

# An Achievable Rate for the Multiple Level Relay Channel<sup>\*†</sup>

Liang-Liang Xie<sup>‡</sup> and P. R. Kumar<sup>§</sup>

Department of Electrical and Computer Engineering, and  
Coordinated Science Laboratory  
University of Illinois, Urbana-Champaign

Submitted to *IEEE Transactions on Information Theory*

November 29, 2003

## Abstract

For the multiple level relay channel, an achievable rate formula is proved. We present a new coding scheme which is different from that in [3]. It is not only simpler, but it is also easier to extend to the multiple level case. Also, generally higher rates can be achieved with this coding scheme in the multiple level relay case than previously known. For a class of degraded channels, this achievable rate is shown to be the exact capacity. An application of the coding scheme to the allcast problem is also discussed.

---

<sup>\*</sup>Please address all correspondence to the second author at the address shown below.

<sup>†</sup>This material is based upon work partially supported by USARO under Contract Nos. DAAD19-00-1-0466 and DAAD19-01010-465, DARPA under Contract Nos. N00014-01-1-0576 and F33615-01-C-1905, AFOSR under Contract No. F49620-02-1-0217, DARPA/AFOSR under Contract No. F49620-02-1-0325, and NSF under Contract Nos. NSF ANI 02-21357 and CCR-0325716.

<sup>‡</sup>Institute of Systems Science, Chinese Academy of Sciences, Beijing 100080, China. This work was conducted while he was visiting the Coordinated Science Laboratory, University of Illinois.

<sup>§</sup>University of Illinois, Coordinated Science Laboratory, 1308 West Main St., Urbana, IL 61801-2307, USA. Email: prkumar@uiuc.edu. Web: <http://black.csl.uiuc.edu/~prkumar>.

# 1 Introduction

The relay channel was introduced by van der Meulen [1], [2]. The simplest case, shown in Figure 1, is the three node scenario where Node 1 functions purely as a relay to help the information transmission from Node 0 to Node 2. An immediate application of this framework, for instance, is in wireless communications, where a node is placed between the source node and the destination node, in order to shorten the distance of a hop, which has implications in terms of the amount of traffic carried, interference, power consumption, etc. In [1] a special discrete memoryless relay channel is even constructed for which no reliable information transmission is possible without the help of the relay node.

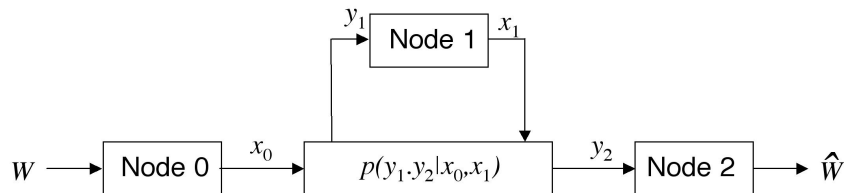


Figure 1: The one relay channel.

The simplest discrete memoryless one relay channel is depicted in Figure 1, where Nodes 0, 1, and 2 are the source, the relay, and the destination, respectively. This channel can be denoted by  $(\mathcal{X}_0 \times \mathcal{X}_1, p(y_1, y_2 | x_0, x_1), \mathcal{Y}_1 \times \mathcal{Y}_2)$ , where  $\mathcal{X}_0, \mathcal{X}_1$  are the transmitter alphabets of Nodes 0 and 1 respectively,  $\mathcal{Y}_1$  and  $\mathcal{Y}_2$  are the receiver alphabets of Nodes 1 and 2 respectively, and a collection of probability distributions  $p(\cdot, \cdot | x_0, x_1)$  on  $\mathcal{Y}_1 \times \mathcal{Y}_2$ , one for each  $(x_0, x_1) \in \mathcal{X}_0 \times \mathcal{X}_1$ . The interpretation is that  $x_0$  is the input to the channel from the source Node 0,  $y_2$  is the output of the channel to the destination Node 2, and  $y_1$  is the output received by the relay Node 1. After processing  $y_1$ , the relay Node 1 sends the input  $x_1$  chosen as a function of its past parameters,

$$x_1(t) = f_t(y_1(t-1), y_1(t-2), \dots), \quad \text{for every } t, \quad (1)$$

where  $f_t(\cdot)$  can be any causal function. Note that a one-step time delay is assumed in (1) to account for the signal processing time by the relay.

For the one relay channel described above, the best result to date is still [3], where the most remarkable conclusion is that the following rate is

achievable:

$$R < \max_{p(x_0, x_1)} \min\{I(X_0; Y_1 | X_1), I(X_0, X_1; Y_2)\}. \quad (2)$$

It is worth noting that the coding scheme to achieve the above rate is *not* simply multi-hop. At first, information goes from the source to the relay, and then from the relay to the destination. However, the destination needs to take into account both the inputs by the source and the relay in order to achieve (2). One can imagine that in the wireless relay channel example mentioned earlier, the destination node can make use of the signal coming directly from the source node, even though it may be relatively weaker than that from the relay node due perhaps to its greater distance from the source.

Moreover, [3] also proved that if  $X_0 \rightarrow (X_1, Y_1) \rightarrow Y_2$  forms a Markov chain, i.e.,

$$p(y_2 | y_1, x_0, x_1) = p(y_2 | y_1, x_1), \quad (3)$$

then the right hand side of (2) is the capacity of this physically degraded relay channel. However, the capacity for the general non-degraded case is still unknown. A recent study [4] of the so-called Gaussian parallel relay channel showed an interesting result: The asymptotically optimal coding scheme dramatically depends on the relative locations of the nodes, or equivalently, on the relative amplitudes of their signal-to-noise ratios (SNRs). To some extent, this result excludes the possibility of the existence of a unifying optimal coding scheme.

Actually, in wireless networks there can be more than one relay node. For instance, the Gaussian parallel relay channel considered in [4] consists of two relay nodes. The general framework would be that there are multiple levels of relays and each level consists of one or more nodes [5, 6]. This general multiple level relay channel was addressed in [5], where the coding scheme of [3] was extended, and an achievable rate formula in a recursive constraint form was proved. Later, by the same coding scheme, and for a special physically degraded Gaussian relay channel, [8] presented a specific formula which was shown to achieve the capacity.

In [6], we proposed a new coding scheme and presented the achievability proof for the Gaussian case. The scheme is simpler and avoids some inconvenient techniques (e.g., the Slepian-Wolf partitioning), although it coincides with [3] in giving the same achievable rate formula for the one relay case. More importantly, this new coding scheme is easier to extend to the multiple level relay case, and generally achieves higher rates than those proved in

[5]. This paper presents the results for the discrete memoryless case, relay systems with feedback, and the degraded versions of these systems.

The advantages of the new coding scheme of [6] have also been recognized by [7], where the corresponding achievable rate formula for the discrete memoryless case is also stated. The paper [7] also goes on to obtain the capacity of some relay channels under fading, which is the first significant capacity result for such channels, and one which may possibly constitute a breakthrough in the field.

Although the proof of the discrete memoryless case follows in a similar style as that of the Gaussian case presented in [6], there are differences: the proof of the Gaussian case can be made even simpler, e.g., in the construction of the codebooks, due to the special character of the Gaussian framework.

Another feature worth much emphasizing is that the decoding of the new scheme is a *simultaneous* typicality check of the signals of the previous several blocks, not in a sequential manner as that of [3]. (The details should be clear in the proof.) This technique is instrumental in making the generation of codebooks much simpler.

In this paper, we focus our discussion on discrete memoryless channels. Some corresponding results for multiple level Gaussian relay channels have been proved in Theorems 3.11-3.12 in [6], which includes the formula in [8] as a special case.

## 2 Models of multiple level relay channels

We begin with a definition of the discrete memoryless multiple level relay channel. Consider a channel with  $M + 1$  nodes. Let the source node be denoted by 0, the destination node by  $M$ , and let the other  $M - 1$  nodes be denoted sequentially as  $1, 2, \dots, M - 1$  in arbitrary order. Assume each Node  $i \in \{0, 1, \dots, M - 1\}$  sends  $x_i(t) \in \mathcal{X}_i$  at time  $t$ , and each Node  $k \in \{1, 2, \dots, M\}$  receives  $y_k(t) \in \mathcal{Y}_k$  at time  $t$ , where the finite sets  $\mathcal{X}_i$  and  $\mathcal{Y}_k$  are the corresponding input and output alphabets for the corresponding nodes. The channel dynamics is described by the following conditional probability mass function:

$$p(y_1, y_2, \dots, y_M | x_0, x_1, \dots, x_{M-1}), \quad (4)$$

for all  $(x_0, \dots, x_{M-1}) \in \mathcal{X}_0 \times \dots \times \mathcal{X}_{M-1}$  and  $(y_1, \dots, y_M) \in \mathcal{Y}_1 \times \dots \times \mathcal{Y}_M$ . Finally, we assume a one-step time delay at every relay node to account for

the signal processing time, so that for all  $i \in \{1, 2, \dots, M - 1\}$ ,

$$x_i(t) = f_{i,t}(y_i(t - 1), y_i(t - 2), \dots), \quad \text{for all } t,$$

where  $f_{i,t}$  can be any causal function. Figure 2 depicts this scenario.

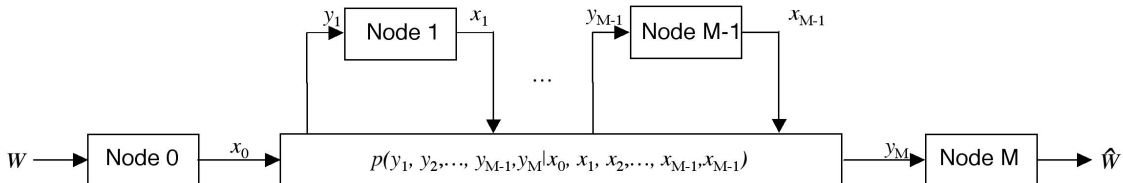


Figure 2: The multiple level relay channel.

**Remark 2.1** *Note that for simplicity in the above channel formulation we have not allowed any output  $y_0$  from the channel to the source Node 0, or any input  $x_M$  to the channel from the destination Node M. The involvement of  $y_0$  and  $x_M$  is equivalent to allowing feedback in the channel since it allows the modeling of systems where all nodes can both hear as well as transmit. Both these seemingly more general formulations are however actually covered by our formulation where node 0 cannot hear, and node M cannot transmit. To see this, simply consider a situation where node  $M - 1$  serves as a surrogate for node M in that node M simply hears  $y_M = (x_{M-1}, y_{M-1})$ . Then nodes  $M - 1$  and M have exactly the same information. Thus  $M - 1$  serves as a transmission-capable surrogate for node M, which is simply a dummy node. In the same vein, suppose that node 1 is a surrogate for node 0 in that node 0 can transmit  $W$  directly (or close to that through a high bandwidth link) which only node 1 can hear, but none of the other nodes. Then node 1 serves as listening-capable surrogate for node 0, which is the only outlet that node 0 has to the rest of the nodes, and node 0 simply becomes a dummy. Thus the system with nodes  $\{0, 1, 2, \dots, M - 2, M - 1, M\}$  where node 0 cannot hear and node M cannot transmit, is simply the same as the system with nodes  $\{1, 2, \dots, M - 2, M - 1\}$ , where node 1 is the source which is also hearing capable, and node  $M - 1$  is the destination which is transmission capable.*

The following definitions of codes and achievable rates are standard.

**Definition 2.1** *A  $(2^{TR}, T, \lambda_T)$  code for a discrete memoryless multiple level relay channel consists of the following:*

1. A random variable  $W$  with  $P(W = k) = \frac{1}{2^{TR}}$ , for every  $k \in \{1, 2, \dots, 2^{TR}\}$ .
2. An encoding function  $F_0 : \{1, 2, \dots, 2^{TR}\} \rightarrow \mathcal{X}_0^T$  for the source Node 0, and relay functions  $f_{i,t} : \mathcal{Y}_i^{t-1} \rightarrow \mathcal{X}_i$ ,  $t = 2, \dots, T$  for all the relay Nodes  $i \in \{1, 2, \dots, M-1\}$ , such that

$$\mathbf{x}_0 := [x_0(1), x_0(2), \dots, x_0(T)] = F_0(W), \quad t = 1, 2, \dots, T; \quad (5)$$

$$x_i(t) = f_{i,t}(y_i(t-1), \dots, y_i(1)), \quad \text{with } x_i(1) \text{ any element in } \mathcal{X}_i. \quad (6)$$

3. A decoding function  $g : \mathcal{Y}_M^T \rightarrow \{1, 2, \dots, 2^{TR}\}$  for the destination Node  $M$ .
4. The maximal probability of error:

$$\lambda_T := \max_{k \in \{1, 2, \dots, 2^{TR}\}} \text{Prob}\{g(\mathbf{y}_M) \neq k | W = k\}, \quad (7)$$

where  $\mathbf{y}_M := [y_M(1), y_M(2), \dots, y_M(T)]$ .

**Definition 2.2** A rate  $R > 0$  is said to be achievable if there exists a sequence of  $(2^{TR}, T, \lambda_T)$  codes such that the maximal probability of error  $\lambda_T$  tends to zero as  $T \rightarrow \infty$ .

We will also consider the *degraded* version of our channel:

**Definition 2.3** A discrete memoryless multiple level relay channel is said to be degraded if

$$p(y_{k+1}, \dots, y_M | y_k, x_0, \dots, x_{k-1}, x_k, \dots, x_{M-1}) = p(y_{k+1}, \dots, y_M | y_k, x_k, \dots, x_{M-1}),$$

for  $k = 1, \dots, M-1$ .

(8)

Equivalently, (8) means that  $(X_0, \dots, X_{k-1}) \rightarrow (Y_k, X_k, \dots, X_{M-1}) \rightarrow (Y_{k+1}, \dots, Y_M)$  forms a Markov chain for every  $k = 1, \dots, M-1$ . In the case of  $M = 2$ , (8) reduces to (3).

**Remark 2.2** If node  $M$  can transmit, then we simply extend the conditioning on both sides of (8) to include  $x_M$ . We note then that a degraded system remains degraded under the embedding procedure of Remark 2.1, and conversely. So our formulation allows the treatment of degraded systems with feedback.

### 3 The main results

We state the main theorem of this paper:

**Theorem 3.1** *For the discrete memoryless multiple level relay channel defined above, the following rate is achievable:*

$$R < \max_{p(x_0, x_1, \dots, x_{M-1})} \min_{1 \leq k \leq M} I(X_0, \dots, X_{k-1}; Y_k | X_k, \dots, X_{M-1}). \quad (9)$$

**Remark 3.1** *Generally, (9) achieves higher rates than the results given in [5]. To see this, first consider the 2-level relay channel example ( $M=3$ ) given in [5], which shows that a rate  $R_0$  is achievable if there exist  $R_1$ ,  $R_2$  and some  $p(x_0, x_1, x_2)$  such that*

$$R_0 < I(X_0; Y_1 | X_1, X_2);$$

and

$$\begin{aligned} R_1 &< I(X_1; Y_2 | X_2), \\ R_0 &< I(X_0; Y_2 | X_1, X_2) + R_1; \end{aligned}$$

and

$$\begin{aligned} R_2 &< I(X_2; Y_3), \\ R_1 &< I(X_1; Y_3 | X_2) + R_2, \\ R_0 &< I(X_0; Y_3 | X_1, X_2) + R_1. \end{aligned}$$

*It is easy to check that this will reduce to (9) if and only if  $I(X_1; Y_2 | X_2) = I(X_1, X_2; Y_3)$ , which generally does not hold. The same reasoning applies to the general case in Theorem 2.1 in [5].*

For the degraded discrete memoryless multiple level relay channel, the following theorem shows that the R.H.S. of (9) is actually the capacity:

**Theorem 3.2** *The capacity of the degraded discrete memoryless multiple level relay channel is*

$$C = \max_{p(x_0, x_1, \dots, x_{M-1})} \min_{1 \leq k \leq M} I(X_0, \dots, X_{k-1}; Y_k | X_k, \dots, X_{M-1}). \quad (10)$$

**Remark 3.2** *In the case of one relay, i.e.,  $M=2$ , Theorem 3.2 reduces to Theorem 1 in [3].*

To illustrate the case of feedback, we now turn to the more general setting where the destination node  $M$  also has an input  $x_M \in \mathcal{X}_M$  to the channel. Then it follows immediately from Theorem 3.1 that the following rate is achievable:

$$R < \max_{x_M \in \mathcal{X}_M} \max_{p(x_0, x_1, \dots, x_{M-1})} \min_{1 \leq k \leq M} I(X_0, \dots, X_{k-1}; Y_k | X_k, \dots, X_{M-1}, x_M). \quad (11)$$

However, we can achieve higher rates than the right hand side of (11) as stated in the following theorem:

**Theorem 3.3** *For the discrete memoryless multiple level relay channel defined above with an additional input  $x_M \in \mathcal{X}_M$  from the destination node  $M$ , the following rate is achievable:*

$$R < \max_{p(x_0, x_1, \dots, x_M)} \min_{1 \leq k \leq M} I(X_0, \dots, X_{k-1}; Y_k | X_k, \dots, X_M). \quad (12)$$

**Remark 3.3** *Noting the embedding procedure of Remark 2.1, and the preservation of the degraded property under embedding as noted in Remark 2.2, it follows that (12) also achieves the capacity for the degraded case where node  $M$  can also transmit, i.e.,  $x_M$  also exists.*

Next, we show that (12) generally achieves larger rates than (11). First, it is obvious that (12) is at least as large as (11), since we can always choose  $\text{Prob}(X_M = x_M) = 1$  for any  $x_M \in \mathcal{X}_M$ . Second, the following example shows that the right hand side of (12) can indeed be larger.

*Example 1.* Consider a simple channel with  $M = 2$ , and  $\mathcal{X}_0 = \mathcal{X}_1 = \mathcal{X}_2 = \mathcal{Y}_1 = \mathcal{Y}_2 = \{0, 1\}$ . The corresponding  $p(y_1, y_2 | x_0, x_1, x_2)$ 's are shown in Table 1:

$(x_0, x_1, x_2) \setminus (y_1, y_2)$	(0,0)	(0,1)	(1,0)	(1,1)
(0,0,0)	1	0	0	0
(0,1,0)	0	1	0	0
(1,0,0)	1	0	0	0
(1,1,0)	0	1	0	0
(0,0,1)	1	0	0	0
(0,1,1)	1	0	0	0
(1,0,1)	0	0	1	0
(1,1,1)	0	0	1	0



Table 1: The channel dynamics  $p(y_1, y_2|x_0, x_1, x_2)$  of the one relay channel in Example 1.

It is easy to check that  $\text{Prob}(Y_1 = 0|x_2 = 0) = 1$  and  $\text{Prob}(Y_2 = 0|x_2 = 1) = 1$ . Hence  $I(X_0; Y_1|X_1, x_2 = 0) = 0$  for every  $p(x_0, x_1)$ , and  $I(X_0, X_1; Y_2|x_2 = 1) = 0$  for every  $p(x_0, x_1)$ . Therefore,

$$\max_{x_2 \in \{0,1\}} \max_{p(x_0, x_1)} \min \{I(X_0; Y_1|X_1, x_2), I(X_0, X_1; Y_2|x_2)\} = 0. \quad (13)$$

Moreover, it is easy to check that  $\text{Prob}(Y_1 = X_0|x_2 = 1) = 1$  and  $\text{Prob}(Y_2 = X_1|x_2 = 0) = 1$ . Hence

$$\begin{aligned} I(X_0; Y_1|X_1, X_2) &= I(X_0; Y_1|X_1, x_2 = 0)\text{Prob}(X_2 = 0) + I(X_0; Y_1|X_1, x_2 = 1)\text{Prob}(X_2 = 1) \\ &= H(X_0|X_1, x_2 = 1)\text{Prob}(X_2 = 1); \\ I(X_0, X_1; Y_2|X_2) &= I(X_0, X_1; Y_2|x_2 = 0)\text{Prob}(X_2 = 0) + I(X_0, X_1; Y_2|x_2 = 1)\text{Prob}(X_2 = 1) \\ &= H(X_1|x_2 = 0)\text{Prob}(X_2 = 0). \end{aligned}$$

Therefore,

$$\begin{aligned} &\max_{p(x_0, x_1, x_2)} \min \{I(X_0; Y_1|X_1, X_2), I(X_0, X_1; Y_2|X_2)\} \\ &= \max_{p(x_0, x_1, x_2)} \min \{H(X_0|X_1, x_2 = 1)\text{Prob}(X_2 = 1), H(X_1|x_2 = 0)\text{Prob}(X_2 = 0)\} \\ &\leq \max_{p(x_2)} \min \{\text{Prob}(X_2 = 1), \text{Prob}(X_2 = 0)\} = \frac{1}{2}, \end{aligned} \quad (14)$$

where the equality in (14) can be achieved by letting  $p(x_0, x_1, x_2) = p(x_0)p(x_1)p(x_2)$ , and  $\text{Prob}(X_0 = 0) = \text{Prob}(X_1 = 0) = \frac{1}{2}$ .

Finally, from (13) and (14), (12) is seen to be strictly larger than (11) in this example.  $\square$

An intuitive interpretation of (9) or (12) comes directly from the coding scheme used in the achievability proof of Theorem 3.1. We can imagine that there is an information flow from the source Node 0 to the destination Node  $M$  along the path  $0 \rightarrow 1 \rightarrow \dots \rightarrow M$ . Each Node  $i$  decodes the information one-step (actually one time block in the coding) before the next Node  $i + 1$ . Hence, by the time the information reaches Node  $k$ , all the upstream nodes (i.e., nodes with smaller index than  $k$ ) have already obtained the same information and can therefore cooperate. This results in the achievability of any rate  $R < I(X_0, \dots, X_{k-1}; Y_k|X_k, \dots, X_M)$ , where the conditioning is due to the same reason: the downstream nodes (with larger index than  $k$ ) get no

more information than Node  $k$ , and therefore their inputs are predictable by Node  $k$ .

Since the order of the nodes, except the source, can be arbitrarily chosen, by the above interpretation of the coding scheme, it follows that to increase (9) one should assign a smaller index to nodes with better “receiving capability.” (Note that it is even not necessary to set the destination to be Node  $M$ .) However, generally, receiving capability is not easily comparable (unless in the degraded case). Of course, we can always try all the permutations to maximize (9) or (12).

What makes the maximization problem even more complicated is the following possibility. We can arrange the nodes into groups, with each group consisting of one or more nodes. The information flow then is along a path formed out of the groups, but in any one group all the nodes have the same level (i.e., they decode the information at the same time). Put into mathematical terms, we have the following theorem.

**Theorem 3.4** *For a discrete memoryless multiple level relay channel with source Node 0, destination Node  $M$ , and the other nodes arranged into  $M - 1$  groups with each Group  $k$  consisting of  $n_k$  nodes,  $k = 1, \dots, M - 1$ , the following rate is achievable:*

$$R < \max_{p(\mathbf{X}_0, \mathbf{X}_1, \dots, \mathbf{X}_M)} \min_{1 \leq k \leq M} \min_{1 \leq i \leq n_k} I(\mathbf{X}_0, \dots, \mathbf{X}_{k-1}; Y_{k,i} | \mathbf{X}_k, \dots, \mathbf{X}_M), \quad (15)$$

where boldface characters are used to denote vectors for each group: e.g.,  $\mathbf{x}_k := (x_{k,1}, x_{k,2}, \dots, x_{k,n_k})$ ,  $k = 0, 1, \dots, M$ , with  $x_{k,i}$  denoting the input of the  $i$ -th node in Group  $k$ . Note that we have set Group 0 := {Node 0} and Group  $M$  := {Node  $M$ }.

For an application of this group relaying in the Gaussian channel case, we refer the reader to Theorem 3.12 in [6], where it is used to study wireless networks under low attenuation.

Finally, a remark on the allcast problem: Suppose the task is for the source Node 0 to send the same information to all the other Nodes  $1, 2, \dots, M$ . As we will see in the proof of Theorem 3.1 in the next section, this task is implicitly achieved by the following relaying scheme: The upstream nodes decode the information before the downstream nodes. Once the final Node  $M$  gets the information, all the other nodes have already obtained the information. Hence, the rate (9), (12) or (15) is also an achievable rate for the allcast problem.

## 4 Proof of Theorem 3.1

We will use the now standard “typical sequences” argument to prove achievability. First we summarize some basic properties of typical sequences that will be used later. For more details, see [9, Sections 8.6 and 14.2].

Let  $(Z_1, Z_2, \dots, Z_m)$  denote a finite collection of discrete random variables with some fixed joint distribution  $p(z_1, z_2, \dots, z_m)$  for  $(z_1, z_2, \dots, z_m) \in \mathcal{Z}_1 \times \mathcal{Z}_2 \times \dots \times \mathcal{Z}_m$ .

**Definition 4.1** *The set  $A_\epsilon^{(T)}$  of  $\epsilon$ -typical  $T$ -sequences  $(\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_m)$  is defined by*

$$A_\epsilon^{(T)}(Z_1, Z_2, \dots, Z_m) := \left\{ (\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_m) : \left| -\frac{1}{n} \log \text{Prob}(\mathbf{s}) - H(S) \right| < \epsilon, \quad \forall S \subseteq \{Z_1, Z_2, \dots, Z_m\} \right\},$$

where each  $\mathbf{z}_i = (z_{i,1}, z_{i,2}, \dots, z_{i,T})$  is a  $T$ -vector,  $i = 1, 2, \dots, m$ , and  $\mathbf{s}$  is defined as follows: If  $S = (Z_{i_1}, Z_{i_2}, \dots, Z_{i_\ell})$ , then  $\mathbf{s} = (\mathbf{z}_{i_1}, \mathbf{z}_{i_2}, \dots, \mathbf{z}_{i_\ell})$  and  $\text{Prob}(\mathbf{s}) = \text{Prob}(\mathbf{z}_{i_1}, \mathbf{z}_{i_2}, \dots, \mathbf{z}_{i_\ell}) = \prod_{t=1}^T p(z_{i_1,t}, z_{i_2,t}, \dots, z_{i_\ell,t})$ .

**Lemma 4.1** *For any  $\epsilon > 0$ , the following hold for sufficiently large  $T$ :*

(i) *Let a  $T$ -sequence  $(\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_m)$  be generated according to*

$$\prod_{t=1}^T p(z_{1,t}, z_{2,t}, \dots, z_{m,t}).$$

*Then*

$$\text{Prob}((\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_m) \in A_\epsilon^{(T)}(Z_1, Z_2, \dots, Z_m)) \geq 1 - \epsilon.$$

(ii) *Let a  $T$ -sequence  $(\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_m)$  be generated according to*

$$\prod_{t=1}^T p(z_{1,t} | z_{2,t}, \dots, z_{m-1,t}) p(z_{m,t} | z_{2,t}, \dots, z_{m-1,t}) p(z_{2,t}, \dots, z_{m-1,t}).$$

*Then*

$$\text{Prob}((\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_m) \in A_\epsilon^{(T)}(Z_1, Z_2, \dots, Z_m)) < 2^{-T(I(Z_1; Z_m | Z_2, \dots, Z_{m-1}) - 6\epsilon)}.$$

Lemma 4.1 (i) and (ii) follow immediately from Theorems 14.2.1 and 14.2.3 in [9], respectively.

The coding scheme used here is different from that of [3]. For example, the Slepian-Wolf partitioning is no longer used. This new coding scheme is easier to extend to the multiple relay case, and generally achieves larger rates (compared with [5]), although it coincides with [3] in the one relay case.

To make the proof easier to follow and allow a better comparison with the coding scheme in [3], first we give the proof for the one relay case ( $M = 2$ ), which contains all the essential ideas. Then we present the straightforward extension to the general case ( $M \geq 2$ ).

### 1. One relay case ( $M = 2$ )

We consider any fixed  $p(x_0, x_1)$ . Throughout the proof, we will use the following marginal or conditional probability mass functions obtained from  $p(x_0, x_1)$  and  $p(y_1, y_2; x_0, x_1)$ :

$$p(x_1), \quad p(x_0|x_1), \quad p(y_1|x_0, x_1), \quad p(y_2|x_0, x_1), \quad p(y_2|x_1),$$

where, for simplicity, the subscripts to distinguish these  $p$  functions are omitted since the exact meaning is obvious from the arguments.

We still use a block coding argument. We consider  $B$  blocks of transmission, each of  $T$  transmission slots. A sequence of  $B - 1$  indices,  $w(b) \in \{1, \dots, 2^{TR}\}$ ,  $b = 1, 2, \dots, B - 1$  will be sent over in  $TB$  transmission slots. (Note that as  $B \rightarrow \infty$ , the rate  $TR(B - 1)/TB$  is arbitrarily close to  $R$  for any  $T$ .)

### Generation of codebooks

The joint codebook is still generated in a backward manner. But here one significant difference is that all the codebooks are of the same length  $2^{TR}$ . (No more  $2^{TR_0}$  as in the proof of [3].)

- 1) Generate at random  $2^{TR}$  i.i.d.  $T$ -sequences in  $\mathcal{X}_1^T$ , each drawn according to  $\text{Prob}(\mathbf{x}_1) = \sum_{t=1}^T p(x_{1,t})$ . Index them as  $\mathbf{x}_1(w_1)$ ,  $w_1 \in \{1, 2, \dots, 2^{TR}\}$ . This is the random codebook for Node 1.
- 2) For each  $\mathbf{x}_1(w_1)$ , generate  $2^{TR}$  conditionally independent  $T$ -sequences  $\mathbf{x}_0(w_0|w_1)$ ,  $w_0 \in \{1, 2, \dots, 2^{TR}\}$ , drawn independently according to

$\text{Prob}(\mathbf{x}_0|\mathbf{x}_1(w_1)) = \sum_{t=1}^T p(x_{0,t}|x_{1,t}(w_1))$ . This defines the joint codebook for Nodes 0, 1:

$$\mathcal{C}_0 := \{\mathbf{x}_0(w_0|w_1), \mathbf{x}_1(w_1)\}. \quad (16)$$

It is apparent from [3] that the reason for this kind of *backward* codebook generation is that the upstream nodes (with smaller index) know what the downstream nodes are going to transmit, and therefore can adjust their own transmission accordingly. The converse is not true because of the unique direction of information flow.

Repeating the above process 1)-2) *independently* once more, we generate another random codebook  $\mathcal{C}_1$  similar to  $\mathcal{C}_0$  in (16). We will use these two codebooks alternately as follows: In block  $b = 1, \dots, B$ , the codebook  $\mathcal{C}_{(b \bmod 2)}$  is used. Hence, in any two consecutive blocks, codewords from different blocks are independent. This is a property we will use in the analysis of the probability of error (see (30)).

Before the transmission, the joint codebooks  $\mathcal{C}_0, \mathcal{C}_1$  are revealed to all the Nodes 0, 1, 2.

## Encoding

At the *beginning* of each block  $b \in \{1, \dots, B\}$ , Node 1 has an estimate (see the decoding section)  $\hat{w}_1(b-1)$  of  $w(b-1)$ , and sends the following  $T$ -sequence from the codebook  $\mathcal{C}_{(b \bmod 2)}$  in the block:

$$\vec{X}_1(b) := \mathbf{x}_1(\hat{w}_1(b-1)). \quad (17)$$

Also in the same block, Node 0 sends the following  $T$ -sequence from the same codebook  $\mathcal{C}_{(b \bmod 2)}$ :

$$\vec{X}_0(b) := \mathbf{x}_0(\hat{w}_0(b)|\hat{w}_0(b-1)), \quad (18)$$

where, obviously the estimate  $\hat{w}_0(b) = w(b)$  for every  $b \in \{1, \dots, B\}$ , since Node 0 is the source. Moreover, we set  $\hat{w}_i(b_1) = w(b_1) = 1$  for every  $b_1 \leq 0$ ,  $i \in \{0, 1, 2\}$ .

Every Node  $k \in \{1, 2\}$  thus receives a  $T$ -sequence

$$\vec{Y}_k(b) := \vec{Y}_k(\vec{X}_0(b), \vec{X}_1(b)) \quad (19)$$

with probability

$$\text{Prob}(\vec{Y}_k(b)|\vec{X}_0(b), \vec{X}_1(b)) = \sum_{t=1}^T p(\vec{Y}_{k,t}(b)|\vec{X}_{0,t}(b), \vec{X}_{1,t}(b)),$$

where  $\vec{Y}_{k,t}(b)$  is the  $t$ -th element of the vector  $\vec{Y}_k(b)$ , and similar definitions hold for  $\vec{X}_{i,t}(b)$ ,  $i = 0, 1$ .

### Decoding

At the *end* of each block  $b \in \{1, \dots, B\}$ , decodings at Node 1 and Node 2 happen simultaneously, but independently:

- Node 1 declares that  $\hat{w}_1(b) = w$  if  $w$  is the unique value in  $\{1, \dots, 2^{TR}\}$  such that in the block  $b$ ,

$$\{\mathbf{x}_0(w|\hat{w}_1(b-1)), \mathbf{x}_1(\hat{w}_1(b-1)), \vec{Y}_1(b)\} \in A_\epsilon^{(T)}(X_0, X_1, Y_1); \quad (20)$$

Otherwise, if no unique  $w$  as above exists, an error is declared with  $\hat{w}_1(b) = 0$ .

- Node 2 declares that  $\hat{w}_2(b-1) = w$  if  $w$  is the unique value in  $\{1, \dots, 2^{TR}\}$  such that in *both* the blocks  $b$  and  $b-1$ ,

$$\begin{aligned} \{\mathbf{x}_1(w), \vec{Y}_2(b)\} &\in A_\epsilon^{(T)}(X_1, Y_2), \quad \text{and} & (21a) \\ \{\mathbf{x}_0(w|\hat{w}_2(b-2)), \mathbf{x}_1(\hat{w}_2(b-2)), \vec{Y}_2(b-1)\} &\in A_\epsilon^{(T)}(X_0, X_1, Y_2). \end{aligned}$$

Otherwise, if no unique  $w$  as above exists, an error is declared with  $\hat{w}_2(b-1) = 0$ . The above scheme implies that the decoding at Node 2 has no intention to estimate  $w(b)$  at the end of block  $b$ .

### Analysis of probability of error

Denote the event that no decoding error is made in the first  $b$  blocks by

$$A_c(b) := \{\hat{w}_k(b_1 - k + 1) = w(b_1 - k + 1), \text{ for all } b_1 \in \{1, \dots, b\} \text{ and } k \in \{1, 2\}\},$$

and let its probability be  $P_c(b) := \text{Prob}(A_c(b))$ , with  $P_c(0) := 1$ .

Then the probability that some decoding error is made at some Node  $k \in \{1, 2\}$  in some block  $b \in \{1, \dots, B\}$  is

$$\begin{aligned}
P_e &:= \text{Prob}(\hat{w}_k(b-k+1) \neq w(b-k+1), \text{ for some } k \in \{1, 2\}, b \in \{1, \dots, B\}) \\
&= \sum_{b=1}^B \text{Prob}(\hat{w}_k(b-k+1) \neq w(b-k+1) \text{ for some } k \in \{1, 2\} | A_c(b-1)) \cdot P_c(b-1) \\
&\leq \sum_{b=1}^B \sum_{k=1}^2 \text{Prob}(\hat{w}_k(b-k+1) \neq w(b-k+1) | A_c(b-1)) \cdot P_c(b-1) \\
&=: \sum_{b=1}^B \sum_{k=1}^2 P_{e,k}(b) \cdot P_c(b-1), \tag{22}
\end{aligned}$$

where  $P_{e,k}(b) := \text{Prob}(\hat{w}_k(b-k+1) \neq w(b-k+1) | A_c(b-1))$ . Hence  $P_{e,k}(b)$  is the probability that a decoding error happens at Node  $k$  in block  $b$ , conditioned on the event that no decoding error was made in the previous  $b-1$  blocks.

Next, we calculate  $P_{e,k}(b)$ ,  $k \in \{1, 2\}$ . Since  $A_c(b-1)$  is presumed to hold, for every Node  $k \in \{1, 2\}$  we have

$$\hat{w}_k(b_1 - k + 1) = w(b_1 - k + 1), \text{ for } 1 \leq b_1 \leq b - 1.$$

Hence, by (17)-(19), the decoding rule (20) is equivalent to

$$\{\mathbf{x}_0(w|w(b-1)), \mathbf{x}_1(w(b-1)), \vec{Y}_1(\mathbf{x}_0(w(b)|w(b-1)), \mathbf{x}_1(w(b-1)))\} \in A_\epsilon^{(T)}(X_0, X_1, Y_1), \tag{23}$$

and the decoding rule (21a)-(21b) is equivalent to

$$\begin{aligned}
&\{\mathbf{x}_1(w), \vec{Y}_2(\mathbf{x}_0(w(b)|w(b-1)), \mathbf{x}_1(w(b-1)))\} \in A_\epsilon^{(T)}(X_1, Y_2), \quad \text{and (24a)} \\
&\{\mathbf{x}_0(w|w(b-2)), \mathbf{x}_1(w(b-2)), \vec{Y}_2(\mathbf{x}_0(w(b-1)|w(b-2)), \mathbf{x}_1(w(b-2)))\} \\
&\quad \in A_\epsilon^{(T)}(X_0, X_1, Y_2). \tag{24b}
\end{aligned}$$

Now, let

$$\begin{aligned}
\mathcal{W}_1(b) &:= \{w \in \{1, \dots, 2^{TR}\} : w \text{ satisfies (23)}\}; \\
\mathcal{W}_{2,0}(b) &:= \{w \in \{1, \dots, 2^{TR}\} : w \text{ satisfies (24a)}\}; \\
\mathcal{W}_{2,1}(b) &:= \{w \in \{1, \dots, 2^{TR}\} : w \text{ satisfies (24b)}\}; \\
\mathcal{W}_2(b) &:= \mathcal{W}_{2,0}(b) \cap \mathcal{W}_{2,1}(b).
\end{aligned}$$

Then,  $P_{e,k}(b)$  is the probability that  $w(b-k+1) \notin \mathcal{W}_k(b)$ , or some  $w' \in \mathcal{W}_k(b)$  but  $w' \neq w(b-k+1)$ , conditioned on the event that no decoding error was made in the previous  $b-1$  blocks. Thus,

$$\begin{aligned} P_{e,k}(b) &= \text{Prob}(w(b-k+1) \notin \mathcal{W}_k(b), \text{ or some } w' \in \mathcal{W}_k(b) \text{ but } w' \neq w(b-k+1) | A_c(b-1)) \\ &\leq \text{Prob}(w(b-k+1) \notin \mathcal{W}_k(b) | A_c(b-1)) \\ &\quad + \text{Prob}(\text{some } w' \in \mathcal{W}_k(b) \text{ but } w' \neq w(b-k+1) | A_c(b-1)) \\ &\leq \frac{\text{Prob}(w(b-k+1) \notin \mathcal{W}_k(b))}{P_c(b-1)} + \frac{\text{Prob}(\text{some } w' \in \mathcal{W}_k(b) \text{ but } w' \neq w(b-k+1))}{P_c(b-1)}. \end{aligned}$$

Hence, by (22),

$$\begin{aligned} P_e &\leq \sum_{b=1}^B \sum_{k=1}^2 [\text{Prob}(w(b-k+1) \notin \mathcal{W}_k(b)) \\ &\quad + \text{Prob}(\text{some } w' \in \mathcal{W}_k(b) \text{ but } w' \neq w(b-k+1))]. \end{aligned} \quad (25)$$

Now, for Node 1 with (23), for  $T$  sufficiently enough, applying Lemma 4.1 (i) with  $(Z_1, \dots, Z_m) = (X_0, X_1, Y_1)$ , we have

$$\text{Prob}(w(b) \notin \mathcal{W}_1(b)) < \epsilon, \quad (26)$$

and for any  $w' \neq w(b)$ , applying Lemma 4.1 (ii), we have

$$\text{Prob}(w' \in \mathcal{W}_1(b)) < 2^{-T(I(X_0; Y_1 | X_1) - 6\epsilon)}.$$

And also for Node 2 with (24a)-(24b), for  $T$  sufficiently enough, applying Lemma 4.1, we have for  $j = 0, 1$ ,

$$\text{Prob}(w(b-1) \notin \mathcal{W}_{2,j}(b)) < \epsilon,$$

and for any  $w' \neq w(b-1)$ ,

$$\text{Prob}(w' \in \mathcal{W}_{2,0}(b)) < 2^{-T(I(X_1; Y_2) - 6\epsilon)}, \quad (27)$$

$$\text{Prob}(w' \in \mathcal{W}_{2,1}(b)) < 2^{-T(I(X_0; Y_2 | X_1) - 6\epsilon)}. \quad (28)$$

Hence,

$$\text{Prob}(w(b-1) \notin \mathcal{W}_2(b)) \leq \sum_{j=0}^1 \text{Prob}(w(b-1) \notin \mathcal{W}_{2,j}(b)) \leq 2\epsilon, \quad (29)$$



and

$$\begin{aligned}
& \text{Prob}(\text{Some } w' \in \mathcal{W}_2(b) \text{ but } w' \neq w(b-1)) \\
\leq & \sum_{\substack{w' \in \{1, \dots, 2^{TR}\} \\ w' \neq w(b-1)}} \text{Prob}(w' \in \mathcal{W}_2(b)) \\
= & \sum_{\substack{w' \in \{1, \dots, 2^{TR}\} \\ w' \neq w(b-1)}} \prod_{j=0}^1 \text{Prob}(w' \in \mathcal{W}_{2,j}(b)) \tag{30} \\
\leq & (2^{TR} - 1)2^{-T(I(X_0, X_1; Y_2) - 12\epsilon)} \tag{31}
\end{aligned}$$

where (30) follows from the independence between the codebooks of any two consecutive blocks, and (31) follows from (27),(28) and the following equation:

$$I(X_0, X_1; Y_2) = I(X_1; Y_2) + I(X_0; Y_2|X_1).$$

For any  $R$  satisfying (9), by choosing  $\epsilon$  small enough, we can make  $T$  large enough such that for any  $\epsilon_1 > 0$ ,

$$\begin{cases} \text{Prob}(\text{some } w' \in \mathcal{W}_1(b) \text{ but } w' \neq w(b)) \leq (2^{TR} - 1)2^{-T(I(X_0; Y_1|X_1) - 6\epsilon)} < \epsilon_1, \\ \text{Prob}(\text{some } w' \in \mathcal{W}_2(b) \text{ but } w' \neq w(b-1)) \leq (2^{TR} - 1)2^{-T(I(X_0, X_1; Y_2) - 12\epsilon)} < \epsilon_1. \end{cases} \tag{32}$$

Hence, by (25),(26), (29) and (32),

$$\begin{aligned}
P_e & \leq \sum_{b=1}^B (3\epsilon + 2\epsilon_1) \\
& \leq 3B\epsilon + 2B\epsilon_1,
\end{aligned}$$

which can be made arbitrarily small by letting  $T \rightarrow \infty$ .

Finally, the argument on choosing one good codebook from many random codebooks and throwing away the worse half of its codewords is standard.

## 2. Multi-relay case. ( $M \geq 2$ )

We consider  $B$  blocks of transmission, each of  $T$  transmission slots. A sequence of  $B - M + 1$  indices,  $w(b) \in \{1, \dots, 2^{TR}\}$ ,  $b = 1, 2, \dots, B - M + 1$  will be sent over in  $TB$  transmission slots. (Note that as  $B \rightarrow \infty$ , the rate  $TR(B - M + 1)/TB$  is arbitrarily close to  $R$  for any  $T$ .)

## Generation of codebooks

- 1) Generate at random  $2^{TR}$  i.i.d.  $T$ -sequences in  $\mathcal{X}_{M-1}^T$ , each drawn according to  $\text{Prob}(\mathbf{x}_{M-1}) = \sum_{t=1}^T p(x_{M-1,t})$ . Index them as  $\mathbf{x}_{M-1}(w_{M-1})$ ,  $w_{M-1} \in \{1, 2, \dots, 2^{TR}\}$ . This is the random codebook for Node  $M-1$ .
- 2) For each  $\mathbf{x}_{M-1}(w_{M-1})$ , generate  $2^{TR}$  conditionally independent  $T$ -sequences  $\mathbf{x}_{M-2}(w_{M-2}|w_{M-1})$ ,  $w_{M-2} \in \{1, 2, \dots, 2^{TR}\}$ , drawn independently according to  $\text{Prob}(\mathbf{x}_{M-2}|\mathbf{x}_{M-1}(w_{M-1})) = \sum_{t=1}^T p(x_{M-2,t}|x_{M-1,t}(w_{M-1}))$ . This defines the joint codebook for Nodes  $M-2, M-1$ :  $\{\mathbf{x}_{M-2}(w_{M-2}|w_{M-1}), \mathbf{x}_{M-1}(w_{M-1})\}$ .
- 3) Continue the process 2) sequentially for Nodes  $i = M-3, M-4, \dots, 0$ , as follows: For each

$$\{\mathbf{x}_{i+1}(w_{i+1}|w_{i+2}, \dots, w_{M-1}), \mathbf{x}_{i+2}(w_{i+2}|w_{i+3}, \dots, w_{M-1}), \dots, \mathbf{x}_{M-1}(w_{M-1})\} \\ =: \mathbf{x}_{i+1}^{M-1}(w_{i+1}, \dots, w_{M-1}),$$

generate  $2^{TR}$  conditionally independent  $T$ -sequence  $\mathbf{x}_i(w_i|w_{i+1}, \dots, w_{M-1})$ , drawn independently according to

$$\text{Prob}(\mathbf{x}_i|\mathbf{x}_{i+1}^{M-1}(w_{i+1}, \dots, w_{M-1})) = \sum_{t=1}^T p(x_{i,t}|x_{i+1,t}^{M-1}(w_{i+1}, \dots, w_{M-1})).$$

Finally, we get a joint codebook for all the transmitter Nodes  $0, 1, \dots, M-1$  as

$$\mathcal{C}_0 = \{\mathbf{x}_0(w_0|w_1, \dots, w_{M-1}), \mathbf{x}_1(w_1|w_2, \dots, w_{M-1}), \dots, \mathbf{x}_{M-1}(w_{M-1})\},$$

with each  $w_i \in \{1, 2, \dots, 2^{TR}\}$ , for  $i = 0, 1, \dots, M-1$ .

Repeating the above process 1)-3) *independently*  $M-1$  times, we generate another  $M-1$  random codebooks  $\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_{M-1}$  similar to  $\mathcal{C}_0$ . We will use these  $M$  codebooks in a sequential way as follows: In block  $b = 1, \dots, B$ , the codebook  $\mathcal{C}_{(b \bmod M)}$  is used. Hence, in any  $M$  consecutive blocks, codewords from different blocks are independent. This is a property we will use in the analysis of the probability of error (see (41)).

Before the transmission, all the joint codebooks  $\mathcal{C}_0, \mathcal{C}_1, \dots, \mathcal{C}_{M-1}$  are revealed to all the nodes  $0, 1, \dots, M$ .

## Encoding

At the *beginning* of each block  $b \in \{1, \dots, B\}$ , every Node  $i \in \{0, \dots, M-1\}$  has estimates (see the decoding section)  $\hat{w}_i(b-k+1)$  of  $w(b-k+1)$ ,  $k \geq i+1$  (with  $\hat{w}_0(b_1) = w(b_1)$  for  $1 \leq b_1 \leq b$ ), and sends the following  $T$ -sequence from the codebook  $\mathcal{C}_{(b \bmod M)}$  in the block:

$$\vec{X}_i(b) := \mathbf{x}_i(\hat{w}_i(b-i)|\hat{w}_i(b-i-1), \dots, \hat{w}_i(b-M+1)), \quad (33)$$

where we set  $\hat{w}_i(b_1) = w(b_1) := 1$  for every  $b_1 \leq 0$ .

Every Node  $k \in \{1, 2, \dots, M\}$  thus receives a  $T$ -sequence

$$\vec{Y}_k(b) := \vec{Y}_k(\vec{X}_0(b), \dots, \vec{X}_{M-1}(b)) \quad (34)$$

with probability:

$$\text{Prob}(\vec{Y}_k(b)|\vec{X}_0(b), \dots, \vec{X}_{M-1}(b)) = \sum_{t=1}^T p(\vec{Y}_{k,t}(b)|\vec{X}_{0,t}(b), \dots, \vec{X}_{M-1,t}(b)),$$

where,  $\vec{X}_{k,t}(b)$  is the  $t$ -th element of the vector  $\vec{Y}_k(b)$ , and similar definitions hold for  $\vec{X}_{i,t}(b)$ ,  $i = 0, \dots, M-1$ .

## Decoding

At the *end* of each block  $b \in \{1, \dots, B\}$ , every Node  $k \in \{1, \dots, M\}$  (for  $b-k+1 \geq 1$ ) declares  $\hat{w}_k(b-k+1) = w$  if  $w$  is the unique value in  $\{1, \dots, 2^{TR}\}$  such that in *all* the blocks  $b-j$ ,  $j = 0, 1, \dots, k-1$ :

$$\{\mathbf{x}_{k-1-j}(w|\hat{w}_k(b-k), \dots, \hat{w}_k(b-j-M+1)), \dots, \mathbf{x}_{M-1}(\hat{w}_k(b-j-M+1)), \vec{Y}_k(b-j)\} \in A_c^{(T)}(X_{k-1-j}, \dots, X_{M-1}, Y_k); \quad (35)$$

Otherwise, if no unique  $w$  as above exists, an error is declared with  $\hat{w}_k(b-k+1) = 0$ .

## Analysis of probability of error

Denote the event that no decoding error is made in the first  $b$  blocks by

$$A_c(b) := \{\hat{w}_k(b_1 - k + 1) = w(b_1 - k + 1), \text{ for all } b_1 \in \{1, \dots, b\} \text{ and } k \in \{1, \dots, M\}\},$$

and let its probability be  $P_c(b) := \text{Prob}(A_c(b))$ , with  $P_c(0) := 1$ .

Then the probability that some decoding error is made at some Node  $k \in \{1, \dots, M\}$  in some block  $b \in \{1, \dots, B\}$  is

$$\begin{aligned}
P_e &:= \text{Prob}(\hat{w}_k(b-k+1) \neq w(b-k+1), \text{ for some } k \in \{1, \dots, M\}, b \in \{1, \dots, B\}) \\
&= \sum_{b=1}^B \text{Prob}(\hat{w}_k(b-k+1) \neq w(b-k+1) \text{ for some } k \in \{1, \dots, M\} | A_c(b-1)) \cdot P_c(b-1) \\
&\leq \sum_{b=1}^B \sum_{k=1}^M \text{Prob}(\hat{w}_k(b-k+1) \neq w(b-k+1) | A_c(b-1)) \cdot P_c(b-1) \\
&=: \sum_{b=1}^B \sum_{k=1}^M P_{e,k}(b) \cdot P_c(b-1), \tag{36}
\end{aligned}$$

where  $P_{e,k}(b) := \text{Prob}(\hat{w}_k(b-k+1) \neq w(b-k+1) | A_c(b-1))$ . Hence  $P_{e,k}(b)$  is the probability that a decoding error happens at Node  $k$  in block  $b$ , conditioned on the event that no decoding error was made in the previous  $b-1$  blocks.

Next, we calculate  $P_{e,k}(b)$ ,  $k \in \{1, \dots, M\}$ . Since  $A_c(b-1)$  is presumed to hold, for every Node  $k \in \{1, \dots, M\}$  we have

$$\hat{w}_k(b_1 - k + 1) = w(b_1 - k + 1), \text{ for } 1 \leq b_1 \leq b - 1.$$

Hence, by (33)-(34), the decoding rule (35) is equivalent to: Each Node  $k \in \{1, \dots, M\}$  (when  $b - k + 1 \geq 1$ ) declares  $\hat{w}_k(b - k + 1) = w$  if  $w$  is the unique value in  $\{1, \dots, 2^{2R}\}$  such that in all the blocks  $b - j$ , for  $j = 0, 1, \dots, k - 1$ :

$$\begin{aligned}
&\left\{ \mathbf{x}_{k-1-j}(w | w(b-k), \dots, w(b-j-M+1)), \right. \\
&\quad \mathbf{x}_{k-j}(w(b-k) | w(b-k-1), \dots, w(b-j-M+1)), \quad \dots, \quad \mathbf{x}_{M-1}(w(b-j-M+1)), \\
&\quad \left. \vec{Y}_k \left( \mathbf{x}_0(w(b-j) | w(b-j-1), \dots, w(b-j-M+1)), \dots, \mathbf{x}_{M-1}(w(b-j-M+1)) \right) \right\} \\
&\qquad \qquad \qquad \in A_\epsilon^{(T)}(X_{k-1-j}, X_{k-j}, \dots, X_{M-1}, Y_k). \tag{37}
\end{aligned}$$

Let

$$\begin{aligned}
\mathcal{W}_{k,j}(b) &:= \{w \in \{1, \dots, 2^{2R}\} : w \text{ satisfies (37)}\}; \\
\mathcal{W}_k(b) &:= \bigcap_{j=0}^{k-1} \mathcal{W}_{k,j}(b).
\end{aligned}$$

Then,  $P_{e,k}(b)$  is the probability that  $w(b-k+1) \notin \mathcal{W}_k(b)$ , or some  $w' \in \mathcal{W}_k(b)$  but  $w' \neq w(b-k+1)$ , conditioned on the event that no decoding error was made in the previous  $b-1$  blocks. Thus,

$$\begin{aligned}
P_{e,k}(b) &= \text{Prob}(w(b-k+1) \notin \mathcal{W}_k(b), \text{ or some } w' \in \mathcal{W}_k(b) \text{ but } w' \neq w(b-k+1) | A_c(b-1)) \\
&\leq \text{Prob}(w(b-k+1) \notin \mathcal{W}_k(b) | A_c(b-1)) \\
&\quad + \text{Prob}(\text{Some } w' \in \mathcal{W}_k(b) \text{ but } w' \neq w(b-k+1) | A_c(b-1)) \\
&\leq \frac{\text{Prob}(w(b-k+1) \notin \mathcal{W}_k(b))}{P_c(b-1)} + \frac{\text{Prob}(\text{some } w' \in \mathcal{W}_k(b) \text{ but } w' \neq w(b-k+1))}{P_c(b-1)}.
\end{aligned}$$

Hence, by (36),

$$\begin{aligned}
P_e &\leq \sum_{b=1}^B \sum_{k=1}^M \left[ \text{Prob}(w(b-k+1) \notin \mathcal{W}_k(b)) \right. \\
&\quad \left. + \text{Prob}(\text{some } w' \in \mathcal{W}_k(b) \text{ but } w' \neq w(b-k+1)) \right]. \tag{38}
\end{aligned}$$

Now, by Lemma 4.1 with  $(Z_1, \dots, Z_m) = (X_{k-1-j}, \dots, X_{M-1}, Y_k)$ , for  $T$  large enough, we have for  $j = 0, 1, \dots, k-1$ ,

$$\text{Prob}(w(b-k+1) \notin \mathcal{W}_{k,j}(b)) < \epsilon,$$

and for any  $w' \neq w(b-k+1)$ ,

$$\text{Prob}(w' \in \mathcal{W}_{k,j}(b)) < 2^{-T(I(X_{k-1-j}, Y_k | X_{k-j}, \dots, X_{M-1}) - 6\epsilon)}. \tag{39}$$

Hence,

$$\begin{aligned}
\text{Prob}(w(b-k+1) \notin \mathcal{W}_k(b)) &\leq \sum_{j=0}^{k-1} \text{Prob}(w(b-k+1) \notin \mathcal{W}_{k,j}(b)) \\
&\leq \sum_{j=0}^{k-1} \epsilon = k\epsilon \leq M\epsilon, \tag{40}
\end{aligned}$$

and

$$\begin{aligned}
& \text{Prob}(\text{some } w' \in \mathcal{W}_k(b) \text{ but } w' \neq w(b-k+1)) \\
& \leq \sum_{\substack{w' \in \{1, \dots, 2^{TR}\} \\ w' \neq w(b-k+1)}} \text{Prob}(w' \in \mathcal{W}_k(b)) \\
& = \sum_{\substack{w' \in \{1, \dots, 2^{TR}\} \\ w' \neq w(b-k+1)}} \prod_{j=0}^{k-1} \text{Prob}(w' \in \mathcal{W}_{k,j}(b)) \tag{41} \\
& \leq (2^{TR} - 1)2^{-T(I(X_0, \dots, X_{k-1}; Y_k | X_k, \dots, X_{M-1}) - 6k\epsilon)} \tag{42}
\end{aligned}$$

where the equation (41) follows from the independence among the codebooks of any  $M$  consecutive blocks, and (42) follows from (39) and the following equation:

$$I(X_0, \dots, X_{k-1}; Y_k | X_k, \dots, X_{M-1}) = \sum_{j=0}^{k-1} I(X_{k-1-j}, Y_k | X_{k-j}, \dots, X_{M-1}).$$

For any  $R$  satisfying (9), by choosing  $\epsilon$  small enough, we can make  $T$  large enough such that for any  $\epsilon_1 > 0$  and all  $k \in \{1, \dots, M\}$ ,

$$\begin{aligned}
& \text{Prob}(\text{some } w' \in \mathcal{W}_k(b) \text{ but } w' \neq w(b-k+1)) \\
& \leq (2^{TR} - 1)2^{-T(I(X_0, \dots, X_{k-1}; Y_k | X_k, \dots, X_{M-1}) - 6k\epsilon)} \\
& < \epsilon_1. \tag{43}
\end{aligned}$$

Hence, by (38), (40) and (43),

$$\begin{aligned}
P_e & \leq \sum_{b=1}^B \sum_{k=1}^M (M\epsilon + \epsilon_1) \\
& \leq BM^2\epsilon + BM\epsilon_1,
\end{aligned}$$

which can be made arbitrarily small by letting  $T \rightarrow \infty$ .

Finally, the argument on choosing one good codebook from many random codebooks and throwing away the worse half of the codewords is standard.

□

## 5 Proof of Theorem 3.2

The achievability is proved in Theorem 3.1. The converse follows immediately from the max-flow min-cut theorem for general multi-node networks stated in Theorem 14.10.1 in [9], where the node set  $S$  is chosen to be  $\{0\}$ ,  $\{0, 1\}$ ,  $\dots$ ,  $\{0, 1, \dots, M-1\}$  sequentially, and with the following equations:

$$\begin{aligned}
& I(X_0, \dots, X_{k-1}; Y_k, \dots, Y_M | X_k, \dots, X_{M-1}) \\
&= I(X_0, \dots, X_{k-1}; Y_k | X_k, \dots, X_{M-1}) \\
&\quad + \sum_{j=k+1}^M I(X_0, \dots, X_{k-1}; Y_j | Y_k, \dots, Y_{j-1}, X_k, \dots, X_{M-1}) \\
&= I(X_0, \dots, X_{k-1}; Y_k | X_k, \dots, X_{M-1}), \quad \text{for } k = 1, \dots, M, \quad (44)
\end{aligned}$$

where (44) follows from

$$\begin{aligned}
& \sum_{j=k+1}^M I(X_0, \dots, X_{k-1}; Y_j | Y_k, \dots, Y_{j-1}, X_k, \dots, X_{M-1}) \\
&= \sum_{j=k+1}^M H(Y_j | Y_k, \dots, Y_{j-1}, X_k, \dots, X_{M-1}) \\
&\quad - \sum_{j=k+1}^M H(Y_j | X_0, \dots, X_{k-1}, Y_k, \dots, Y_{j-1}, X_k, \dots, X_{M-1}) \\
&= H(Y_{k+1}, \dots, Y_M | Y_k, X_k, \dots, X_{M-1}) - H(Y_{k+1}, \dots, Y_M | Y_k, X_0, \dots, X_{M-1}) \\
&= 0, \quad \text{for } k = 1, \dots, M-1,
\end{aligned}$$

with the last equation following immediately from (8) and the definition of conditional entropy.  $\square$

## 6 Proof of Theorem 3.3

In this section, we prove Theorem 3.3.

We add a “virtual” Node  $M+1$  to the channel with output  $y_{M+1} \equiv y_M$ , but with no input. Hence this node will not affect the dynamics of the channel in any way. Then by Theorem 3.1, the following rate is achievable from Node 0 to Node  $M+1$ :

$$R < \max_{p(x_0, x_1, \dots, x_M)} \min_{1 \leq k \leq M+1} I(X_0, \dots, X_{k-1}; Y_k | X_k, \dots, X_M). \quad (45)$$

By the coding scheme stated in the proof of Theorem 3.1, the rate above is also achieved from Node 0 to Node  $M$ . (This is also obvious since  $y_{M+1} \equiv y_M$  and there is no  $x_{M+1}$ .)

Now by (45), to prove Theorem 3.1, we only need to show that the following inequality always holds:

$$I(X_0, \dots, X_M; Y_{M+1}) \geq I(X_0, \dots, X_{M-1}; Y_M | X_M).$$

Fortunately, by the construction that  $Y_{M+1} = Y_M$ , it follows immediately that

$$\begin{aligned} I(X_0, \dots, X_M; Y_{M+1}) &= I(X_0, \dots, X_M; Y_M) \\ &= I(X_M; Y_M) + I(X_0, \dots, X_{M-1}; Y_M | X_M) \\ &\geq I(X_0, \dots, X_{M-1}; Y_M | X_M). \end{aligned}$$

□

Finally, the proof of Theorem 3.4 is similar to that of Theorem 3.1 and is omitted.

## 7 Concluding remarks

We have provided a new coding scheme for the multiple level relay problem. This scheme is not only simpler, but it also easily extends to the case of multiple levels of relaying. This allows us to develop a simple expression for an achievable rate that is generally higher than that in [5]. For degraded channels, our result is exact.

## References

- [1] E. van der Meulen, “Three-terminal communication channels,” *Adv. Appl. Prob.*, vol. 3, pp. 120–154, 1971.
- [2] E. van der Meulen, *Transmission of information in a T-terminal discrete memoryless channel*. PhD thesis, Dept. of Statistics, University of California, Berkeley, 1968.
- [3] T. Cover and A. El Gamal, “Capacity theorems for the relay channel,” *IEEE Trans. Inform. Theory*, vol. 25, pp. 572–584, 1979.



- [4] B. Schein and R. Gallager, "The Gaussian parallel relay network," in *Proceedings of the 2000 IEEE Int. Symp. Inform. Theory*, (Sorrento, Italy), p. 22, 2000.
- [5] P. Gupta and P. R. Kumar, "Towards an information theory of large networks: an achievable rate region." Preprint, September 28, 2001. [http://black.csl.uiuc.edu/prkumar/ps\\_files/net\\_info.ps](http://black.csl.uiuc.edu/prkumar/ps_files/net_info.ps).
- [6] L.-L. Xie and P. R. Kumar, "A network information theory for wireless communication: scaling laws and optimal operation," *IEEE Trans. Inform. Theory*, 2002. Submitted.
- [7] G. Kramer, M. Gastpar, and P. Gupta, "Capacity theorems for wireless relay channels." Presented at 41st Allerton Conference on Communications, Control, and Computing, October 1–3, 2003. Monticello, Illinois.
- [8] A. Reznik, S. Kulkarni, and S. Verdú, "Capacity and optimal resource allocation in the degraded Gaussian relay channel with multiple relays." Manuscript, 2002.
- [9] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley and Sons, 1991.