat{\mathbf{q}}(\cdot|x) \| \mathbf{p}_Y) \\ & \leq \max_{x \in \mathcal{X}} \sup_{\hat{\mathbf{Q}} \in \mathcal{A}(\beta)} D(\hat{\mathbf{q}}(\cdot|x) \| \tilde{\pi}) \end{aligned}$$

where $\hat{\mathbf{q}}(\cdot|x)$ is the row of $\hat{\mathbf{Q}}$ corresponding to x . Let \mathcal{E}_i denote the essential indices of $\hat{\mathbf{Q}}$ corresponding to $\tilde{\mathbf{P}}_i$. Call $\hat{\mathbf{q}}_i \triangleq \hat{\mathbf{q}}(\cdot|x)$, for some $x \in \mathcal{E}_i$. By symmetry of the set $\mathcal{A}(\beta)$, the last maximization over all $\hat{\mathbf{Q}} \in \mathcal{A}(\beta)$ can equivalently be expressed as maximization over all $\hat{\mathbf{q}}_i = \tilde{\pi}_i + \mathbf{a}$, under the following constraints, for all i, y : i) $|a_y| \leq \beta$, ii) $a_y = 0$ if $y \notin \mathcal{E}_i$, iii) $\sum_y a_y = 0$, and iv) $\beta < \tilde{\pi}_i(y)$. The last constraint is automatically satisfied assuming ℓ large enough, because $\tilde{\mathbf{P}}_i$ is primitive and hence $\tilde{\pi}_i(y) > 0$. With this, we can continue the above chain of inequalities

$$\begin{aligned} C(\tilde{\mathbf{Q}}^\ell) &\leq \max_{i \in \{1, \dots, D\}} \sup_{\mathbf{a}} \sum_{y \in \mathcal{E}_i} (\tilde{\pi}_i(y) + a_y) \log \left(D + \frac{D a_y}{\tilde{\pi}_i(y)} \right) \\ &\leq \log(D) + \max_{i \in \{1, \dots, D\}} \sup_{\mathbf{a}} \sum_{y \in \mathcal{E}_i} (\tilde{\pi}_i(y) + a_y) \frac{a_y}{\tilde{\pi}_i(y)} \\ &\leq \log(D) + \beta^2 \max_{i \in \{1, \dots, D\}} \sum_{y \in \mathcal{E}_i} \frac{1}{\tilde{\pi}_i(y)} \end{aligned}$$

where we used that $\log(1+x) \leq x$. Call

$$A \triangleq \max_{i \in \{1, \dots, D\}} \sum_{y \in \mathcal{E}_i} \frac{1}{\tilde{\pi}_i(y)}.$$

Note that $A < \infty$ since all components of $\tilde{\pi}_i$ are strictly positive for all i (all $\{\tilde{\mathbf{P}}_i\}$ are primitive). We conclude that

$$\begin{aligned} \liminf_{L \rightarrow \infty} -\frac{1}{L} \log (C(\mathbf{Q}^L) - \log D) \\ &\geq \liminf_{L \rightarrow \infty} -\frac{1}{L} \log \left(C(\tilde{\mathbf{Q}}^{\lfloor L/d \rfloor}) - \log D \right) \\ &\geq \liminf_{L \rightarrow \infty} -\frac{1}{L} \log \left(A b (\lfloor L/d \rfloor)^2 |\lambda_2(\mathbf{Q})|^{2d \lfloor L/d \rfloor} \right) \\ &= -2 \log |\lambda_2(\mathbf{Q})|. \end{aligned}$$

We now show that

$$\limsup_{L \rightarrow \infty} E_L(\mathbf{Q}) \leq -2 \log |\lambda_2(\mathbf{Q})|.$$

To do that, we use the lower bound

$$C(\tilde{\mathbf{Q}}^\ell) \geq I(\tilde{\pi}, \tilde{\mathbf{Q}}^\ell)$$

where $\tilde{\pi}$ is defined as in (45).

Let $\tilde{q}^\ell(\cdot|x)$ be a row of $\tilde{\mathbf{Q}}^\ell$, where the row index x is an essential index for the primitive matrix $\tilde{\mathbf{P}}_{i(x)}$ with stationary distribution $\tilde{\pi}_{i(x)}$, and $\mathcal{E}_{i(x)}$ is the set of all essential indices corresponding to $\tilde{\mathbf{P}}_{i(x)}$. Using (41) we have

$$\tilde{q}^\ell(y|x) = \begin{cases} \tilde{\pi}_{i(x)}(y) + \beta_{x,y}, & \text{if } y \in \mathcal{E}_{i(x)} \\ 0, & \text{else} \end{cases}$$

with

$$\beta_{x,y} \triangleq b_{x,y}(\ell) |\lambda_2(\tilde{\mathbf{P}}_{i(x)})|^\ell \quad (46)$$

for some subexponential $b_{x,y}(\ell)$. Note that since both $\tilde{\mathbf{Q}}^\ell$ and $\tilde{\mathbf{Q}}^\infty$ in (41) are stochastic

$$\sum_{y \in \mathcal{E}_{i(x)}} \beta_{x,y} = 0. \quad (47)$$

Using the fact that $\tilde{\pi}$ in (45) satisfies $\tilde{\pi} \tilde{\mathbf{Q}}^\ell = \tilde{\pi}$, the inequality $\log(1+x) \geq x - \frac{3}{4}x^2$ which is valid for all $x \geq -\varepsilon$ for some $\varepsilon > 0$, and (47), we get that

$$\begin{aligned} C(\tilde{\mathbf{Q}}^\ell) &\geq I(\tilde{\pi}, \tilde{\mathbf{Q}}^\ell) \\ &= \sum_x \tilde{\pi}(x) \sum_y \tilde{q}^\ell(y|x) \log \frac{\tilde{q}^\ell(y|x)}{\tilde{\pi}(y)} \\ &= \sum_x \sum_{y \in \mathcal{E}_{i(x)}} \tilde{\pi}(x) (\tilde{\pi}_{i(x)}(y) + \beta_{x,y}) \log \left(D + \frac{D \beta_{x,y}}{\tilde{\pi}_{i(x)}(y)} \right) \\ &\geq \log(D) + \sum_x \sum_{y \in \mathcal{E}_{i(x)}} \tilde{\pi}(x) (\tilde{\pi}_{i(x)}(y) + \beta_{x,y}) \\ &\quad \times \left(\frac{\beta_{x,y}}{\tilde{\pi}_{i(x)}(y)} - \frac{3\beta_{x,y}^2}{4\tilde{\pi}_{i(x)}(y)^2} \right) \\ &= \log(D) + \sum_x \sum_{y \in \mathcal{E}_{i(x)}} \frac{\tilde{\pi}(x)}{\tilde{\pi}_{i(x)}(y)} \left(\frac{1}{4} \beta_{x,y}^2 - \frac{3}{4\tilde{\pi}_{i(x)}(y)} \beta_{x,y}^3 \right). \end{aligned}$$

Since all $\{\tilde{\mathbf{P}}_i\}$ are primitive, all $\tilde{\pi}_i(y)$ are strictly positive and hence, for ℓ large enough, $|\beta_{x,y}/\tilde{\pi}_{i(x)}(y)| \leq \varepsilon$. Moreover, the $\beta_{x,y}^2$ term dominates the $\beta_{x,y}^3$ term for large ℓ , and thus we conclude using (46) that

$$\begin{aligned} \limsup_{L \rightarrow \infty} -\frac{1}{L} \log (C(\mathbf{Q}^L) - \log D) \\ &\leq \limsup_{L \rightarrow \infty} -\frac{1}{L} \log \left(C(\tilde{\mathbf{Q}}^{\lceil L/d \rceil}) - \log D \right) \\ &\leq \limsup_{L \rightarrow \infty} -\frac{2 \lceil L/d \rceil}{L} \log (\max_i |\lambda_2(\tilde{\mathbf{P}}_i)|) \\ &= -\frac{2}{d} \log |\lambda_2(\tilde{\mathbf{Q}})| \\ &= -2 \log |\lambda_2(\mathbf{Q})|. \quad \square \end{aligned}$$

Theorem III.5 follows from the two lemmas by noting that we can lower-bound $C(\mathbf{Q}^L)$ by restricting the support of the input distribution to contain only essential indices of \mathbf{Q} . Or, equivalently, by reducing \mathbf{Q} to the stochastic matrix $\hat{\mathbf{Q}}$, containing only the essential indices of \mathbf{Q} .

APPENDIX III

WORST CHANNEL UNDER A TRACE CONSTRAINT

The next lemma provides conditions for a channel \mathbf{W} to be the worst under certain constraints. Consider a set Ω of (x, y) pairs and define

$$\mathcal{W} \triangleq \left\{ \mathbf{W} \in \mathcal{S}_{n,n} : \sum_{(x,y) \in \Omega} w(y|x) \geq b \right\}$$

for some constant b .

Lemma III.1: \mathbf{W} is the solution to the minimization problem

$$\min_{\mathbf{W} \in \mathcal{W}} I(\mathbf{p}, \mathbf{W})$$

if and only if \mathbf{W} is feasible (i.e., $\mathbf{W} \in \mathcal{W}$) and satisfies the following conditions:

$$p(\alpha) \log \frac{w(\gamma|\alpha)}{\sum_{\tilde{x}} p(\tilde{x})w(\gamma|\tilde{x})} \begin{cases} = \mu(\alpha) + \mathbf{1}_\Omega(\alpha, \gamma)\lambda_b, & \text{if } w(\gamma|\alpha) > 0 \\ \geq \mu(\alpha) + \mathbf{1}_\Omega(\alpha, \gamma)\lambda_b, & \text{if } w(\gamma|\alpha) = 0 \end{cases} \quad (48)$$

for all $\alpha \in \mathcal{X}$, $\gamma \in \mathcal{Y}$, for some vector μ , and for some $\lambda_b \geq 0$ such that

$$\left(\sum_{(x,y) \in \Omega} w(y|x) - b \right) \lambda_b = 0. \quad (49)$$

Proof: The set of matrices \mathcal{W} is defined by linear inequalities and equalities and is hence a convex set. This implies, together with the fact that mutual information is convex in the channel transition probability matrix, that

$$\min_{\mathbf{W} \in \mathcal{W}} I(\mathbf{p}, \mathbf{W})$$

is a convex minimization problem in \mathbf{W} , and the Kuhn–Tucker conditions are necessary and sufficient for optimality [33, Ch. 5.5.3]. The Lagrangian of the minimization problem is given by

$$\sum_{x,y} p(x)w(y|x) \log \frac{w(y|x)}{\sum_{\tilde{x}} p(\tilde{x})w(y|\tilde{x})} - \sum_{x,y} w(y|x)\lambda_a(x,y) - \left(\sum_{(x,y) \in \Omega} w(y|x) - b \right) \lambda_b - \sum_x \left(\sum_y w(y|x) - 1 \right) \mu(x). \quad (50)$$

The derivative of (50) with respect to $w(\gamma|\alpha)$ is

$$p(\alpha) \log \frac{w(\gamma|\alpha)}{\sum_{\tilde{x}} p(\tilde{x})w(\gamma|\tilde{x})} - \lambda_a(\alpha, \gamma) - \mathbf{1}_{\{(\alpha,\gamma) \in \Omega\}} \lambda_b - \mu(\alpha). \quad (51)$$

The Kuhn–Tucker conditions for optimality require then that (51) is equal to zero for all α, γ , and for $\lambda_a(\alpha, \gamma), \lambda_b \geq 0$ together with the constraint that $\mathbf{W} \in \mathcal{W}$. Moreover, we have complementary slackness, that is, (49) and

$$w(\gamma|\alpha)\lambda_a(\alpha, \gamma) = 0 \quad (52)$$

for all (α, γ) . As all the $\lambda_a(\alpha, \gamma)$ are nonnegative, (51) and (52) can be combined to yield (48). \square

Example III.1: Consider the set $\mathcal{W} \subset \mathcal{S}_{n,n}$ of stochastic matrices \mathbf{W} with $\text{tr}(\mathbf{W}) \geq nb$, and $\mathbf{p} = \frac{1}{n}\mathbf{1}^T$.

If $1 \geq b > 1/n$, then

$$\mathbf{W}^* = \frac{1-b}{n-1}\mathbf{1}\mathbf{1}^T + \frac{nb-1}{n-1}\mathbf{I}$$

is an element of \mathcal{W} and satisfies the Kuhn–Tucker conditions given in Appendix III Lemma III.1. Hence, \mathbf{W}^* is the worst channel in the set \mathcal{W} .

If $b \leq 1/n$, then the worst channel is (trivially)

$$\mathbf{W}^* = \frac{1}{n}\mathbf{1}\mathbf{1}^T. \quad \diamond$$

ACKNOWLEDGMENT

The authors would like to thank J. Körner for pointing out the connection with common information; C. Chekuri, O. Lévêque, B. Rimoldi, E. Telatar, and D. Shah for many helpful discussions; and the anonymous reviewers whose constructive comments helped improving the final version of this manuscript.

REFERENCES

- [1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
- [2] R. Ahlswede, N. Cai, S.-Y. R. Li, and R. W. Yeung, “Network information flow,” *IEEE Trans. Inf. Theory*, vol. 46, no. 4, pp. 1204–1216, Jul. 2000.
- [3] S.-Y. R. Li, R. W. Yeung, and N. Cai, “Linear network coding,” *IEEE Trans. Inf. Theory*, vol. 49, no. 2, pp. 371–381, Feb. 2003.
- [4] P. Sanders, S. Egner, and L. Tolhuizen, “Polynomial time algorithms for network information flow,” in *Proc. 15th ACM Symp. Parallel Algorithms and Architectures*, San Diego, CA, Jun. 2003, pp. 286–294.
- [5] R. Koetter and M. Médard, “Beyond routing: An algebraic approach to network coding,” in *Proc. IEEE INFOCOM*, New York, Jun. 2002, vol. 1, pp. 122–130.
- [6] C. Fragouli and E. Soljanin, “Information flow decomposition for network coding,” *IEEE Trans. Inf. Theory*, vol. 52, no. 3, pp. 829–848, Mar. 2006.
- [7] C. E. Shannon, “The zero error capacity of a noisy channel,” *IEEE Trans. Inf. Theory*, vol. IT-2, no. 3, pp. 8–19, Sep. 1956.
- [8] P. Gupta and P. Kumar, “The capacity of wireless networks,” *IEEE Trans. Inf. Theory*, vol. 46, no. 2, pp. 388–404, Mar. 2000.
- [9] M. K. Simon, “On the capacity of a cascade of identical discrete memoryless nonsingular channels,” *IEEE Trans. Inf. Theory*, vol. IT-16, no. 1, pp. 100–102, Jan. 1970.
- [10] R. A. Silverman, “On binary channels and their cascade,” *IEEE Trans. Inf. Theory*, vol. IT-1, no. 3, pp. 19–27, Dec. 1955.
- [11] A. B. Kieley and J. T. Coffey, “On the capacity of a cascade of channels,” *IEEE Trans. Inf. Theory*, vol. 39, no. 4, pp. 1310–1321, Jul. 1993.
- [12] D. Lun, M. Médard, and M. Effros, “On coding for reliable communication over packet networks,” in *Proc. 42nd Annu. Allerton Conf. Communications, Control, and Computing*, Monticello, IL, Oct. 2004.
- [13] P. Pakzad, C. Fragouli, and A. Shokrollahi, “Coding schemes for line networks,” in *Proc. IEEE Int. Symp. Information Theory*, Adelaide, Australia, Sep. 2005, pp. 1853–1857.
- [14] U. Niesen, C. Fragouli, and D. Tuninetti, “On the capacity of an infinite cascade of channels,” in *Proc. Allerton Conf. Communications, Control, and Computing*, Monticello, IL, Sep. 2005.
- [15] J. Körner and A. Orlitsky, “Zero-error information theory,” *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2207–2229, Oct. 1998.
- [16] E. Seneta, *Non-Negative Matrices and Markov Chains*. New York: Springer-Verlag, 1981.
- [17] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1985.
- [18] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1991.
- [19] C. E. Shannon, “Some geometrical results in channel capacity,” *Nachrichtentechnische Z.*, vol. 10, pp. 1–4, Jan. 1957.
- [20] R. T. Rockafellar, *Convex Analysis*. Princeton, NJ: Princeton Univ. Press, 1996.
- [21] M. Saks and A. Condon, “A limit theorem for sets of stochastic matrices,” *Linear Algebra and Applications*, vol. 381, pp. 61–76, 2004.
- [22] P. Gács and J. Körner, “Common information is far less than mutual information,” *Probl. Contr. Inf. Theory*, vol. 2, no. 2, pp. 149–162, 1973.
- [23] S. Wolf and J. Wullschlegler, “Zero-error information and applications in cryptography,” in *Proc. IEEE Inf. Theory Workshop*, San Antonio, TX, Oct. 2004, pp. 1–6.

- [24] R. Renner, S. Wolf, and J. Wullschleger, "The single-serving channel capacity," in *Proc. IEEE Int. Symp. Information Theory*, Seattle, WA, Jul. 2006, pp. 1424–1427.
- [25] L. Lovász, "On the Shannon capacity of a graph," *IEEE Trans. Inf. Theory*, vol. IT-25, no. 1, pp. 1–7, Jan. 1979.
- [26] V. Vazirani, *Approximation Algorithms*. New York: Springer-Verlag, 2001.
- [27] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.
- [28] D. Tuninetti and C. Fragouli, "On the throughput improvement due to limited complexity processing at relay nodes," in *Proc. IEEE Int. Symp. Information Theory*, Adelaide, Australia, Sep. 2005, pp. 1081–1085.
- [29] D. Tuninetti, U. Niesen, and C. Fragouli, "On capacity of line networks," in *Proc. ITA 2006*, San Diego, CA, Jan. 2006 [Online]. Available: <http://ita.ucsd.edu/workshop/06/talks/papers/157.pdf>
- [30] W. Rudin, *Principles of Mathematical Analysis*. New York: McGraw-Hill, 1976.
- [31] M. Chiang and S. Boyd, "Geometric programming duals of channel capacity and rate distortion," *IEEE Trans. Inf. Theory*, vol. 50, no. 2, pp. 245–258, Feb. 2004.
- [32] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1981.
- [33] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.