ON A THEORY OF NETWORK EQUIVALENCE

Ralf Koetter
ralf.koetter@tum.de

Michelle Effros
effros@caltech.edu

Muriel Médard
medard@mit.edu

Abstract—We describe an equivalence result for network capacity. Roughly, our main result is as follows. Given a network of noisy, independent, memoryless links, a collection of demands can be met on the given network if and only if it can be met on another network where each noisy link is replaced by a noiseless bit pipe with throughput equal to the noisy link capacity. This result had heretofore only been known for the case of multicast connections.

I. INTRODUCTION

Network information theory has two natural facets reflecting different approaches to networking. On the one hand, networks are considered in the graph theoretic setup consisting of nodes and links connecting them. Typical concepts would, for example, be information flows and routing issues. A typical network model would not include noisy links but the links connecting nodes in a network are, by and large, noise-free bit pipes which can be used error free up to a certain capacity. On the other hand, multiterminal information theory deals with noisy channels, or rather the stochastic nature of input and output variables at nodes in a network. Here the interplay of transmissions in a network leads to a quite different set of questions dealing with fundamental limits of communication. The capacity regions of broadcast channels, multiple access channels, or the interference channel are all examples of questions that are addressed in the context of multiterminal information theory and that appear to have no obvious equivalent in networks consisting of error free bit pipes. Nevertheless, the two views of networking are two natural facets of the same problem, namely communication through networks. The objective of this paper is to explore the relationship between these two worlds.

Establishing viable bridges between these two areas shows to be surprisingly fertile. For example questions about feedback in multiterminal systems are quite nicely expressed in networks of error free links. Separation issues — in particular separation between network coding and channel coding — have natural answers, revealing many network problems as combinatorial rather than statistical, even in noisy networks. In fact, many problems in network information theory appear to be reducible to solving a central network coding problem described as follows: Given a network of error free rate-constrained bit pipes, is a given set of demands satisfieable or not. In certain situations, most notably a multicast demand, this question has nice and simple answers, however, the general case is wide open. In fact, it is suspected that the central combinatorial network coding problem is hard, however NP hardness is only established for the class of linear network coding [1]. Nevertheless, there are algorithms available that, with running time that is exponential in the number of nodes, solve just this problem [2]. This possibility to, in principle, characterize the rate region of a combinatorial network coding problem will be a cornerstone for our investigations.

It thus appears that any research into network information theory must acknowledge the fact that fully characterizing the combinatorial network coding problem is out of reach [3]. Nevertheless, moderate size networks can be solved quite efficiently. The situation is not unlike issues in complexity theory, where a lot of research is being devoted to show that one problem is essentially as difficult as another one without being able to give precise expressions as to how difficult a problem is in absolute terms. Inspired by this analogy we thus resort in this paper to characterizing the relationship of a number of network coding problems to the central combinatorial network coding problem. This characterization is in fact all we need if we want to address separation issues in networks but also other questions, such as a degree-of-freedom or high signal to noise ratio analysis, reveal interesting insights.

It is interesting to note that the reduction of a network information theoretic question to its combinatorial essence is also at the heart of some related recent publications, see e.g. [4]. While our approach is quite different in terms of technique and also results, we believe it to be no coincidence that in both cases the reduction of a problem to its combinatorial essence is a central step.

II. INTUITION AND SUMMARY OF RESULTS

To clarify the benefits of the proposed approach, consider the problem of finding the capacity region of a network of independent noisy channels, for example a network of Gaussian channels. When there are general demands on such a network, i.e. any two nodes in the network may want to exchange information, this problem seems
to be completely out of reach. Yet, for special demands on the network, the situation is different. For example, the case of a single unicast demand in the noise free network was solved by Ford and Fulkerson [5]. The case of a single multicast demand was solved for noise-free networks by Ahlswede et al. [6] and for noisy links by Cai et al. [7]. Cai’s paper shows a separation between the combinatorial and statistical problems. This result is proven by first finding an outer bound for the desired rate region and then showing that that outer bound is reachable using channel coding to replace noisy links by lossless bit pipes and then network coding across those bit pipes. That approach cannot be used to extend this result to arbitrary demands since the cut-set outer bound is no longer tight and finding the true outer bound for each new demand structure is an NP hard problem.

While the separated strategy is always achievable, it is clearly not optimal in many networks. For example, consider a pair of identical, independent channels each of capacity $C$ running in parallel from a shared source node to a shared sink node. We consider two strategies for reliable communication across this network. In the first, we communicate across the two channels using identical channel codes with $2^{nR}$ codewords of blocklength $n$ for some $R < C$; here each channel may see as many as $2^{nR}$ distinct inputs. In the second, we design a single codebook with $2^{2nR}$ codewords of blocklength $2n$; we send the first $n$ symbols of each codeword across the first channel and the remaining symbols across the second channel. With this approach, each channel may see as many as $2^{2nR}$ distinct inputs. The first strategy is the separation approach, which operates each channel below its capacity, allowing asymptotically lossless transmission across each link. The second strategy operates each channel at twice its capacity, failing to achieve reliable communication across either link but achieving reliable communication across the network when the channel outputs are decoded together. Since the second strategy doubles the code’s blocklength, it improves the code’s error exponent. Thus joint channel and network coding can improve the error exponent. It remains to be shown whether joint channel and network coding is ever required to achieve network capacity.

We can express our main result in the following statement, which we prove in Section IV in a more formal setting. Given a network of noisy, independent, memoryless links, a collection of demands can be met on the given network if and only if it can be met on another network where each noisy link is replaced by a noiseless bit pipe with throughput equal to the noisy link capacity. This result subsumes the separation result of [7] in the special case of multicase demands but does so without resorting to calculating capacity regions since that approach is infeasible for our more general problem.

This claim has a number of surprisingly powerful consequences. Not only would it show separation between channel and network coding for arbitrary networks and arbitrary demands, also, many network information theoretic questions are naturally asked in the light of this combinatorial perspective. For example, the classical result that feedback does not increase the capacity of a point-to-point link now can be proven in two ways. The first is the information theoretic argument that shows that the channel has no information that is useful to the transmitter that the transmitter does not already know. The second simply observes that the min-cut between transmitter and receiver is the same with or without feedback, which is obvious from the given equivalence. Most importantly, this statement reveals that at the heart of information theory lie combinatorial problems involving finding the rate region for error-free networks.

Since the prior approach of finding the outer bound and proving it achievable is out of the question, we build an equivalence theorem instead. Intuitively, we prove equivalence between networks A and B by showing that if anyone shows us a way to operate network A at one rate point, then we can find a way to operate network B at the same rate point and vice versa. Note that this never answers the question of whether a particular rate point is in the rate region or not. Operating codes designed for bit pipes in the noisy channel network is straightforward using a separated strategy. The other direction is harder since a noisy network allows a far richer algorithmic behavior. It is known that a noiseless bit-pipe of a given throughput can emulate any discrete memoryless channel of lesser capacity [8], so a network of bit pipes may be operated as if it were a network of noisy links. Yet applying this result seems to be difficult. Difficulties arise with continuous random variables, timing questions, and proving continuity of rate regions. Worst of all, since we do not know which strategy achieves the network capacity, we must be able to emulate all of them. We prove our main claim directly, without exploiting [8].

III. THE SETUP

Our notation is similar to that of Cover and Thomas [9]. A multiterminal network comprises $m$ nodes with associated random variables $X^{(i)} \in \mathcal{X}^{(i)}$ which are transmitted from node $i$ and $Y^{(i)} \in \mathcal{Y}^{(i)}$ which are received at node $i$. The network is assumed to be memoryless, so it is characterized by a conditional probability distribution $p(y|x) = p(y^{(1)}, \ldots, y^{(m)}|x^{(1)}, \ldots, x^{(m)})$. We operate
the network over $n$ time steps with the goal of communicating message $W^{(i-j)} \in W^{(i-j)} = \{1, \ldots, 2^n R^{(i-j)}\}$ from source node $i$ to sink node $j$, where $R^{(i-j)}$ is called the rate of the transmission. The vector of rates $R^{(i-j)}$ is denoted by $R$. We denote the random variables transmitted and received at node $i$ at time-step $t$ by $X^{(i)}_t$ and $Y^{(i)}_t$. A network is thus written as a triple $(\prod_{i=1}^m X^{(i)}_t, \eta(y|x), \prod_{i=1}^m Y^{(i)}_t)$ with the additional constraint that random variable $X^{(i)}_1$ is a function of random variables $\{Y^{(1)}_1, \ldots, Y^{(i-1)}_1, W^{(i-1)}, \ldots, W^{(i-m)}\}$ alone.

While this characterization is very general, it does not exploit any information about the network’s structure.

The structure is given as a hypergraph $G$ with node set $V = \{1, \ldots, m\}$ and hyperedge set $E$. Let $\mathcal{P}(V)$ denote the power set of $V$. Each directed hyperedge $e \in E$ takes the form $e = [V_1, V_2], V_1, V_2 \in \mathcal{P}(V)$. Thus $E \subseteq \mathcal{P}(V) \times \mathcal{P}(V)$. For example, a point to point channel is described by a single transmitter $V_1 = \{i\}$ and a single receiver $V_2 = \{j\}$, a broadcast channel is described by a single transmitter $V_1 = \{i\}$ and multiple receivers $V_2 = \{j_1, \ldots, j_k\}$, a multiple access channel is described by multiple transmitters $V_1 = \{i_1, \ldots, i_k\}$ and a single receiver $V_2 = \{j\}$, and so on. This paper focuses on point-to-point transmissions, where $|V_1| = |V_2| = 1$; we treat broadcast and multiple access transmission in future work. We keep the hyperedge notation to highlight the observation that the strategy employed here generalizes beyond the case of point-to-point links. Though separation does not hold in general for hyperlinks, the same approach applies to prove bounds on more general network components.

The indegree $d_{\text{in}}(i)$ and outdegree $d_{\text{out}}(i)$ of node $i$ in hypergraph $G$ are defined as $d_{\text{in}}(i) = |\{V_1, V_2 \in E, i \in V_2\}|$ and $d_{\text{out}}(i) = |\{V_2, V_1, V_2 \in E, i \in V_1\}|$. If a node has outdegree or indegree larger than one then $X^{(i)}_t = \prod_{i=d_{\text{out}}(i)} X^{(i)}_t$ and $Y^{(i)}_t = \prod_{d_{\text{in}}(i)} Y^{(i)}_t$ and the inputs and outputs of nodes at time $t$ are given by $X^{(i)}_t = \{X^{(i)}_1, \ldots, X^{(i,d_{\text{out}}(i))}_t\}$ and $Y^{(i)}_t = \{Y^{(i)}_1, \ldots, Y^{(i,d_{\text{in}}(i))}_t\}$. Given a hyperedge $e = [V_1, V_2]$, we use $V_1(e)$ and $V_2(e)$ to denote the input and output ports of hyperedge $e$ and $X(V_1(e))$ and $Y(V_2(e))$ to denote the sets of input and output random variables of hyperedge $e$. For example, consider a link $e$ from node $i$ to node $j$. Then the input to channel $e = \{i, j\}$ at node $i$ is $X^{(i,s)}$ for some index $s \in \{1, \ldots, d_{\text{out}}(i)\}$ while the output from channel $\{i, j\}$ at node $j$ is $Y^{(j,r)}$ for some index $r \in \{1, \ldots, d_{\text{in}}(j)\}$. In this case, $V_1(e) = (i, s)$ and $V_2(e) = (j, r)$, giving $X(V_1(e)) = X^{(i,s)}$ and $Y(V_2(e)) = Y^{(j,r)}$ as desired. The collection of sets $\{V_1(e) : e \in E\}$ and $\{V_2(e) : e \in E\}$ forms a partition of the set of random variables $X^{(1,1)}, \ldots, X^{(m,d_{\text{out}}(m))}$.

When the characterization corresponds to network $G = (V, E)$, we factor $p(y|x)$ to give the characterization:

$$
\left(\prod_{i=1}^m X^{(i)}_t, \prod_{e \in E} p(Y(V_2(e))|X(V_1(e))), \prod_{i=1}^m Y^{(i)}_t\right)
$$

again with the additional property that random variable $X^{(i)}_t$ is a function of random variables $\{Y^{(1)}_1, \ldots, Y^{(i-1)}_1, W^{(i-1)}, \ldots, W^{(i-m)}\}$ alone. In this paper we want to investigate some information theoretic aspects of replacing factors in the factorization of $p(y|x)$.

**Definition 1** Let a network $\mathcal{N} \equiv \left(\prod_{i=1}^m X^{(i)}_t, \prod_{e \in E} p(y(V_2(e))|x(V_1(e))), \prod_{i=1}^m Y^{(i)}_t\right)$ be given corresponding to a hypergraph $G = (V, E)$. A blocklength-$n$ solution $S(\mathcal{N})$ to this network is defined as a set of encoding and decoding functions:

$$X^{(i)}_t : (\mathcal{Y}(i))^{t-1} \times \prod_{j=1}^m W^{(i-j)} \rightarrow X^{(i)}_t$$

$$W^{(j-i)} : (\mathcal{Y}(j))^{n} \times \prod_{i=1}^m W^{(i-j)} \rightarrow W^{(j-i)}$$

mapping $(Y^{(1)}_1, \ldots, Y^{(i)}_t, W^{(i-1)}, \ldots, W^{(i-m)})$ to $X^{(i)}_t$ for each $i \in V$ and $t \in \{1, \ldots, n\}$ and mapping $(Y^{(1)}_1, \ldots, Y^{(i)}_t, W^{(i-1)}, \ldots, W^{(i-m)})$ to $W^{(j-1)}$ for each $i, j \in V$. The solution $S(\mathcal{N})$ is called a $(\lambda, R)$-solution, denoted $(\lambda, R) - S(\mathcal{N})$, if the specified encoding and decoding functions imply $Pr(W^{(i-j)} \neq W^{(i-j)}) < \lambda$ for all source and sink pairs $i, j$.

**Definition 2** The rate region $\delta(\mathcal{N}) \subseteq \mathbb{R}_+^{m(m-1)}$ of a network $\mathcal{N}$ is the closure of all rate vectors $R$ such that for any $\lambda > 0$, a $(\lambda, R) - S(\mathcal{N})$ solution exists.\(^1\)

**IV. Results**

The goal of this paper is not to give the capacity regions of networks with respect to various demands, which is

---

\(^1\)It here suffices to find a solution for any blocklength $n$. The blocklength required may vary with $\lambda$, $R$, and $\mathcal{N}$. 

An intractable problem owing to its combinatorial nature. Rather we wish to develop equivalence relationships between capacity regions of networks. Given the existence of a solution \((\lambda, R) - S(N)\) of some blocklength \(n\) for a network \(N\) we will try to imply statements for the existence of a solution \((N', R') - S(N')\) of some blocklength \(n'\) for a network \(N'\).

Assume, for example, a network contains a hyperedge \(\bar{e} = \{i, \{j\}\}\) which is a link between nodes \(i\) and \(j\). The input and output random variables are \(X(V_1(e)) = X^{(i,s)}\) and \(Y(V_2(e)) = Y^{(j,r)}\), where \(s \in \{1, \ldots, d_{out}(i)\}\) and \(r \in \{1, \ldots, d_{in}(j)\}\) are the indices of \(X^{\bar{e}}(i)\) and \(Y^{\bar{e}}(j)\) corresponding to edge \(e\). The transition probability for the network thus factors as:

\[
p(y^{(j,r)}|x^{(i,s)}; \bar{e}) = \prod_{e \in E} p(y^{V_2(e)}|x^{V_1(e)})\]

Let another network \(N'\) be given with random variables \((X^{(i,s)}, Y^{(j,r)})\) replacing \((X^{(i,s)}, Y^{(j,r)})\) in \(N\). We have replaced the link characterized by \(p(y^{(j,r)}|x^{(i,s)})\) with another link characterized by \(\hat{p}(y^{(j,r)}|x^{(i,s)})\). When \(I(X^{\bar{e}}; Y^{\bar{e}}) \leq I(X^{\bar{e}}; Y^{\bar{e}})\), we want to prove that the existence of a \((\lambda, R) - S(N)\) solution implies the existence of a \((N', R') - S(N')\) solution, where \(\lambda'\) can be made arbitrarily small if \(\lambda\) can. Since node \(j\) need not decode \(Y^{\bar{e}}(j)\), channel capacity is not necessarily a relevant characterization of the link’s behavior. For example a Gaussian channel from \(i\) to \(j\) might contribute a real-valued estimation of the input random variable; a binary erasure channel that replaces it cannot immediately deliver the same functionality.

Our proof does not invent a coding scheme. Instead, we demonstrate a technique for operating any coding scheme for \(N'\) on the network \(N\). Since there exists a coding scheme for \(N\) that achieves any point in the interior of \(B(N)\), this proves that \(B(N) \subseteq B(N')\). Since we don’t know the form of an optimal code for \(N\), our method must work for all possible codes on \(N\). For example, it must succeed even when the code for \(N\) is time-varying. As a result, we cannot apply typicaly arguments across time. We introduce instead a notion of stacking in order to exploit averaging arguments across multiple uses of the same network rather than trying to exploit such arguments across time.

The \(N\)-fold stacked network \(N^N\) is the network \(N\) repeated \(N\) times. That is, \(N^N\) has \(N\) copies of each vertex \(v \in V\) and \(N\) copies of each hyperedge \(e \in E\). All copies of a node are allowed to operate together to form the channel inputs and message reconstructions at each node. We show, however, in Theorem 1 that it is always possible to achieve the network capacity with a code that operates the same solution independently in each layer, and that is the approach that we eventually employ.

Since the vertex and edge sets of \(N^N\) are multisets and not sets, the stacked network is not a network and new definitions are required. We carry over notation and variable definitions from network \(N\) to the stacked network \(N^N\) by underlining the variable names. So for any \(i, j \in V\), \(W^{(i-j)} \in W^{(i-j)} - S(N)\) is the \(N\)-dimensional vector of messages that originate at source \(i\) and are to be transmitted to sink \(j\), and \(X^{(i)} \in X^{(i)} - S(N^N)\) and \(Y^{(i)} \in Y^{(i)} - S(N^N)\) are the \(N\)-dimensional vectors of channel inputs and channel outputs, respectively, for node \(i\) at time \(t\).

The variables in the \(\ell\)-th layer of the stack are denoted by an argument \(\ell\), for example \(W^{(i-j)}(\ell)\) is the message from node \(i\) to node \(j\) in the \(\ell\)-th layer of the stack and \(X^{(i)}(\ell)\) is the layer-\(\ell\) channel input from node \(i\) at time \(t\). Since the message \(W^{(i-j)}\) is an \(N\)-dimensional vector of messages, when \(W^{(i-j)} \in W^{(i-j)} - S(N)\), \(W^{(i-j)} \in W^{(i-j)} - S(N^N)\). We therefore define the rate \(R^{(i-j)}\) for a stacked network to be \(\log |W^{(i-j)}|/(nN)\); this normalization makes the rate of a network and its corresponding stacked network comparable.

**Definition 3** Let a network \(N^N - S(N^N)\) be given corresponding to a hypergraph \(G = (V, E)\). Let \(N^N\) be the \(N\)-fold stacked network for \(N\). A blocklength-\(n\) solution \(S(N^N)\) to this network is defined as a set of encoding and decoding functions:

\[
N^N \equiv \left( \prod_{i=1}^m X^{(i)}, \prod_{e \in E} p(y^{V_2(e)}|x^{V_1(e)}), \prod_{i=1}^m Y^{(i)} \right)
\]

mapping \((Y^{(i)}_1, \ldots, Y^{(i)}_n, W^{(i-1)}, \ldots, W^{(i-m)})\) to \(X^{(i)}\) for each \(t \in \{1, \ldots, n\}\) and \(i \in V\) and mapping \((Y^{(1)}_1, \ldots, Y^{(n)}_n, W^{(i-1)}, \ldots, W^{(i-m)})\) to \(W^{(i-j)}(\ell)\) for each \(i, j \in V\). The solution \(S(N^N)\) is called a \((\lambda, R)\)-solution for \(N^N\), denoted \((\lambda, R) - S(N^N)\), if the encoding and decoding functions imply \(\Pr(\cup_j W^{(i-j)}(\ell) \neq W^{(i-j)}(\ell)) < \lambda\) for all source and sink pairs \(i, j\).

According to the given definition, the \(N\) copies of each node \(v \in V\) work together in the \(N\)-fold stacked network – using the outgoing messages and channel outputs from all \(N\) layers of the network to create the channel input on each layer. This seems to make the \(N\)-fold stacked network \(N^N\) considerably more powerful than the network \(N\) from which it was derived. Notice, however, that
the error criterion for the stacked mappings is also more stringent; an error occurs in the message from node $i$ to node $j$ if an error occurs in one or more of the layers.

**Definition 4** The rate region $\mathcal{R}(\mathcal{N}) \subset \mathbb{R}^m_{\geq 0}$ of a stacked network $\mathcal{N}$ is the closure of all rate vectors $\mathcal{R}$ such that $S(\mathcal{N})$ solution exists for any $\lambda > 0$ and all $N$ sufficiently large.\(^2\)

**Theorem 1** The rate regions $\mathcal{R}(\mathcal{N})$ and $\mathcal{R}(\mathcal{N})$ are identical. Further, there exist solutions of type $(2^{-N\delta}, \mathcal{R}) - S(\mathcal{N})$ for each $\mathcal{R} \in \text{int}(\mathcal{R}(\mathcal{N}))$.

**Proof.** We first show that $\mathcal{R}(\mathcal{N}) \subseteq \mathcal{R}(\mathcal{N})$. Let $\lambda = \text{int}(\mathcal{R}(\mathcal{N}))$. Then for any $\lambda > 0$, there exists a $(\lambda, \mathcal{R}) - S(\mathcal{N})$ solution to the stacked network $\mathcal{N}$. Let $n$ be the blocklength of $S(\mathcal{N})$. We use $S(\mathcal{N})$ to build a blocklength $nN$ solution $\tilde{S}(\mathcal{N})$ for network $\mathcal{N}$. The solution $\tilde{S}(\mathcal{N})$ implements the $N$ layers of each time step of $S(\mathcal{N})$ in $N$ sequential time steps in $\mathcal{N}$. Showing that the error probability and rate of $\tilde{S}(\mathcal{N})$ on $\mathcal{N}$ equal the error probability and rate of $S(\mathcal{N})$ on $\mathcal{N}$ gives the desired result. The following description makes this precise.

Let $f^{(i-j)} : \{1, \ldots, 2^nN R^{(i-j)}\} \rightarrow \{1, \ldots, 2^nR^{(i-j)}\}$ be the natural one-to-one mappings between a single sequence of $NnR^{(i-j)}$ bits and $N$ consecutive subsequences each of $nR^{(i-j)}$ bits. Let $g^{(i-j)}$ be the inverse of $f^{(i-j)}$. We use $f^{(i-j)}$ to map messages from the message alphabet of the rate-$R^{(i-j)}$ blocklength-$nN$ code $S(\mathcal{N})$ to the message alphabet for the $N$-layer, rate-$R^{(i-j)}$, blocklength-$n$ code $S(\mathcal{N})$. The mapping is one-to-one since in each sequence the total number of bits transmitted from node $i$ to node $j$ is $NnR^{(i-j)}$. For each $k \in \{1, \ldots, n\}$, let

$$X^{(i)}(k) = (X^{(i)}(k-N+1), \ldots, X^{(i)}(k))$$

$$Y^{(i)}(k) = (Y^{(i)}(k-N+1), \ldots, Y^{(i)}(k))$$

be vectors containing the channel inputs and outputs at node $i$ for $N$ consecutive time steps beginning at time $(k-1)N + 1$. We define the solution $S(\mathcal{N})$ as

$$X^{(i)}(k) = X^{(i)}(k-N+1), \ldots, Y^{(i)}(k-N+1),$$

$$W^{(i-j)} = g^{(i-j)}(W^{(i-j)}(1), \ldots, W^{(i-j)}(n),$$

$$f^{(i-j)}(W^{(i-j)}(1), \ldots, W^{(i-j)}(n)).$$

Since $S(\mathcal{N})$ satisfies the causality constraints and operates precisely the mappings from $S(\mathcal{N})$ on $\mathcal{N}$, the solution $S(\mathcal{N})$ achieves the same rate and error probability on $\mathcal{N}$ as the solution $S(\mathcal{N})$ achieves on $\mathcal{N}$.

For the converse, the job is more difficult. A solution $(\lambda, \mathcal{R}) - S(\mathcal{N})$ needs to achieve an error probability of at most $\lambda$ for every $(i, j)$ pair in a network. A solution $(\lambda, \mathcal{R}) - S(\mathcal{N})$ also needs to achieve an error probability of at most $\lambda$ for each $(i, j)$, but here the error event is a union over errors in each of the $N$ layers with $N$ growing arbitrarily large.

Let $\mathcal{R} \in \text{int}(\mathcal{R}(\mathcal{N}))$, and fix some $\tilde{R} \in \text{int}(\mathcal{R}(\mathcal{N}))$ for which $\tilde{R}^{(i-j)} > R^{(i-j)}$ for all $i, j$. Set $\rho = \min_{i,j}(\tilde{R}^{(i-j)} - R^{(i-j)})$. For reasons that will become clear later, we wish to find a solution with error probability $\lambda$ and blocklength $n$ satisfying

$$\max_{i,j} \tilde{R}^{(i-j)}(\lambda + h(\lambda)/n < \rho.$$
channel output satisfies
\[ I(U^{(i-j)}(t); \tilde{U}^{(i-j)}(t)) = H(U^{(i-j)}(t)) - H(U^{(i-j)}(t); \tilde{U}^{(i-j)}(t)) \]
\[ = n R^{(i-j)} - H(U^{(i-j)}(t); \tilde{U}^{(i-j)}(t)) \]
\[ > n R^{(i-j)} - (\lambda n R^{(i-j)} + h(\lambda)) \]

by Fano’s inequality. For each \((i, j)\) we now apply a \(2^{n R^{(i-j)}(N)}\)-channel code across the \(N\) uses of this discrete memoryless channel corresponding to the \(N\) layers of our stacked network. The channel code is designed at random using the uniform distribution on the input alphabet \(U^{(i-j)} = \{1, \ldots, 2^{n R^{(i-j)}}\}^N\). The channel code’s encoder maps each block of messages \(W^{(i-j)} = (W^{(i-j)}(1), \ldots, W^{(i-j)}(N))\) to a block of channel inputs \(U^{(i-j)} = (U^{(i-j)}(1), \ldots, U^{(i-j)}(N))\). Note that the channel code’s rate per channel use \(n R^{(i-j)}\) is strictly less than the channel’s mutual information, precisely \(I(U^{(i-j)}(t); \tilde{U}^{(i-j)}(t)) > n R^{(i-j)} - (\lambda n R^{(i-j)} + h(\lambda)) > 0\); thus applying the strong coding theorem for discrete memoryless channels, we obtain an error probability that is bounded by \(2^{-N^\delta}\), where \(\delta\) is an increasing function of the gap \(\min_{i,j} [I(U^{(i-j)}(t); \tilde{U}^{(i-j)}(t)) - n R^{(i-j)}]\). \(\blacksquare\)

The proof of Theorem 1 shows not only that \(\mathcal{R}(N) = \mathcal{R}(N)\) but also that for any rate \(R \in \text{int}(\mathcal{R}(N))\) there exists a \((2^{-N^\delta}, R) - \mathcal{S}(\mathcal{N})\) solution for \(\mathcal{N}\) that first channel codes each message \(W^{(i-j)}\) and then sends the channel coded description \(\tilde{U}^{(i-j)}\) of \(W^{(i-j)}\) through the stacked network using the same solution \(\mathcal{S}(\mathcal{N})\) independently in each layer of the stacked network \(\mathcal{N}\). We henceforth restrict our attention to codes of this type; since the proof of Theorem 1 shows that these codes can obtain all rates in the interior of \(\mathcal{R}(\mathcal{N})\), there is no loss of generality in this restriction.

With this restriction, if \(U^{(i-j)}(1), \ldots, U^{(i-j)}(N)\) are independent and identically distributed (i.i.d.), then for each time \(t\)
\[ \{X^{(1)}(t), \ldots, X^{(m)}(t), Y^{(1)}(t), \ldots, Y^{(m)}(t)\} \]
are also i.i.d. for \(\ell \in \{1, \ldots, N\}\) since the solutions in the layers of \(\mathcal{N}\) are independent and similar. Note that our message inputs \(U^{(i-j)}(1), \ldots, U^{(i-j)}(N)\) are not the messages \(W^{(i-j)}(1), \ldots, W^{(i-j)}(N)\) (which are independent and uniformly distributed over \(\mathcal{W}^{(i-j)}(t)\) by assumption) but the channel coded versions of those messages. To maintain the desired i.i.d. structure across \(U^{(i-j)}(1), \ldots, U^{(i-j)}(N)\), we employ a random channel code design that draws each of the \(2^{n R^{(i-j)}}\) codewords in the codebook for \(i, j\) independently and uniformly at random from the set \(\tilde{U}^{(i-j)}\). Using this random code, the distribution on the message set \(\tilde{U}^{(i-j)}\) remains i.i.d. as desired. We put off the choice of a particular instance of this code until the proof of Theorem 4. Our argument employs random coding at multiple junctures. We make the decision about the instances of all random codes at once after all random codes are in place.

We now focus on the first factor in the transition probabilities,
\[ p(y(V^e(\ell))|x(V^e(\ell))) \prod_{e \in E \setminus \{e\}} p(y(V^e(\ell))|x(V^e(\ell))). \]

We here treat the case where \(p(y(V^e(\ell))|x(V^e(\ell))) = p(y(i,s)|x(j,r))\) corresponds to a single link from node \(i\) to node \(j\); we treat factors corresponding to hyperedges in future work. For simplicity, we drop the superscripts \((i, s)\) and \((j, r)\), referring only to random variables \(X, Y\) and their realizations \(x, y\) in the sequel.

Consider a \((2^{-N^\delta}, R) - \mathcal{S}(\mathcal{N})\) solution achieved by channel coding each message and then applying the same solution \(\mathcal{S}(\mathcal{N})\) in each layer of \(\mathcal{N}\) as in Theorem 1. Let \(n\) be the blocklength of solution \(\mathcal{S}(\mathcal{N})\), and for each \(t \in \{1, \ldots, n\}\), use \(A_{i,t}^N\) to denote the set of jointly typical channel input-output pairs with respect to the distribution \(p_t(x, y) = p_t(x)p(y|x)\) is the distribution imposed across the channel of interest at time \(t\) when we run solution \(\mathcal{S}(\mathcal{N})\) on the network \(\mathcal{N}\) and \(p_t(x, y)\) is the resulting product distribution across the layers. We next bound the fraction of jointly typical pairs \((x, y) \in A_{i,t}^N\) that give a large conditional error probability in the operation of the network according to solution \(\mathcal{S}(\mathcal{N})\).

For the set of mappings specified by the solution, let
\[ B_{i,t}^N(\tilde{\lambda}) = \{ (x_N, y_N) \in A_{i,t}^N : \Pr \left( \bigcup_{a,b} \{ W^{(a-b)} \neq \tilde{W}^{(a-b)} \} \right) \}
\[ \geq \tilde{\lambda} \}
\]
The definition of \(B_{i,t}^N(\tilde{\lambda})\) relies on the conditional probability of an error in running the solution \(\mathcal{S}(\mathcal{N})\) on the network \(\mathcal{N}\) of noisy channels. Lemma 2 bounds both the fraction of jointly typical channel input-output pairs that have high conditional error probability and the probability, with respect to the i.i.d. distribution \(p_t(x, y)\), of the remaining jointly typical pairs.

**Lemma 2** Assuming the existence of a \((2^{-N^\delta}, R) - \mathcal{S}(\mathcal{N})\) solution for a stacked network \(\mathcal{N}\), we have
\[ \frac{|B_{i,t}^N(\tilde{\lambda})|}{|A_{i,t}^N|} < \frac{m^2 2^{-N(\delta - 2\epsilon)}}{(1 - \epsilon)\tilde{\lambda}} \quad \text{and} \]
\[ \Pr(A_{i,t}^N \setminus B_{i,t}^N(\tilde{\lambda})) > (1 - \epsilon) - \frac{m^2 2^{-N(\delta - 4\epsilon)}}{(1 - \epsilon)\tilde{\lambda}} \]
for $N$ sufficiently large.³

Proof. Let $E = \bigcup_{n,t} \{ W^{(a-b)} \neq \tilde{W}^{(a-b)} \}$ denote the error event. Then $\Pr(E|B_{i,t}^{(N)}(\tilde{\lambda}) \cap E) \leq \Pr(A_{i,t}^{(N)} \cap E) \leq \Pr(E)$ which, by the union bound and lemma assumption, is at most $m^22^{-N\delta}$. (The assumption is itself valid for rates $R$ of interest by Theorem 1.) Thus

$$m^22^{-N\delta} \geq \Pr(E|B_{i,t}^{(N)}(\tilde{\lambda})) \Pr(B_{i,t}^{(N)}(\tilde{\lambda})) > \frac{\lambda |B_{i,t}^{(N)}(\tilde{\lambda})|2^{-N(H(X,Y)+\epsilon)}}{|A_{i,t}^{(N)}|}.$$

Then $|A_{i,t}^{(N)}|2^{-N(H(X,Y)+\epsilon)} > \Pr(A_{i,t}^{(N)}) > 1 - \epsilon$ for large $N$ by the AEP gives the first bound. For the second inequality,

$$\Pr(A_{i,t}^{(N)} \setminus B_{i,t}^{(N)}(\tilde{\lambda})) = \Pr(A_{i,t}^{(N)}) - \Pr(B_{i,t}^{(N)}(\tilde{\lambda})) > (1 - \epsilon) - |B_{i,t}^{(N)}(\tilde{\lambda})|2^{-N(H(X,Y)+\epsilon)} = (1 - \epsilon) - |A_{i,t}^{(N)}|2^{-N(H(X,Y)+\epsilon)}|B_{i,t}^{(N)}(\tilde{\lambda})|/|A_{i,t}^{(N)}| \geq (1 - \epsilon) - 2^N\epsilon|B_{i,t}^{(N)}(\tilde{\lambda})|/|A_{i,t}^{(N)}|.$$

In network $N$, the input $X_t$ to a channel $p(y|x)$ will usually result in a jointly typical channel output $Y_t$ with respect to the distribution $p_t(x,y)$. We argue in Theorem 4 that the functionality of the network $N$ remains unchanged if we replace the $N$ uses of the channel $p(y|x)$ at time $t$ by any other means to pick a jointly typical $Y_t$. Consider the source code with encoder $\alpha_{N,t}: X^N \rightarrow \{1, \ldots, 2^{NR}\}$ and decoder $\beta_{N,t}: \{1, \ldots, 2^{NR}\} \rightarrow Y^N$. The random decoder design draws $2^{NR}$ codewords $\beta_{N,t}(1), \ldots, \beta_{N,t}(2^{NR})$ i.i.d. from the distribution $p_t(y_N) = \prod_{i=1}^N p_t(y_i)$, where $p_t(y) = \sum_x p_t(x,y)$ is the marginal at time $t$ on the channel output $Y$ in a single layer of $N$ under the fixed solution $S(N)$. The encoder maps each vector $x_N$ to any index $k$ for which the pair $(x_N, \beta_{N,t}(k))$ is in the set $A_{i,t}^{(N)} \setminus B_{i,t}^{(N)}(\tilde{\lambda})$. If there is more than one such index, the choice is made uniformly at random. If there is no such index, $\alpha_{N,t}(x_N)$ is set to 1. Let function $K_t(x_N, y_N)$ (cf. steps 13.92-13.101 in [9]) be defined as $K_t(x_N, y_N) = 1$ if $(x_N, y_N) \in A_{i,t}^{(N)} \setminus B_{i,t}^{(N)}(\tilde{\lambda})$ and $K_t(x_N, y_N) = 0$ otherwise. Lemma 3 bounds the conditional probability

$$\hat{p}_t(y_N|x_N) = \Pr(\beta_{N,t}(\alpha_{N,t}(x_N)) = y_N|x_N = x_N)$$

established by the source code. We show that for $(x_N, y_N) \in A_{i,t}^{(N)} \setminus B_{i,t}^{(N)}(\tilde{\lambda})$, the conditional probability of $y_N$ given $x_N$ remains virtually unchanged by the replacement of the channel by a matching source code. Let

$$F_t(x_N) = \{|y_N: K_t(x_N, y_N) = 1\}|.$$

The upper bound in Lemma 3 is restricted to vectors $x_N$ for which $F_t(x_N) > (1 - \epsilon)2^{N(H(Y|X)-\epsilon)}$. The probability of observing an $x_N$ that meets this property approaches 1 as $N$ grows without bound, so we restrict our attention to this case and include all other $x_N$ among the error events. The derivation of Lemma 3 is the topic of the appendix.

Lemma 3 If $F_t(x_N) > (1 - \epsilon)2^{N(H(Y|X)-\epsilon)}$, then

$$\hat{p}_t(y_N|x_N) < p_t(y_N|x_N) < 2^{6Ne}/(1 - \epsilon)$$

for all $(x_N, y_N) \in A_{i,t}^{(N)} \setminus B_{i,t}^{(N)}(\tilde{\lambda})$. If $R > I(X; Y)$, then

$$\hat{p}_t(y_N|x_N) > (1 - \epsilon)p_t(y_N|x_N)2^{-6Ne}$$

for any $(x_N, y_N) \in A_{i,t}^{(N)} \setminus B_{i,t}^{(N)}(\tilde{\lambda})$ and any $\epsilon' > 0$ provided that $N$ is sufficiently large.

We finally are ready to state the main theorem. Since Theorem 4 treats the entire network rather than just a single link, where useful for clarity we reintroduce the superscripts $(i,s)$ and $(j,r)$ to our channel input $x$ and channel output $y$, respectively.

Theorem 4 Let networks $N$ and $\hat{N}$ be defined as

$$N = \left( X^{(1,1)} \times \cdots \times X^{(i,s)} \times \cdots \times X^{(m,d_{out}(m))}, \right.$$

$$p(y^{(j,r)}|x^{(i,s)}) \prod_{e \in E(N)} p(y^{(V_2(e))}|x^{(V_1(e))}),$$

$$\left. Y^{(l,1)} \times \cdots \times Y^{(j,r)} \times \cdots \times Y^{(m,d_{in}(m))} \right)$$

$$\hat{N} = \left( \hat{X}^{(1,1)} \times \cdots \times \hat{X}^{(i,s)} \times \cdots \times \hat{X}^{(m,d_{out}(m))}, \right.$$

$$\delta(x^{(i,s)} - y^{(j,r)}) \prod_{e \in E(N)} p(y^{(V_2(e))}|x^{(V_1(e))}),$$

$$\left. Y^{(l,1)} \times \cdots \times Y^{(j,r)} \times \cdots \times Y^{(m,d_{in}(m))} \right),$$

where $\hat{X}^{(i,s)}$, $\delta(x^{(i,s)} - y^{(j,r)})$, $Y^{(j,r)}$ is a noiseless bit pipe that in $n$ channel uses noiselessly maps $nR$ bits from its channel input to its channel output for some $R > C = \max_{p(x^{i,s})} I(X^{(i,s)}; Y^{(j,r)})$. Then $D(N) \subseteq D(\hat{N})$.³

³Given a set $A$, we use $|A|$ to denote the cardinality of $A$ if $A$ is countable and the volume of $A$ if $A$ is uncountably infinite. We likewise use notation $H(X, Y)$ for both discrete and differential entropy. We assume that $H(X, Y) < \infty$ so that $|A_{i,t}^{(N)}| < \infty$ by the AEP.
This proves the desired equivalence between the network of noisy channels and the network of noiseless, capacititated links since for any \( R < C, \mathcal{R}(N) \supseteq \mathcal{R}(\hat{N}) \) by Shannon’s channel coding theorem.

\begin{proof}
Let \( \mathcal{N} \) and \( \hat{\mathcal{N}} \) be the stacked networks for \( N \) and \( \hat{N} \). Since \( \mathcal{R}(\hat{N}) = \mathcal{R}(\mathcal{N}) \) and \( \mathcal{R}(\mathcal{N}) = \mathcal{R}(\hat{N}) \) by Theorem 1, it suffices to prove \( \mathcal{R}(\hat{N}) \subseteq \mathcal{R}(\mathcal{N}) \). For any \( R \in \text{int}(\mathcal{R}(\mathcal{N})) \) and any \( \lambda > 0 \), there exists a \((\lambda, R)\)-\( \mathcal{S}(\hat{N}) \) solution for network \( \hat{N} \). We must prove that for any \( \lambda > 0 \) there exists a \((\lambda, R)\)-\( \mathcal{S}(\mathcal{N}) \) solution for network \( \mathcal{N} \).

By Theorem 1, for any \( R \in \text{int}(\mathcal{R}(\mathcal{N})) \), there exists a \((2^{-N\delta}, R)\)-\( \mathcal{S}(\mathcal{N}) \) solution for \( N \) that applies the same solution \( \mathcal{S}(N) \) in every layer of \( \mathcal{N} \). Let \( n \) be the block-length of this solution, and for each time \( t \in \{1, \ldots, n\} \), let \( p_t(x_N^N, y_N^N) = \prod_{i=1}^N p_t(x_i, y_i) \) be the distribution on \( (X^{(i,s)})^N \times (Y^{(j,r)})^N \) at time \( t \) that results from running this \((2^{-N\delta}, R)\)-\( \mathcal{S}(\mathcal{N}) \) solution on network \( \mathcal{N} \).

For each \( t \in \{1, \ldots, n\} \), randomly design a source code \((\alpha_N, \beta_N)\) with \( 2^{N\delta} \) codewords drawn i.i.d. according to distribution \( p_t(x_N^N, y_N^N) \) and an encoder \( \alpha_{N,t} \) that maps each \( x_N^N \) to an index \( k \) for which \( (x_N^N, \beta_{N,t}(k)) \in A_{\mathcal{N}}^{(N)} \setminus B_{\mathcal{N}}^{(N)}(\hat{\lambda}) \) if available and index 1 otherwise. Here \( A_{\mathcal{N}}^{(N)} \) is the typical set defined with respect to distribution \( p_t(x_N^N, y_N^N) \) while \( B_{\mathcal{N}}^{(N)}(\hat{\lambda}) \) is the subset of those values for which the conditional error probability in the operation of \( \mathcal{N} \) is greater than \( \hat{\lambda} \).

We run the given \((2^{-N\delta}, R)\)-\( \mathcal{S}(\mathcal{N}) \) solution for \( N \) on \( \mathcal{N} \) using our source codes \( \{(\alpha_{N,t}, \beta_{N,t})\}_{t=1}^n \).

Precisely, node \( i \) encodes \( X^{(i,s)}_t \) to

\[
\hat{X}^{(i,s)}_t = \alpha_{N,t}(X^{(i,s)}_t),
\]

before transmission across channel \( (X^{(i,s)}, \delta(x^{(i,s)} - y^{(j,r)}), Y^{(j,r)}) \) to node \( j \) decodes channel output \( \hat{Y}^{(j,r)}_t = \beta_{N,t}(\hat{Y}^{(j,r)}_t) \)

before application of its usual mappings. We denote the resulting solution for \( \mathcal{N} \) by \( \mathcal{S}(\mathcal{N}) \).

We offer the following analysis for the error probability achieved by \( \mathcal{S}(\mathcal{N}) \) on \( \mathcal{N} \). The probability as calculated includes the randomness of all \( n \) source code designs in addition to the operation of the remaining random channels. For each \( t \in \{1, \ldots, n\} \), define \( c_1(t) \) and \( c_2(t) \) as

\[
c_1(t) \overset{\text{def}}{=} \Pr \left( (X^{(i,s)}_t, Y^{(j,r)}_t) \notin A_{\mathcal{N},t}^{(N)} \setminus (B_{\mathcal{N},t}^{(N)}(\hat{\lambda})) \right), \quad (X^{(i,s)}_t, Y^{(j,r)}_t) \notin A_{\mathcal{N},t}^{(N)} \setminus (B_{\mathcal{N},t}^{(N)}(\hat{\lambda})) \quad \forall t' < t
\]

\[
c_2(t) \overset{\text{def}}{=} \Pr \left( (X^{(i,s)}_t, Y^{(j,r)}_t) \notin A_{\mathcal{N},t}^{(N)} \setminus (B_{\mathcal{N},t}^{(N)}(\hat{\lambda})) \right), \quad (X^{(i,s)}_t, Y^{(j,r)}_t) \notin A_{\mathcal{N},t}^{(N)} \setminus (B_{\mathcal{N},t}^{(N)}(\hat{\lambda})) \quad \forall t' < t
\]

Further, define \( c_3 \) and \( c_4 \) as

\[
c_3 \overset{\text{def}}{=} \Pr \left( (X^{(i,s)}_t, Y^{(j,r)}_t) \notin A_{\mathcal{N},t}^{(N)} \setminus (B_{\mathcal{N},t}^{(N)}(\hat{\lambda})) \right), \quad \forall t \leq n
\]

\[
c_4 \overset{\text{def}}{=} \Pr \left( (X^{(i,s)}_t, Y^{(j,r)}_t) \notin A_{\mathcal{N},t}^{(N)} \setminus (B_{\mathcal{N},t}^{(N)}(\hat{\lambda})) \right), \quad \forall t \leq n
\]

Then

\[
\Pr \left( \cup_{t=1}^n \{ (X^{(i,s)}_t, Y^{(j,r)}_t) \notin A_{\mathcal{N},t}^{(N)} \setminus (B_{\mathcal{N},t}^{(N)}(\hat{\lambda})) \} \right)
\]

\[
\leq \sum_{t=1}^n c_1(t) c_2(t) + c_3 c_4,
\]

where the second inequality follows from the first since \( c_2(t) \) for all \( t \) and \( c_3 \) are probabilities and therefore bounded above by 1.

To bound the probability \( c_1(t) \) for each \( t \), we note that we are interested in the event \( (X^{(i,s)}_t, Y^{(j,r)}_t) \notin A_{\mathcal{N},t}^{(N)} \setminus (B_{\mathcal{N},t}^{(N)}(\hat{\lambda})) \) in the case where all prior source code input-output pairs are in the desired set. We bound these probabilities as

\[
c_1(t) = \Pr((X_t, \beta_N(\alpha_N(X_t))) \notin A_{\mathcal{N},t}^{(N)} \setminus (B_{\mathcal{N},t}^{(N)}(\hat{\lambda})), \quad (X_{t'}, \beta_N(\alpha_N(X_{t'}))) \in A_{\mathcal{N},t'}^{(N)} \setminus (B_{\mathcal{N},t'}^{(N)}(\hat{\lambda})), \quad \forall t' < t
\]

\[
< \sum_{x_N^N \in A^N} \hat{p}_t(x_N^N) \prod_{i=1}^n \prod_{y_N^N \in \beta_N^N} p_t(y_N^N, K_t(x_N^N, y_N^N)) 2^{N\delta}.
\]

The given bound reflects the probability that no codeword \( y_N^N \in \{\beta_N, \ldots, \beta_N(2^{N\delta})\} \) is jointly typical with the observed \( x_N^N \). The probability \( \hat{p}_t(x_N^N) \) captures the impact of observing \( (X_{t'}, \beta_N(\alpha_N(X_{t'}))) \) for \( t' < t \).

\[8\]
Next recall that \((1 - ab)^k \leq 1 - a + e^{-bk}\) and
\[
p_t(y^N) = p(y^N|x^N) p_t(y^N)p_t(x^N) \\
> p(y^N|x^N)2^{-N(I(x^N;Y^N) + 3\epsilon)}
\]
for all \((x^N, y^N) \in A_{t}^{(N)}\) [9]. Here \(I(x^N;Y^N)\) is the mutual information for the distribution \(p_t(x,y) = p_t(x)p(y|x)\) across the given channel at time \(t\). Note that \(I(x^{(i,s)}_t;Y^{(j,r)}_t) \leq C < R\). Thus,
\[
c_1(t) \\
< \sum_{y^N} \hat{p}_t(x^N) \left(1 - \sum_{y^N} p(y^N|x^N) K_t(x^N, y^N) - e^{-2N(R - I(x^{(i,s)}_t;Y^{(j,r)}_t) - 3\epsilon)}\right) \\
< \epsilon + 2^{i(t-1)6N^2} m^{22 - N(\delta - 4\epsilon)} + e^{-2N(R - I(x^{(i,s)}_t;Y^{(j,r)}_t) - 3\epsilon)}
\]
since observing typical outcomes at all prior time steps cannot increase the probability of an atypical outcome at time \(t\). The bound on the probability of each element in \(A_{t}^{(N)} \setminus B_{\epsilon,t}^{(N)}(\hat{\lambda})\) follows from Lemma 2 and the application of Lemma 3 for each \(t' < t\). Finally,
\[
(x^{(i,s)}, \beta_{N,t}(\alpha_{N,t}(x^{(i,s)}))) \in A_{t}^{(N)} \setminus B_{\epsilon,t}^{(N)}(\hat{\lambda})
\]
for all \(t\) implies \(c_1(\epsilon) \leq \hat{\lambda}\), giving
\[
Pr\left(\bigcup_{(a,b)} W(a-b) \neq \tilde{W}(a-b)\right) \\
\leq \left[\sum_{t=1}^{n} c_1(t) + c_4\right] \\
\leq ne + \sum_{t=1}^{n} \left(2^{(t-1)6N^2} m^{22 - N(\delta - 4\epsilon)} + e^{-2N(R - I(x^{(i,s)}_t;Y^{(j,r)}_t) - 3\epsilon)} + \hat{\lambda}\right) \\
\leq ne + \frac{nm^{22 - N(\delta - (6n - 2)\epsilon)}}{(1 - e)\lambda} + \hat{\lambda} \\
+ ne^{-2N(R - I(x^{(i,s)}_t;Y^{(j,r)}_t) - 3\epsilon)} + \hat{\lambda},
\]
Recall that \(n\) and \(\delta\) are the blocklength and error exponent associated with our solution \(S(N)\); these values were fixed a priori. The remaining parameters \(\epsilon\), \(N\) and \(\hat{\lambda}\) are chosen for the design of our source code. The given error bound can be made arbitrarily small by setting \(\epsilon < \delta/(6n - 2)\) and \(\hat{\lambda}\) sufficiently small and then letting \(N\) grow without bound.

The given error probability is actually an expectation with respect to a large family of randomly drawn codes. Now consider the single distribution that simultaneously governs all independent random code choices. Since the expected error probability with respect to this distribution can be made arbitrarily small, there must be a single instance of all codes that does at least as well. Note that we here consider a single distribution governing over all random code choices and therefore choose the instance of all codes simultaneously rather than sequentially. This is necessary to guarantee good joint performance of the many codes.

**Remark 2** It is interesting to specify the choice of parameters in the above proof. Assume we want to guarantee the existence of a \((\hat{\lambda}, \hat{R})\) \(- S(N)\) solution for an arbitrary \(\hat{\lambda}\) and \(\hat{R} \in \text{int}(\mathcal{A}(N))\). Since we have \(\mathcal{R} \in \text{int}(\mathcal{A}(N))\) there exists \(\mathcal{R}' \in \mathcal{A}(N)\) with \(\mathcal{R}' > \mathcal{R}\), and we choose \(\rho\) in Theorem 1 accordingly as \(\min\{\mathcal{R}' - \mathcal{R}\}\). Once \(\rho\) is chosen, we choose \(\lambda\) and \(n\) so that the condition \(\rho > \max_{N,t}(\{R^{(i,j)}\})\lambda + h(\lambda)/n\) is satisfied for a \((\lambda, \mathcal{R})\) \(- S(N)\) solution of blocklength \(n\). Note that \(R^{(i,j)}\) is less than capacity of the channel \(p(W^{(i-j)}|W^{(i-j)})\) imposed by this solution, so \(\delta > 0\). We finally choose \(\epsilon < \min\{\hat{\lambda}/(6n - 2), \delta/(3 + n)\}\) and \(\hat{\lambda} < \hat{\lambda}/3\) and choose \(N\) sufficiently large. The resulting error probability satisfies \(P_e < \hat{\lambda}\).

**Appendix**

In this appendix, we bound the conditional distribution \(\hat{p}_t(y^N|x^N)\) that results from the application of source code \((\alpha_{N,t}, \beta_{N,t})\). A single instance of this and all other random codes in the network will be chosen when all random codes are in place, but for now the joint distribution on inputs and outputs is taken with respect to a random code. Lemma 5 characterizes \(\hat{p}_t(y^N|x^N)\) as a function of the probability \(q_t(x^N) = \sum_{y^N} K(x^N, y^N)p_t(y^N)\) that a single codeword drawn at random is jointly typical with vector \(x^N\).

**Lemma 5** Let \((\alpha_{N,t}, \beta_{N,t})\) be a random source code with codewords
\[
\{\beta_{N,t}(1), \ldots, \beta_{N,t}(2^{2NR})\}
\]
drawn i.i.d. according to distribution \(p_t(y^N)\) and encoder \(\alpha_{N,t}(x^N)\) chosen uniformly at random from the set
\[
\{k : (x^N, \beta_{N,t}(k)) \in A_{t}^{(N)} \setminus B_{\epsilon,t}^{(N)}(\hat{\lambda})\}
\]
if that set is not empty and \(\alpha_{N,t}(x^N) = 1\) otherwise. Then for any \((x^N, y^N) \in A_{t}^{(N)} \setminus B_{\epsilon,t}^{(N)}(\hat{\lambda})\),
\[
\hat{p}_t(y^N|x^N) = p_t(y^N) \frac{1 - q_t(x^N)}{q_t(x^N)}.\]
Proof. Recall that \( q_t(x^N) \) is the probability that a single randomly drawn codeword \( Y^N \) satisfies \( (x^N,Y^N) \in A_{e,t}^N \setminus B_{e,t}^N(\bar{\lambda}) \). Given our random code design, for any \((x^N,y^N) \in A_{e,t}^N \setminus B_{e,t}^N(\bar{\lambda})\),
\[
\hat{p}_t(y^N|x^N) = \sum_{j=1}^{2NR} \sum_{k=1}^j \binom{2NR}{j} \binom{j}{k} (1 - q_t(x^N) )^{2NR-j} (q_t(x^N) - p_t(y^N))^{j-k} (p_t(y^N))^k \frac{\partial}{\partial b}[(a+b)^j - a^j] 
\]
\[
= p_t(y^N) \sum_{j=1}^{2NR} \binom{2NR}{j} \frac{1}{j} (1 - q_t(x^N) )^{2NR-j} - \frac{1}{2} (q_t(x^N))^j - \frac{1}{2} \frac{\partial}{\partial b}[(a+b)^j - a^j] 
\]
\[
= p_t(y^N) \sum_{j=1}^{2NR} \binom{2NR}{j} \frac{1}{j} (1 - q_t(x^N) )^{2NR-j} - \frac{1}{2} (q_t(x^N))^j - \frac{1}{2} \frac{1}{j} (q_t(x^N))^j 
\]
\[
= p_t(y^N) \frac{1 - (1 - q_t(x^N) )^{2NR} }{q_t(x^N)} .
\]

Since the vectors that contribute to the probability \( q_t(x^N) \) are approximately equally probable, it is useful to understand the size of set \( \{y^N: K_t(x^N,y^N) = 1\} \). Let
\[
F_t(x^N) = | \{y^N: K_t(x^N,y^N) = 1\} | 
\]
and
\[
C_t(x^N) = \{x^N: F_t(x^N) > (1 - \epsilon)^{2N(H(Y|X)-\epsilon)} \}.
\]

Lemma 6 shows that for any \( \epsilon > 0 \), \( \Pr(C_t(x^N)) \) approaches 1 as \( N \) grows without bound.

**Lemma 6** Let \( X_1, \ldots, X_N \) be drawn i.i.d. according to distribution \( p_t(x) \). Then for any \( \epsilon > 0 \),
\[
\lim_{N \to \infty} \Pr(C_t(x^N)) = 1.
\]
Since all vectors $y^N$ contributing to the probability $q_t(x^N)$ are approximately equally probable, for all $x^N \in C^N$, we bound $q_t(x^N)$ as
\[
q_t(x^N) > F_t(x^N) 2^{-N(H(Y)+\epsilon)} > (1-\epsilon) 2^{-N(I(X;Y)+3\epsilon)}.
\]

Since $\hat{\mu}_t(y^N|x^N)$ is decreasing in $q_t(x^N)$ and $(x^N, y^N)$ are typical by assumption, we bound $\hat{\mu}_t(y^N|x^N)$ as
\[
\hat{\mu}_t(y^N|x^N) = p_t(y^N) 1 - (1 - q_t(x^N))^2^{NR} q_t(x^N)
< p_t(y^N|x^N) p_t(x^N)p_t(y^N) p_t(x^N, y^N)
\cdot 1 - (1 - (1 - \epsilon) 2^{-N(I(X;Y)+3\epsilon)} 2^{NR})
\cdot (1 - \epsilon) 2^{-N(I(X;Y)+3\epsilon)}
< p_t(y^N|x^N) 2^{-N(I(X;Y)-3\epsilon)}
\cdot 1
\cdot (1 - \epsilon) 2^{-N(I(X;Y)+3\epsilon)}
= p_t(y^N|x^N) 2^{6N\epsilon}/(1 - \epsilon).
\]

The derivation of the lower bound is similar. For any $x^N \in X^N$,
\[
1 > F_t(x^N) 2^{-N(H(Y)+2\epsilon)}
\]
and therefore
\[
F_t(x^N) < 2^{N(H(Y)+2\epsilon)}.
\]

Thus
\[
q_t(x^N) < F_t(x^N) 2^{-N(H(Y)-\epsilon)} < 2^{-N(I(X;Y)-3\epsilon)},
\]
and for all $(x^N, y^N) \in A^N \setminus B^N(\lambda)$, we bound $\hat{\mu}_t(y^N|x^N)$ as
\[
\hat{\mu}_t(y^N|x^N) = p_t(y^N) 1 - (1 - q_t(x^N))^2^{NR} q_t(x^N)
\geq p_t(y^N|x^N) p_t(x^N)p_t(y^N) p_t(x^N, y^N)
\cdot 1 - (1 - 2^{-N(I(X;Y)+3\epsilon)} 2^{NR})
\cdot 2^{-N(I(X;Y)+3\epsilon)}
\geq p_t(y^N|x^N) 2^{-N(I(X;Y)-3\epsilon)}
\cdot 1 - e^{-2^{-N(I(X;Y)+3\epsilon)} 2^{NR}}
\cdot 2^{-N(I(X;Y)+3\epsilon)}
\geq (1 - \epsilon') p_t(y^N|x^N) 2^{-6N\epsilon}
\]
for any $\epsilon' > 0$ provided $R > I(X;Y)$ and $N$ is sufficiently large.

\begin{thebibliography}{99}

\end{thebibliography}