On the Reliability Function of the Discrete Memoryless Relay Channel

Vincent Y. F. Tan, Member, IEEE

Abstract

Bounds on the reliability function for the discrete memoryless relay channel are derived using the method of types. Two achievable error exponents are derived based on partial decode-forward and compress-forward which are well-known superposition block-Markov coding schemes. The derivations require combinations of the techniques involved in the proofs of Csizszár-Körner-Marton’s packing lemma for the error exponent of channel coding and Marton’s type covering lemma for the error exponent of source coding with a fidelity criterion. The decode-forward error exponent is evaluated on Sato’s relay channel. From this example, it is noted that to obtain the fastest possible decay in the error probability for a fixed effective coding rate, one ought to optimize the number of blocks in the block-Markov coding scheme assuming the blocklength within each block is large. An upper bound on the reliability function is also derived using ideas from Haroutunian’s lower bound on the error probability for point-to-point channel coding with feedback.

Index Terms

Relay channel, Error exponents, Reliability function, Method of types, Block-Markov coding, Partial decode-forward, Compress-forward, Cutset bound, Haroutunian exponent

I. INTRODUCTION

We derive bounds on the reliability function for the discrete memoryless relay channel. This channel, introduced by van der Meulen in [1], is a point-to-point communication system that consists of a sender $X_1$, a receiver $Y_3$ and a relay with input $Y_2$ and output $X_2$. See Fig. 1. The capacity is not known in general but there exist several coding schemes that are optimal for certain classes of relay channels, e.g., physically degraded. These coding schemes, introduced in the seminal work by Cover and El Gamal [2] include decode-forward (DF), partial decode-forward (PDF) and compress-forward (CF). Using PDF, the capacity of the relay channel $C$ is lower bounded as

$$C \geq \max \min \{I(U; X_2; Y_3), I(U; Y_2; X_2) + I(X_1; Y_2; X_2|U)\}$$

where the maximization is over all $P_{UX_1X_2}$. The auxiliary random variable $U$ with cardinality $|U| \leq |X_1||X_2|$ represents a part of the message that the relay decodes; the rest of the message is decoded by the receiver. DF is a special case of PDF in which $U = X_1$ and instead of decoding part of the message as in PDF, the relay decodes the entire message. In CF, a more complicated coding scheme, the relay sends a description of $Y_2$ to the receiver. This description is denoted as $\hat{Y}_2$. The receiver then uses $Y_3$ as side information à la Wyner-Ziv [3, Ch. 11] [4] to reduce the rate of the description. One form of the CF lower bound is given as [5]

$$C \geq \max \min \{I(X_1; \hat{Y}_2Y_3|X_2), I(X_1X_2; Y_3) − I(Y_2; \hat{Y}_2X_1X_2Y_3)\}$$

where the maximization is over $P_{X_1}, P_{X_2}$ and $P_{Y_2|X_2}$ and $|\hat{Y}_2| \leq |X_2||Y_2| + 1$. Both PDF and CF involve block-Markov coding [2] in which the channel is used $N = nb$ times over $b$ blocks, each involving an independent message to be sent and the relay codeword in block $j$ depends statistically on the message from block $j − 1$. The best known upper bound on the capacity is the so-called cutset bound [2]

$$C \leq \max \min \{I(X_1X_2; Y_3), I(X_1; Y_2Y_3|X_2)\}$$

where the maximization is over all $P_{X_1X_2}$. 

The author is with the Department of Electrical and Computer Engineering (ECE) and the Department of Mathematics at the National University of Singapore (Email: vtan@nus.edu.sg).

This paper was presented in part at the 2013 International Symposium on Information Theory in Istanbul, Turkey.

The work of the author is supported in part by NUS startup grant R-263-000-A98-750/133 and in part by A*STAR, Singapore.
In addition to capacities, in information theory, error exponents are also of tremendous interest. They quantify the exponential rate of decay of the error probability when the rate of the code is below capacity or the set of rates is strictly within the capacity region. Such results allow us to provide approximate bounds on the blocklength needed to achieve a certain rate (or set of rates) and so provide a means to understand the tradeoff between rate(s) and error probability. In this paper, we derive achievable error exponents based on two superposition block-Markov coding schemes for the discrete memoryless relay channel. These bounds are positive for all rates below (1) and (2). We also derive an upper bound on the reliability function that is positive for all rates below the cutset bound in (3). We evaluate the exponent based on PDF for the Sato relay channel [6].

A. Main Contributions

We now elaborate on our three main contributions in this paper—all concerning bounds on the reliability function.

For PDF, of which DF is a special case, by using maximum mutual information (MMI) decoding [7], [8], we show that the analogue of the random coding error exponent (i.e., an error exponent that is similar in style to the one presented in [8 Thm. 10.2]) is universally attainable. That is, the decoder does not need to know the channel statistics. This is in contrast to the recent work by Bradford-Laneman [9] in which the authors employed maximum-likelihood (ML) decoding with the sliding window decoding technique introduced by Carleial [10] and also used by Kramer-Gastpar-Gupta [11] for relay networks. In [9], the channel needs to be known at the decoder but one advantage the sliding window has over backward decoding [12], [13] (which we use) is that it ameliorates the problem of excessive delay. To prove this result, we generalize the techniques used to prove the packing lemmas in [8], [14]–[17] so that they are applicable to the relay channel.

For CF, we draw inspiration from [18] in which the authors derived achievable error exponents for Wyner-Ahlswede-Körner coding (lossless source coding with coded side information) [3, Ch. 10] and Wyner-Ziv (lossy source coding with decoder side information) [3, Ch. 11] coding. We handle the combination of covering and packing in a similar way as [18] in order to derive an achievable error exponent for CF. In addition, a key technical contribution is the taking into account of the conditional correlation between \( \hat{Y}_2 \) and \( X_1 \) (given \( X_2 \)) using a bounding technique introduced by Scarlett-Martinez-Guillén i Fàbregas [19], [20] called, in our words, the one-at-a-time union bound. This bound is reviewed in Section II-E. We also leverage Csiszár’s \( \alpha \)-decoder [21] which specializes, in the point-to-point case, to ML decoding [22, Ch. 5] and MMI decoding [7].

For the upper bound on the reliability function, we draw on ideas from Haroutunian’s lower bound on the error probability for channel coding with feedback [23]. Our proof leverages on work by Palaiyanur [24]. We show that the upper bound can be expressed similarly to Haroutunian’s bound with the constraint that the minimization is over all transition matrices for which the cutset bound is no larger than the rate of transmission. This is the first time an upper bound on the reliability function for relay channels has been derived. At a very high level, we cast the relay channel as a point-to-point channel from \( (X_1, X_2) \) to \( (Y_2, Y_3) \) with feedback and make use of techniques developed by Haroutunian [23], [25] and Palaiyanur [24].

B. Related Work

The work that is most closely related to the current one are the papers by Bradford-Laneman [9] and Nguyen-Rasmussen [26] who derived random coding error exponents for DF based on Gallager’s Chernoff-bounding techniques [22]. The latter paper considers a streaming setting as well. We generalize their results to PDF and we use MMI decoding which has the advantage of being universal. Our techniques, in contrast to all previous
works on error exponents for relaying, hinge on the method of types which we find convenient in the discrete memoryless setting. The results are also intuitive and can be interpreted easily. For PDF, our work leverages on techniques used to prove various forms of the packing lemmas for multiuser channels in, for example, in the monograph by Haroutunian-Haroutunian-Harutyunyan [17]. It also uses a change-of-measure technique introduced by Hayashi [27] for proving second-order coding rates in channel coding. This change-of-measure technique allows us to use a random constant composition code ensemble and subsequently analyze this ensemble as if it were an i.i.d. ensemble without any loss in the error exponents sense.

For CF, since it is closely related to Wyner-Ziv coding [4], we leverage on the work of Kelly-Wagner [18] who derived an achievable exponents for Wyner-Ahlsvede-Körner coding [3] Ch. 10 and Wyner-Ziv [3] Ch. 11 coding. In a similar vein, Moulin-Wang [28] and Dasarathy-Draper [29] derived lower bounds for the error exponents of Gel’fand-Pinsker coding [3, Ch. 7] and content identification respectively. These works involve analyzing the combination of both packing and covering error events.

We also note that the authors in [30] and [31] presented achievable error exponents for various schemes for the additive white Gaussian noise (AWGN) relay channel and backhaul-constrained parallel relay networks respectively but these works do not take into account block-Markov coding [2]. Similarly, [32] analyzes the error exponents of Gel’fand- Pinsker coding [3, Ch. 7] and content identification respectively. These works involve analyzing the combination of both packing and covering error events.

C. Structure of Paper

This paper is structured as follows. In Section [II] we state our notation, some standard results from the method of types [8], [14] and the definitions of the discrete memoryless relay channel, and the reliability function. To make this paper as self-contained as possible, prior reliability function results for the point-to-point channel and for lossy source coding are also reviewed in this section. In Sections [III] and [IV] we state and prove error exponent theorems for PDF and CF respectively. In Section [V] we state an upper bound on the reliability function. In these three technical sections, the proofs of the theorems are provided in the final subsections (Subsections [III]-C, [IV]-D and [V]-C) and can be omitted at a first reading. We evaluate the DF exponent on the Sato relay channel in Section [VI]. Finally, we conclude our discussion in Section [VII] where we also mention several other avenues of research.

II. PRELIMINARIES

A. General Notation

We generally adopt the notation from Csiszár and Körner [8] with a few minor modifications. Random variables (e.g., X) and their realizations (e.g., x) are in capital and small letters respectively. All random variables take values on finite sets, denoted in calligraphic font (e.g., X). For a sequence \( x^n = (x_1, \ldots, x_n) \in X^n \), its type is the distribution \( P(x) = \frac{1}{n} \sum_{i=1}^{n} 1\{x = x_i\} \) where \( 1\{\text{clause}\} \) is 1 if the clause is true and 0 otherwise. All logs are with respect to base 2 and we use the notation \( \exp(t) \) to mean \( 2^t \). Finally, \( |a|^+ := \max\{a,0\} \) and \( |a| := \{1, \ldots, |a|\} \) for any \( a \in \mathbb{R} \).

The set of distributions supported on X is denoted as \( \mathcal{P}(X) \). The set of all sequences \( x^n \) of type P is the type class of P and is denoted as \( T_P := \{x^n \in X^n : x^n \text{ has type } P\} \). For a distribution \( P \in \mathcal{P}(X) \) and a stochastic matrix \( V : X \to Y \), we denote the joint distribution interchangeably as \( P \times V \) or \( PV \). This should be clear from the context. For \( x^n \in T_P \), the set of sequences \( y^n \in Y^n \) such that \( (x^n, y^n) \) has joint type \( P \times V \) is the V-shell \( T_V(x^n) \). Let \( \mathcal{V}_n(Y; P) \) be the family of stochastic matrices \( V : X \to Y \) for which the V-shell of a sequence of type \( P \in \mathcal{P}(X) \) is not empty. The elements of \( \mathcal{V}_n(Y; P) \) are called conditional types compatible with \( P \) (or simply conditional types if \( P \) is clear from the context).

Information-theoretic quantities are denoted in the usual way. For example, \( I(X; Y) \) and \( I(P, V) \) denote the mutual information where the latter expression makes clear that the joint distribution of \( (X, Y) \) is \( P \times V \). In addition, \( \hat{I}(x^n \land y^n) \) is the empirical mutual information of \( (x^n, y^n) \), i.e., if \( x^n \in T_P \) and \( y^n \in T_V(x^n) \), then, \( \hat{I}(x^n \land y^n) = I(P, V) \). For a distribution \( P \in \mathcal{P}(X) \) and two stochastic matrices \( V : X \to Y, W : X \times Y \to Z \), \( I(V, W | P) \) is the conditional mutual information \( I(Y; Z | X) \) where \( (X, Y, Z) \) is distributed as \( P \times V \times W \).
We will also often use the asymptotic notation $\approx$ to denote equality to first-order in the exponent. That is, for two positive sequences $\{a_n, b_n\}_{n=1}^{\infty}$, we say that $a_n \approx b_n$ if and only if $\lim_{n \to \infty} n^{-1} \log \frac{a_n}{b_n} = 0$. Also, we will use $\leq$ to denote inequality to first-order in the exponent. That is, $a_n \leq b_n$ if and only if $\limsup_{n \to \infty} n^{-1} \log \frac{b_n}{a_n} \leq 0$. Finally, $a_n = \Theta(b_n)$ if and only if there exists constants $0 < c_1 \leq c_2 < \infty$ such that $c_1 b_n \leq a_n \leq c_2 b_n$ for $n$ sufficiently large.

**B. The Method of Types**

We also summarize some known facts about types that we use extensively in the sequel. The following lemma summarizes key results in [8] Ch. 2.

**Lemma 1** (Basic Properties of Types). Fix a type $P \in \mathcal{P}_n(\mathcal{X})$ and sequence $x^n \in \mathcal{P}_n(\mathcal{X})$. Also fix a conditional type $V \in \mathcal{V}_n(\mathcal{Y}; P)$ and a sequence $y^n \in \mathcal{T}_V(x^n)$. For any stochastic matrix $W : \mathcal{X} \to \mathcal{Y}$, we have

1. $|\gamma_n(Y; P)| \leq (n+1)^{x|y|}$
2. $(n+1)^{-|x|y} \exp(nH(V|P)) \leq |\mathcal{T}_V(x^n)| \leq \exp(nH(V|P))$
3. $W^n(y^n|x^n) = \exp[-n(D(V||W|P) + H(V|P))]$
4. $(n+1)^{-|x|y} \exp[-nD(V||W|P)] \leq W^n(\mathcal{T}_V(x^n)|x^n) \leq \exp[-nD(V||W|P)]$

The following lemmas are implicit in the results in [8] Ch. 10. We provide formal statements and proofs in Appendices A and B for completeness. Lemma 2 is a conditional version of the following statement: Let $X^n$ be a length-$n$ sequence drawn uniformly at random from $\mathcal{T}_P$. The probability that an arbitrary sequence $\tilde{y}^n$ of marginal type $Q(y) = \sum_x P(x)V(y|x)$ (marginal consistency) lies in the $V'$-shell of $X^n$ is roughly $\exp[-nI(P,V')]$.

**Lemma 2** (Joint Typicality for Types I). Let $P \in \mathcal{P}_n(\mathcal{X}_1)$, $V \in \mathcal{V}_n(\mathcal{X}_2; P)$ and $V' \in \mathcal{V}_n(\mathcal{Y}; P \times V)$. Define $W(y|x_1) := \sum_{x_2} P(x_1)V(x_2|x_1)V'(y|x_1,x_2)$. Then for any $x^n_1 \in \mathcal{T}_P$, if $X^n_2$ is uniformly drawn from the shell $\mathcal{T}_V(x^n_1)$ and $\tilde{y}^n$ is any element of $\mathcal{T}_W(x^n_1)$,

$$
\frac{1}{p_1(n)} \exp[-nI(V,V'|P)] \leq \mathbb{P} [\tilde{y}^n \in \mathcal{T}_V(x^n_1,X^n_2)] \leq p_2(n) \exp[-nI(V,V'|P)],
$$

where $p_1(n)$ and $p_2(n)$ are polynomial functions of $n$ depending only on the cardinalities of the alphabets.

**Lemma 3** is a conditional version of the following statement: Let $\tilde{y}^n$ be a fixed length-$n$ sequence from $\mathcal{T}_{Q}$.

Let $X^n$ be a random length-$n$ sequence drawn uniformly at random from the marginal type class $\mathcal{T}_P$ where $P(x) = \sum_y Q(y)W(y|x)$ (marginal consistency). Then, the probability that $X^n$ lies in the $V'$-shell of $\tilde{y}^n$ is upper bounded by $\exp[-nI(P,W)]$ up to a polynomial term.

**Lemma 3** (Joint Typicality for Types II). Let $P \in \mathcal{P}_n(\mathcal{X}_1)$, $V \in \mathcal{V}_n(\mathcal{X}_2; P)$ and $V' \in \mathcal{V}_n(\mathcal{Y}; P)$. Let $W : \mathcal{Y} \times \mathcal{X}_1 \to \mathcal{X}_2$ be any (marginally consistent) channel satisfying $\sum_y W(x_2|y,x_1)V'(y|x_1) = V(x_2|x_1)$. Fix $x^n_1 \in \mathcal{T}_P$. Let $X^n_2$ be uniformly distributed in $\mathcal{T}_V(x^n_1)$. For any $y^n \in \mathcal{T}_V(x^n_1)$, we have

$$
\mathbb{P} [X^n_2 \in \mathcal{T}_W(y^n,x^n_1)] \leq p_3(n) \exp[-nI(V',W|P)],
$$

where $p_3(n)$ is a polynomial function of $n$ depending only on the cardinalities of the alphabets.

**C. The Relay Channel and Definition of Reliability Function**

In this section, we recall the definition of the relay channel and the notion of the reliability function for a channel.

**Definition 1.** A 3-node discrete memoryless relay channel (DM-RC) is a tuple $(\mathcal{X}_1 \times \mathcal{X}_2, W, \mathcal{Y}_2 \times \mathcal{Y}_3)$ where $\mathcal{X}_1, \mathcal{X}_2, \mathcal{Y}_2$ and $\mathcal{Y}_3$ are finite sets and $W : \mathcal{X}_1 \times \mathcal{X}_2 \to \mathcal{Y}_2 \times \mathcal{Y}_3$ is a stochastic matrix. The sender (node 1) wishes to communicate a message $M$ to the receiver (node 3) with the help of the relay node (node 2). See Fig. 7.

**Definition 2.** A $(2^{nR}, n)$-code for the DM-RC consists of a message set $\mathcal{M} = [2^{nR}]$, an encoder $f : \mathcal{M} \to \mathcal{X}_1^n$ that assigns a codeword to each message, a sequence of relay encoders $g_i : \mathcal{Y}_2^{i-1} \to \mathcal{X}_2, i \in [n]$ each assigning a symbol to each past received sequence and a decoder $\varphi : \mathcal{Y}_3^n \to \mathcal{M}$ that assigns an estimate of the message to each channel output. The rate of this code is $R$. 
We denote the $i$-th component of the vector $f(m) \in X^n_1$ as $f_i(m)$. We assume that $M$ is uniformly distributed on $\mathcal{M}$ and the channel is memoryless. More precisely, this means that the current received symbols $(Y_2, Y_3)$ are conditionally independent of the message and the past symbols $(M, X_1^{i-1}, X_2^{i-1}, Y_1^{i-1}, Y_3^{i-1})$ given the current transmitted symbols $(X_1, X_2)$, i.e.,

$$W^n(y_2^n, y_3^n | x_1^n, x_2^n) = \prod_{i=1}^n W(y_2, y_3 | x_1, x_2).$$ (6)

Fix a $(2^{nR}, n)$-code given by the message set $\mathcal{M} = [2^{nR}]$, encoder $f$, the relay encoders $(g_1, g_2, \ldots, g_n)$ and the decoder $\varphi$ yielding disjoint decoding regions $D_m = \varphi^{-1}(m) \subset \mathcal{Y}_3^n$. Let $g^n : \mathcal{Y}_2^n \to \mathcal{X}_2^n$ denote the concatenation of the relay encoders, i.e., $g^n(y_2^n) := (x_2^n, g_2(y_2), g_3(y_2^n), \ldots, g_n(y_2^n))$, where $x_2^n \in \mathcal{X}_2$ is any fixed symbol. (It does not matter which $x_2^n \in \mathcal{X}_2$ is fixed because $n$ is allowed to tend to infinity.) For a given DM-RC $W$ and coding functions $(f, g^n, \varphi)$, define $P_W((Y_2^n, Y_3^n) = (y_2^n, y_3^n) | M = m)$ to be the probability that $(Y_2^n, Y_3^n) = (y_2^n, y_3^n)$ when message $m$ is sent under channel $W$, i.e.,

$$P_W((Y_2^n, Y_3^n) = (y_2^n, y_3^n) | M = m) := \prod_{i=1}^n W(y_2, y_3 | f_i(m), g_i(y_2^{i-1})).$$ (7)

In addition, define the marginal probability

$$P_W(Y_3^n = y_3^n | M = m) := \sum_{y_2^n \in \mathcal{Y}_2^n} P_W((Y_2^n, Y_3^n) = (y_2^n, y_3^n) | M = m).$$ (8)

**Definition 3.** Let the average error probability given the code $(\mathcal{M}, f, g^n, \varphi)$ be defined as

$$P_e(W; \mathcal{M}, f, g^n, \varphi) := \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} P_W(Y_3^n \in D_m^c | M = m),$$ (9)

where $D_m^c := \mathcal{Y}_3^n \setminus D_m$. We also denote the average error probability more succinctly as $P_e(W)$ or $P(\hat{M} \neq M)$ where the dependencies on the code are suppressed.

We now define the reliability function formally.

**Definition 4.** The reliability function [8] for the DM-RC $W$ is defined as

$$E(R) := \sup \left\{ \liminf_{n \to \infty} \frac{1}{n} \log \frac{1}{P(\hat{M} \neq M)} : M \in \mathcal{M} = [2^{nR}] \right\}$$ (10)

where $M \in \mathcal{M} = [2^{nR}]$ and the supremum is taken over all sequences of $(2^{nR}, n)$ codes for the DM-RC.

As in [9], we use block-Markov coding to send a message $M$ representing $NR_{\text{eff}} = nbR_{\text{eff}}$ bits of information over the DM-RC. We use the channel $N$ times and this total blocklength is partitioned into $b$ correlated blocks each of length $n$. We term $n$, a large integer, as the *per-block blocklength*. The number of blocks $b$ is fixed and regarded as a constant (does not grow with $n$). We discuss the effect of $b = b_n$ growing in Section 21-B. The message is split into $b - 1$ sub-messages $M_j, j \in [b - 1]$, each representing $nR$ bits of information. Thus, the effective rate of the code is

$$R_{\text{eff}} = \frac{(b - 1)nR}{N} = \frac{(b - 1)nR}{nb} = \frac{b - 1}{b}R.$$ (11)

We also say that $R > R_{\text{eff}}$ is the *per-block rate*. Under this coding setup, we wish to provide lower bounds on the reliability function. We also prove an upper bound on the reliability function.

**D. Background on Error Exponents via the Method of Types**

In this section, we provide a brief summary of Csiszár-Körner-style [8] error exponents for channel coding [16] and lossy source coding [35]. For a more comprehensive exposition, see Csiszár’s review paper [14].

For channel coding, the packing lemma [8, Lem. 10.1] allows us to show that for every distribution $P \in \mathcal{P}(\mathcal{X})$, the following exponent, called the *random coding error exponent*, is achievable

$$E(R, P) := \min_{V: \mathcal{X} \to \mathcal{Y}} D(V || W | P) + |I(P, V) - R|^+. $$ (12)
Roughly speaking, the term $D(V\|W|P)$ represents the atypicality of the channel $W : \mathcal{X} \rightarrow \mathcal{Y}$, namely that it behaves like $V$. (More precisely, for finite $n$, the conditional type of $y^n$ given $x^n$ is represented by $V$.) The term $|I(P, V) - R|^+$ represents the deviation of the code rate from the rate that the channel can support $I(P, V)$. Besides the proof using the packing lemma [8, Thm. 10.2], one can show that $E_\alpha(R, P)$ is achievable by Gallager’s Chernoff bounding techniques [22, Ch. 5] or by considering a family of decoding rules which involve maximizing a function $\alpha(P, V)$ [21]. Here, $P$ represents the type of the codewords of a constant composition code and $V$ represents the conditional type of the channel output given a particular codeword. If $\alpha(P, V) = I(P, V)$ this corresponds to MMI decoding; if $\alpha(P, V) = H(V|P)$, this corresponds to minimum conditional entropy decoding (these two decoding strategies are identical for constant composition codes); while if $\alpha(P, V) = D(V\|W|P) + H(V|P)$ this corresponds to ML decoding. Notice that ML decoding depends on knowledge of the true channel $W$. For PDF, we will use MMI decoding and obtain an exponent that is similar to (12). For CF, we find it convenient to use a combination of MMI and ML decoding in addition to other techniques such as the one-at-a-time union bound described in Section IV.

It was shown by Haroutunian [25] that for every distribution $P \in \mathcal{P}(\mathcal{X})$, the following is an upper bound to the reliability function

$$E_{sp}(R, P) := \min_{V: X \rightarrow Y \colon I(P, V) \leq R} D(V\|W|P). \quad (13)$$

This is customarily called the sphere-packing exponent. An alternative expression is given by Shannon-Gallager-Berlekamp [36]. In the presence of feedback, Haroutunian [23] also proved the following upper bound to the reliability function

$$E_H(R) := \min_{V: X \rightarrow Y \colon C(V) \leq R} \max_P D(V\|W|P) \quad (14)$$

where $C(V) := \max_P I(P, V)$ is the Shannon capacity of $V$. This is called the Haroutunian exponent. It is known that $E_H(R) = \max_P E_{sp}(R, P)$ for output symmetric channels but in general $E_H(R) > \max_P E_{sp}(R, P)$ [23].

Our upper bound on the reliability function is similar to the Haroutunian exponent with the exception that $C(V)$ in (14) is replaced by the cutset bound $C_e(V) := \max_P \min \{I(X_1X_2; Y_3), I(X_1; Y_2Y_3|X_2)\}$ where $V$ is the conditional distribution of $(Y_2, Y_3)$ given $(X_1, X_2)$.

In our proof of the CF exponent in Section IV, we will make use of a technique Marton [35] developed to analyze the error exponent for compressing discrete memoryless sources with a fidelity criterion. In contrast to channel coding, this exponent is known for all rates and is given by the Marton exponent

$$F(P, R, \Delta) = \inf_{Q: R(Q, \Delta) > R} D(Q\|P) \quad (15)$$

where $P \in \mathcal{P}(\mathcal{X})$ is the distribution of the source, $R$ is code rate, $\Delta$ is the distortion level and $R(Q, \Delta)$ is the rate-distortion function. Marton’s exponent in (15) is intuitive in view of Sanov’s theorem [8, Ch. 2]: If the coding rate $R$ is below $R(Q, \Delta)$ for some $Q$, which represents the type of the realization of the source $X^n \sim P^n$, then the error decays with exponent $D(Q\|P)$. A key lemma used to prove the direct part of (15) is called the type covering lemma proved by Berger [37]. Also see [8, Lem. 9.1]. We use a refined version of this technique and combine that with other packing techniques to prove the CF exponent in Section IV.

E. The “One-At-A-Time Union Bound”

As mentioned in the Introduction, we use a modification of the union bound in our proof of the error exponent for CF. This is called (in our words) the one-at-a-time union bound and was used for error exponent analyses in [19], [20]. To describe this technique, first recall the truncated union bound which says that if $E_m, 1 \leq m \leq M$ is a collection of events with the same probability, then

$$\mathbb{P}\left(\bigcup_{m=1}^M E_m\right) \leq \min \{1, M\mathbb{P}(E_1)\}. \quad (16)$$

The one-at-a-time union bound says that if we have two independent sequences of identically distributed random variables $A_m, 1 \leq m \leq M$ and $B_k, 1 \leq k \leq K$, then the probability that any of the pairs $(A_m, B_k)$ belongs to...
some set $\mathcal{D}$ can be bounded as

$$\mathbb{P}\left(\bigcup_{m=1}^{M} \bigcup_{k=1}^{K} \{(A_m, B_k) \in \mathcal{D}\}\right) \leq \min\left\{1, M \mathbb{P}\left(\bigcup_{k=1}^{K} \{(A_1, B_k) \in \mathcal{D}\}\right)\right\} \leq \min\left\{1, M \mathbb{E}\left[\mathbb{P}\left(\bigcup_{k=1}^{K} \{(A_1, B_k) \in \mathcal{D}\} \mid A_1\right)\right]\right\} \leq \min\left\{1, M \mathbb{E}\left[\min\{1, K \mathbb{P}\left(\{(A_1, B_1) \in \mathcal{D}\} \mid A_1\}\right)\right]\right\}. \quad (17)$$

In (17) and (19), we applied the truncated union bound in (16). Clearly, applying the union bounds in the other order ($k$ first, then $m$) yields another upper bound. These bounds are usually better than a simple application of the union bound on both indices jointly, i.e.,

$$\mathbb{P}\left(\bigcup_{m,k} \{(A_m, B_k) \in \mathcal{D}\}\right) \leq \min\{1, MK \mathbb{P}\left(\{(A_1, B_1) \in \mathcal{D}\}\right)\}. \quad (20)$$

## III. PARTIAL DECODE-FORWARD (PDF)

We warm up by deriving two achievable error exponents using PDF. In PDF, the relay decodes part of the message in each block. For block $j \in [b]$, the part of the message that is decoded by the relay is indicated as $M_j'$ and the remainder of the message is $M_j''$. Thus, $M_j = (M_j', M_j'')$. We will state the main theorem in Section III-A, provide some remarks in Section III-B and prove it in Section III-C.

### A. Analogue of the Random Coding Error Exponent

The analogue of the random coding exponent is presented as follows:

**Theorem 1** (Random Coding Error Exponent for Partial Decode-Forward). Fix $b \in \mathbb{N}$, auxiliary alphabet $\mathcal{U}$ and a rate $R$. Let $R = R' + R''$ for two non-negative rates $R'$ and $R''$. Fix a joint distribution $Q_{X_2} \times Q_{U|X_2} \times Q_{X_1|U X_2} \in \mathcal{P}(\mathcal{X}_2 \times \mathcal{U} \times \mathcal{X}_1)$. These distributions induce the following virtual channels:

$$W_{Y_2|U X_2}(y_2|u, x_2) := \sum_{x_1, y_3} W(y_2, y_3|x_1, x_2)Q_{X_1|U X_2}(x_1|u, x_2), \quad (21)$$

$$W_{Y_3|U X_2}(y_3|u, x_2) := \sum_{x_1, y_2} W(y_2, y_3|x_1, x_2)Q_{X_1|U X_2}(x_1|u, x_2), \quad (22)$$

$$W_{Y_3|U X_1 X_2}(y_3|u, x_1, x_2) := \sum_{y_2} W(y_2, y_3|x_1, x_2), \quad \forall u \in \mathcal{U}. \quad (23)$$

The following is a lower bound on the reliability function:

$$E(R_{\text{eff}}) \geq E_{\text{pdf}}^{(b)}(R_{\text{eff}}) := \frac{1}{b} \min\{F(R'), G(R'), \tilde{G}(R'')\}, \quad (24)$$

where $F(R'), G(R')$ and $\tilde{G}(R'')$ are constituent error exponents defined as

$$F(R') := \min_{V: \mathcal{U} \times \mathcal{X}_2 \rightarrow \mathcal{Y}_2} D(V||W_{Y_2|U X_2}Q_{U X_2}) + |I(Q_{U|X_2}, V|Q_{X_2}) - R'|^+ \quad (25)$$

$$G(R') := \min_{V: \mathcal{U} \times \mathcal{X}_2 \rightarrow \mathcal{Y}_3} D(V||W_{Y_3|U X_2}Q_{U X_2}) + |I(Q_{U|X_2}, V) - R'|^+ \quad (26)$$

$$\tilde{G}(R'') := \min_{V: \mathcal{U} \times \mathcal{X}_1 \rightarrow \mathcal{Y}_3} D(V||W_{Y_3|U X_1 X_2}Q_{U X_1 X_2}) + |I(Q_{X_1|U X_2}, V|Q_{U X_2}) - R''|^+ \quad (27)$$

The proof of this result is based on a modification of the techniques used in the packing lemma [8], [14]–[17] and is provided in Section III-C.
B. Remarks on the Error Exponent for Partial Decode-Forward

A few comments are in order concerning Theorem 1.

1) Firstly, since $Q_{X_2}$, $Q_{Y|X_2}$ and $Q_{X_1|UX_2}$ as well as the splitting of $R$ into $R'$ and $R''$ are arbitrary, we can maximize the lower bounds in (24) over these free parameters. Particularly, for a fixed split $R' + R'' = R$, if

$$R' < I(U; Y_2|X_2)$$
$$R' < I(UX_2; Y_3)$$
$$R'' < I(X_1; Y_3|UX_2),$$

for some $Q_{X_2}$, $Q_{Y|X_2}$ and $Q_{X_1|UX_2}$, then $F(R')$, $G(R')$, and $\hat{G}(R'')$ are positive. Hence, the error probability decays exponentially fast if $R$ satisfies the PDF lower bound in (1). In fact, $|U|$ is also a free parameter. As such we may let $|U| \to \infty$. It is not obvious that a finite $U$ is optimal (in the sense that $E_{\text{pdf}}^b(R_{\text{eff}})$ does not improve by increasing $|U|$). This is because we cannot apply the usual cardinality bounding techniques based on the support lemma [8, App. C]. A method to prove that the cardinalities of auxiliary random variables can be bounded for error exponents was presented in [38, Thm. 2] but the technique is specific to the multiple-access channel and does not appear to carry over to our setting in a straightforward manner.

2) Secondly, we can interpret $F(R')$ as the error exponent for decoding a part of the message $M_j$ (block $j$) at the relay and $G(R')$ and $\hat{G}(R'')$ as the error exponents for decoding the whole of the message $M_j = (M'_j, M''_j)$ at the receiver. Setting $U = X_1$ and $R'' = 0$ recovers DF for which the error exponent (without the sliding-window modification) is provided in [9]. Indeed, now we only have two exponents $F(R)$ and $G(R)$ corresponding respectively to the errors at the relay and the decoder. For the overall exponent to be positive, we require

$$R < \min\{I(X_1; Y_2|X_2), I(X_1X_2; Y_3)\}$$

which corresponds exactly to the DF lower bound [3, Thm. 16.2]. However, the form of the exponents is different from that in [9]. Ours is in the Csiszár-Körner [8] style while [9] presents exponents in the Gallager [22] form. Also see item 7 below.

3) The exponent in (24) demonstrates a tradeoff between the effective rate and error probability: for a fixed $R$, as the number of blocks $b$ increases, $R_{\text{eff}}$ increases but because of the division by $b$, $E_{\text{pdf}}^b(R_{\text{eff}})$ decreases. Varying $R$ alone, of course, also allows us to observe this tradeoff. Now in capacity analysis, one usually takes the number of blocks $b$ to tend to infinity so as to be arbitrary close to a $*$-forward lower bound [4, Sec. 16.4.1]. If we let $b$ increase with the per-block blocklength $n$, then the decay in the error probability would no longer be exponential because we divide by $b$ in (24) so $E_{\text{pdf}}^b(R_{\text{eff}})$ is arbitrarily small for large enough $n$. For example if $b_n = \Theta(n^{1/2})$, then the error probability decays as $\exp(-\Theta(nb_n^{-1})) = \exp(-\Theta(n^{1/2}))$. However, the effective rate $R_{\text{eff}}$ would come arbitrarily close to the per-block rate $R$. If $R$ is also allowed to tend towards a $*$-forward lower bound, this would be likened to operating in the moderate- instead of large-deviations regime as the rate of decay of $P_e(W)$ is subexponential but the effective rate of the code is arbitrarily close to a $*$-forward lower bound. See [39], [40] for a partial list of works on moderate-deviations in information theory.

4) In general, the sliding window technique [10], [11] may yield potentially better exponents. We do not explore this extension here but note that the improvements can be obtained by appealing to the techniques in [9, Props. 1 and 2]. In the sliding window decoding technique, the receiver estimates a message such that two typicality conditions are simultaneously satisfied. See [3, Sec. 18.2.1, pp. 463]. This corresponds, in our setting, to maximizing two separate empirical mutual information quantities simultaneously, which is difficult to analyze.

5) We also derived expurgated error exponents for PDF using the technique outlined in [8, Ex. 10.2 and 10.18] together with ML decoding. These appear similarly to their classical forms and so are not included in this paper.

6) In the proof, we use a random coding argument and show that averaged over a random code ensemble, the probability of error is desirably small. We do not first assert the existence of a good codebook via the classical packing lemma [8, Lem. 10.1] (or its variants [14]–[17]) then upper bound the error probability given the non-random codebook. We also use a change-of-measure technique that was also used by Hayashi [27] for
proving second-order coding rates in channel coding. This change-of-measure technique allows us to use a random constant composition code ensemble and subsequently analyze this ensemble as if it were an i.i.d. ensemble, thus simplifying the analysis.

7) As is well known [8 Ex. 10.24], we may lower bound the exponents in (25–27) in the Gallager form [22, 41], which is more amenable to computation. Furthermore, the Gallager-form lower bound is tight for capacity-achieving input distributions in the point-to-point setting. For the DM-RC using the PDF scheme, for a fixed joint distribution \( Q_{X_2} \times Q_{U|X_2} \times Q_{Y_1|X_2} \in \mathcal{P}(X_2 \times U \times X_1) \),

\[
F(R') \geq \max_{\rho \in [0, 1]} \left\{ -\rho R' - \log \sum_{x_2, y_2} Q_{X_2}(x_2) \left[ \sum_u Q_{U|X_2}(u|x_2) W_{Y_2|U,X_2}(y_2|u, x_2)^{1+\rho} \right]^{1+\rho} \right\},
\]

(32)

\[
G(R') \geq \max_{\rho \in [0, 1]} \left\{ -\rho R' - \log \sum_{u, x_2} Q_{UX_2}(u, x_2) \left[ \sum_{y_2} Q_{Y_2|UX_2}(y_2|u, x_2)^{1+\rho} \right]^{1+\rho} \right\},
\]

(33)

\[
\bar{G}(R'') \geq \max_{\rho \in [0, 1]} \left\{ -\rho R'' - \log \sum_{u, x_2, y_3} Q_{UX_2}(u, x_2) \left[ \sum_{x_1} Q_{X_1|UX_2}(x_1|u, x_2) W_{Y_3|UX_1,X_2}(y_3|u, x_1, x_2)^{1+\rho} \right]^{1+\rho} \right\}.
\]

(34)

We compute these exponents for the Sato relay channel [6] in Section VI.

C. Proof of Theorem 7

Proof: We code over \( b \) blocks each of length \( n \) (block-Markov coding [22]). Fix rates \( R' \) and \( R'' \) satisfying \( R' + R'' = R \). We fix a joint type \( Q_{X_2} \times Q_{U|X_2} \times Q_{Y_1|X_2} \in \mathcal{P}(X_2 \times U \times X_1) \). Split each message \( M_j, j \in [b-1] \) of rate \( R \) into two independent parts \( M_j' \) and \( M_j'' \) of rates \( R' \) and \( R'' \) respectively.

Generate \( k' = \exp(nR') \) sequences \( x_2^n(m_{j-1}', m_{j-1}'') \in [k'] \) uniformly at random from the type class \( T_{Q_{X_2}} \). For each \( m_{j-1}' \in [k'] \) (with \( m_0' = 1 \)), generate \( k'' = \exp(nR'') \) sequences \( u^n(m_j', m_{j-1}'') \) uniformly at random from the \( Q_{U|X_2} \)-shell \( \mathcal{T}_{Q_{U|X_2}}(x_2^n(m_{j-1}'')) \). Now for each \( (m_{j-1}'', m_j') \in [k''] \), generate \( k'' = \exp(nR'') \) sequences \( x_2^n(m_j', m_{j-1}'') \) uniformly at random from the \( Q_{X_1|UX_2} \)-shell \( \mathcal{T}_{Q_{X_1|UX_2}}(u^n(m_j', m_{j-1}''), x_2^n(m_{j-1}'')) \). This gives a random codebook. Note that the \( x_2^n(m_{j-1}'') \) sequences need not be distinct if \( R' \geq H(Q_{X_2}) \) (and similarly for the \( u^n(m_j', m_{j-1}'') \) and \( x_2^n(m_j', m_{j-1}'') \) sequences) but this does not affect the subsequent arguments. We bound the error probability averaged over realizations of this random codebook.

The sender and relay cooperate to send \( m_j' \) to the receiver. In block \( j \), relay transmits \( x_2^n(m_{j-1}'') \) and transmits \( x_1^n(m_j', m_{j-1}'') \) (with \( m_0' = m_0'' = 1 \) by convention).

At the \( j \)-th step, the relay does MMI decoding [7] for \( m_j' \) given \( m_{j-1}' \) and \( y_2^n(j) \). More precisely, it declares that \( \hat{m}_j' \) is sent if

\[
\hat{m}_j' = \arg \max_{m_j' \in [\exp(nR')]} \hat{I}(u^n(m_j'|m_{j-1}'') \land y_2^n(j)|x_2^n(m_{j-1}'')).
\]

By convention, set \( m_0' = 1 \). Recall that \( \hat{I}(u^n(m_j'|m_{j-1}'') \land y_2^n(j)|x_2^n(m_{j-1}'')) \) is the conditional mutual information \( I(\hat{U}; \hat{Y}_2|\hat{X}_2) \) where the dummy random variables \( (\hat{X}_2, \hat{U}, \hat{Y}_2) \) have joint type given by the sequences \( (u^n(m_j'|m_{j-1}''), x_2^n(m_{j-1}''), y_2^n(j)) \). After all blocks are received, the decoder performs backward decoding [12, 13] by using the MMI decoder [7]. In particular, it declares that \( \hat{m}_j' \) is sent if

\[
\hat{m}_j' = \arg \max_{m_j' \in [\exp(nR')]} \hat{I}(u^n(m_j'|m_{j-1}''), x_2^n(m_j'') \land y_3^n(j+1)).
\]

(36)

After all \( m_j', j \in [b-1] \) have been decoded in step (36), the decoder then decodes \( M_j'' \) using MMI [7] as follows:

\[
\hat{m}_j'' = \arg \max_{m_j'' \in [\exp(nR'')]} \hat{I}(x_1^n(m_j', m_{j-1}'') \land y_2^n(j)|u^n(m_j'|\hat{m}_j'-1), x_2^n(\hat{m}_j'-1)).
\]

(37)

In steps (35), (36) and (37) if there exists more than one message attaining the arg max, then pick any one uniformly at random. We assume, by symmetry, that \( M_j = (M_j', M_j'') = (1, 1) \) is sent for all \( j \in [b-1] \). The line of analysis in [9 Sec. III] yields

\[
P(\hat{M} \neq M) \leq (b-1)(\epsilon_R + \epsilon_{D,1} + \epsilon_{D,2})
\]

(38)
where for any $j \in [b - 1]$,
\[
\epsilon_R := \mathbb{P}(\tilde{M}_j' \neq 1 | \tilde{M}_{j-1}' = 1)
\]
(39)
is the error probability in decoding at the relay and for any $j \in [b - 1]$,
\[
\epsilon_{D,1} := \mathbb{P}(\tilde{M}_{j+1}' \neq 1 | \tilde{M}_{j+1}' = 1, M_{j+1} = 1), \quad \text{and}
\]
\[
\epsilon_{D,2} := \mathbb{P}(\tilde{M}_{j}' \neq 1 | M_{j}' = 1, M_{j-1}' = 1)
\]
(40)
(41)
and are the error probabilities of decoding $M_{j+1}$ and $M_{j}'$ at the decoder. Since $b$ is assumed to be constant, it does not affect the exponential dependence of the error probability in (38). So we just bound $\epsilon_R$, $\epsilon_{D,1}$ and $\epsilon_{D,2}$. Since all the calculations are similar, we focus on $\epsilon_R$ leading to the error exponent $F(R')$ in (25).

An error occurs in step (35) (conditioned on neighboring blocks being decoded correctly so $\tilde{m}_{j-1}' = m_{j-1}' = 1$) if and only if there exists some index $\tilde{m}_j' \neq 1$ such that the empirical conditional information computed with respect to $u^n(\tilde{m}_j'|1)$ is higher than that of $u^n(1|1)$, i.e.,
\[
\epsilon_R = \mathbb{P}(\exists \tilde{m}_j' \neq 1 : \hat{I}(U^n(\tilde{m}_j'|1) \wedge Y^n_2(j)|X^n_2(1)) \geq \hat{I}(U^n(1|1) \wedge Y^n_2(j)|X^n_2(1))).
\]
(42)
We can condition this on various values of $(u^n, x^n_2) \in \mathcal{T}_{Q_{UX_2}}$ as follows,
\[
\epsilon_R = \sum_{(u^n, x^n_2) \in \mathcal{T}_{Q_{UX_2}}} \frac{1}{|\mathcal{T}_{Q_{UX_2}}|} \beta_n(u^n, x^n_2),
\]
(43)
where
\[
\beta_n(u^n, x^n_2) := \mathbb{P}(\exists \tilde{m}_j' \neq 1 : \hat{I}(U^n(\tilde{m}_j'|1) \wedge Y^n_2(j)|X^n_2(1)) \\
\geq \hat{I}(U^n(1|1) \wedge Y^n_2(j)|X^n_2(1)) | (U^n(1|1), X^n_2(1)) = (u^n, x^n_2)).
\]
(44)
It can be seen that $\beta_n(u^n, x^n_2)$ does not depend on $(u^n, x^n_2) \in \mathcal{T}_{Q_{UX_2}}$ so we simply write $\beta_n = \beta_n(u^n, x^n_2)$ and we only have to upper bound $\beta_n$. Because $x^n_1(1, 1|1)$ is drawn uniformly at random from $\mathcal{T}_{Q_{X_1|UX_2}}(u^n, x^n_2)$,
\[
\beta_n = \sum_{x^n_1(1, 1|1) \in \mathcal{T}_{Q_{X_1|UX_2}}(u^n, x^n_2)} \frac{1}{|\mathcal{T}_{Q_{X_1|UX_2}}(u^n, x^n_2)|} \sum_{y^n_2} W^n(y^n_2|x^n_1, x^n_2) \mu_n(y^n_2).
\]
(45)
where
\[
\mu_n(y^n_2) := \mathbb{P}(\exists \tilde{m}_j' \neq 1 : \hat{I}(U^n(\tilde{m}_j'|1) \wedge Y^n_2(j)|X^n_2(1)) \\
\geq \hat{I}(U^n(1|1) \wedge Y^n_2(j)|X^n_2(1)) | (U^n(1|1), X^n_2(1)) = (u^n, x^n_2), Y^n_2(j) = y^n_2).
\]
(46)
Note that the event before conditioning in $\mu_n(y^n_2)$ does not depend on $\{X^n_1(1, 1|1) = x^n_1\}$ so we drop the dependence on $x^n_1$ from the notation $\mu_n(y^n_2)$. Intuitively $(U^n(1|1), X^n_2(1)) = (u^n, x^n_2)$ is the “cloud center” and $X^n_1(1, 1|1) = x^n_1$ the “satellite codeword”. Since $x^n_1 \in \mathcal{T}_{Q_{X_1|UX_2}}(u^n, x^n_2)$, our knowledge of the precise $x^n_1$ does not increase our knowledge of $U^n(\tilde{m}_j'|1), \tilde{m}_j' \neq 1$, which is the only source of randomness in the probability in (46). We continue to bound $\beta_n$ as follows:
\[
\beta_n \leq \sum_{x^n_1 \in \mathcal{T}_{Q_{X_1|UX_2}}(u^n, x^n_2)} (n + 1) |\mathbb{I}(|X_1||X_2)| \exp(-nH(Q_{X_1|UX_2}|Q_{UX_2})) \sum_{y^n_2} W^n(y^n_2|x^n_1, x^n_2) \mu_n(y^n_2)
\]
(47)
\[
= \sum_{x^n_1 \in \mathcal{T}_{Q_{X_1|UX_2}}(u^n, x^n_2)} (n + 1) |\mathbb{I}(|X_1||X_2)| Q^n_{X_1|UX_2}(x^n_1|u^n, x^n_2) \sum_{y^n_2} W^n(y^n_2|x^n_1, x^n_2) \mu_n(y^n_2)
\]
(48)
\[
\leq (n + 1) |\mathbb{I}(|X_1||X_2)| \sum_{y^n_2} \mu_n(y^n_2) \sum_{x^n_1} Q^n_{X_1|UX_2}(x^n_1|u^n, x^n_2) W^n(y^n_2|x^n_1, x^n_2)
\]
(49)
\[
\leq (n + 1) |\mathbb{I}(|X_1||X_2)| \sum_{y^n_2} W^n_{Y_2|UX_2}(y^n_2|u^n, x^n_2) \mu_n(y^n_2),
\]
(50)
where (47) follows by lower bounding the size of $\mathcal{T}_{Q_{X_1|UX_2}}(u^n, x^n_2)$ (Lemma 1), (48) follows by the fact that the $Q^n_{X_1|UX_2}(\cdot|u^n, x^n_2)$-probability of any sequence in $\mathcal{T}_{Q_{X_1|UX_2}}(u^n, x^n_2)$ is exactly $\exp(-nH(Q_{X_1|UX_2}|Q_{UX_2}))$ (Lemma 1), (49) follows by dropping the constraint $x^n_1 \in \mathcal{T}_{Q_{X_1|UX_2}}$, and (50) follows from the definition of
We have to perform the calculation leading to (50) because each $X^n_1$ is drawn uniformly at random from $T_{Q_{X_1|X_2}}(u^n, x^n_2)$ and not from the product measure $Q^n_{X_1|X_2}(\cdot | u^n, x^n_2)$, which would simplify the calculation and the introduction of the product (memoryless) channel $W^n_{Y_2|X_2}$. This change-of-measure technique (from constant-composition to product) was also used by Hayashi [27, Eqn. (76)] in his work on second-order coding rates for channel coding.

Hence it essentially remains to bound $\mu_n(y^n_2)$ in (46). By applying the union bound, we have

$$
\mu_n(y^n_2) \leq \min\{1, \exp(nR')\tau_n(y^n_2)\}
$$

(51)

where

$$
\tau_n(y^n_2) := \Pr\left(\hat{I}^{(2)}(U^n(2) \land Y^n_2(j)|X^n_2(1)) \geq \hat{I}(U^n(1) \land Y^n_2(j)|X^n_2(1)) \mid (U^n(1), X^n_2(1)) = (u^n, x^n_2), Y^n_2(j) = y^n_2\right).
$$

(52)

The only randomness now is in $U^n(2)$. Denote the conditional type of $y^n_2$ given $x^n_2$ as $P_{y^n_2|x^n_2}$. Now consider reverse channels $\tilde{V} : Y_2 \times X_2 \to U$ such that $\sum_{y_2} \tilde{V}(u|y_2, x_2) P_{y^n_2|x^n_2}(y_2|x_2) = Q_{U|X_2}(u|x_2)$ for all $(u, x_2, y_2)$, i.e., they are marginally consistent with $Q_{U|X_2}$. Denote this class of reverse channels as $\mathcal{R}(Q_{U|X_2})$. We have,

$$
\tau_n(y^n_2) = \sum_{\tilde{V} \in \mathcal{R}(Q_{U|X_2})} \Pr(U^n(2) \land \tilde{V}(y^n_2, x^n_2))
$$

(53)

By Lemma 3 (with the identifications $P \leftarrow Q_{X_2}, V \leftarrow Q_{U|X_2}, V' \leftarrow P_{y^n_2|x^n_2}$ and $W \leftarrow \tilde{V}$), for any $\tilde{V} \in \mathcal{R}(Q_{U|X_2})$,

$$
\Pr(U^n(2) \mid \tilde{V}(y^n_2, x^n_2)) \leq \exp(-nI(P_{y^n_2|x^n_2} \mid \tilde{V}(Q_{X_2})).
$$

(54)

As a result,

$$
\tau_n(y^n_2) \leq \sum_{\tilde{V} \in \mathcal{R}(Q_{U|X_2})} \exp(-nI(P_{y^n_2|x^n_2} \mid \tilde{V}(Q_{X_2}))
$$

(55)

$$
\leq \sum_{\tilde{V} \in \mathcal{R}(Q_{U|X_2})} \exp(-n\hat{I}(u^n \land y^n_2|x^n_2))
$$

(56)

$$
\leq \exp(-n\hat{I}(u^n \land y^n_2|x^n_2))
$$

(57)

Plugging this back into (51) yields,

$$
\mu_n(y^n_2) \leq \min\left\{1, \exp\left(-n(\hat{I}(u^n \land y^n_2|x^n_2) - R')\right)\right\}
$$

(58)

$$
\approx \exp\left[-n \left|\hat{I}(u^n \land y^n_2|x^n_2) - R'\right|^+\right].
$$

(59)

Plugging this back into (50) yields,

$$
\beta_n \leq \sum_{y^n_2} W^n_{Y_2|U,X_2}(y^n_2 | u^n, x^n_2) \exp\left[-n \left|\hat{I}(u^n \land y^n_2|x^n_2) - R'\right|^+\right]
$$

(60)

$$
= \sum_{V \in \mathcal{R}(Q_{Y_2|U,X_2})} W^n_{Y_2|U,X_2}(V | u^n, x^n_2) \exp\left[-n |I(Q_{U|X_2}, V|Q_{X_2}) - R'|^+\right]
$$

(61)

$$
\leq \sum_{V \in \mathcal{R}(Q_{Y_2|U,X_2})} \exp(-nD(V \mid W^n_{Y_2|U,X_2} | Q_{X_2})) \exp\left[-n |I(Q_{U|X_2}, V|Q_{X_2}) - R'|^+\right]
$$

(62)

$$
\approx \exp\left[-n \min_{V \in \mathcal{R}(Q_{Y_2|U,X_2})} \left(D(V \mid W^n_{Y_2|U,X_2} | Q_{X_2}) + |I(Q_{U|X_2}, V|Q_{X_2}) - R'|^+\right)\right]
$$

(63)

By appealing to continuity of exponents [8, Lem. 10.5] (minimum over conditional types is arbitrarily close to the minimum over conditional distributions for $n$ large enough), we obtain the exponent $F(R')$ in (25).
IV. COMPRESS-FORWARD (CF)

In this section, we state and prove an achievable error exponent for CF. CF is more complicated than PDF because the relay does vector quantization on the channel outputs \(Y^n_2\) and forwards the description to the destination. This quantized version of the channel output is denoted by \(\hat{Y}^n_2 \in \tilde{Y}^n_2\) and the error here is analyzed using covering techniques Marton introduced for deriving the error exponent for rate-distortion [35] and reviewed in Section II-D. Subsequently, the receiver decodes both the quantization index (a bin index associated with the quantized signal \(\hat{Y}^n_2\)) and the message index. This combination of covering and packing leads to a more involved analysis of the error exponent that needs to leverage on ideas in Kelly-Wagner [18] where the error exponent for Wyner-Ziv coding [4] was derived. It also leverages on a recently-developed proof technique by Scarlett-Guillén i Fàbregas [19] known as the one-at-a-time union bound given in [19]. This is particularly useful to analyze the error when two indices (of correlated signals) are to be simultaneously decoded given a channel output. At a high level, we operate on a conditional type-by-conditional type basis for the covering step at the relay. We also use an \(\alpha\)-decoding rule [21] for decoding the messages and the bin indices at the receiver.

This section is structured as follows: In Section IV-A, we provide basic definitions of the quantities that are used to state and prove the main theorem. The main theorem is stated in Section IV-B. A detailed set of remarks to help in understanding the quantities involved in the main theorem is provided in Section IV-C. Finally, the proof is provided in Section IV-D. The notation for the codewords follows that in El Gamal and Kim [3, Thm. 16.4].

A. Basic Definitions

Before we are able to state our result as succinctly as possible, we find it convenient to first define several quantities upfront. For CF, the following types and conditional types will be kept fixed and hence can be optimized over eventually: input distributions \(Q_{X_1} \in \mathcal{P}_n(X_1)\), \(Q_{X_2} \in \mathcal{P}_n(X_2)\) and test channel \(Q_{Y_2|Y_2} \in \mathcal{P}_n(Y_2; Q_{Y_2|X_1} Q_{X_2})\) for some (adversarial) channel realization \(Q_{Y_1|X_2} \in \mathcal{P}_n(Y_2; Q_{X_2})\) to be specified later.

1) Auxiliary Channels: Let the auxiliary channel \(W_{Q_{X_1}} : \mathcal{X}_2 \rightarrow \mathcal{Y}_2\) be defined as

\[
W_{Q_{X_1}}(y_2, y_3|x_2) := \sum_{x_1} W(y_2, y_3|x_1, x_2)Q_{X_1}(x_1).
\]  

(64)

This is simply the original relay channel averaged over \(Q_{X_1}\). With a slight abuse of notation, we denote its marginals using the same notation, i.e.,

\[
W_{Q_{X_1}}(y_3|x_2) := \sum_{y_2} W_{Q_{X_1}}(y_2, y_3|x_2),
\]

(65)

\[
W_{Q_{X_1}}(y_2|x_2) := \sum_{y_3} W_{Q_{X_1}}(y_2, y_3|x_2).
\]

(66)

Define another auxiliary channel \(W_{Q_{Y_2|X_2}} : \mathcal{X}_1 \times \mathcal{X}_2 \rightarrow \hat{Y}_2 \times \mathcal{Y}_3\)

\[
W_{Q_{Y_2|X_2}}(\hat{y}_2, y_3|x_1, x_2) := \sum_{y_2} W(y_3|x_1, x_2, y_2)Q_{Y_2|X_2}(\hat{y}_2|y_2, x_2)Q_{X_1}(y_2|x_2)
\]

(67)

where \(W(y_3|x_1, x_2, y_2) = W(y_2, y_3|x_1, x_2)/\sum_{y_2} W(y_2, y_3|x_1, x_2)\) is the \(\mathcal{Y}_3\)-conditional distribution of the DM-RC. Note that \(W_{Q_{Y_2|X_2}}\) is simply the original relay channel averaged over both channel realization \(Q_{Y_1|X_2}\) and test channel \(Q_{Y_2|X_2}\). Hence, if the realized conditional type of the relay input is \(Q_{Y_2|X_2}\) and we fixed the test channel to be \(Q_{Y_2|X_2}\) (to be chosen dependent on \(Q_{Y_2|X_2}\)), then we show that, effectively, the channel from \(\mathcal{X}_1 \times \mathcal{X}_2\) to \(\hat{Y}_2 \times \mathcal{Y}_3\) behaves as \(W_{Q_{Y_2|X_2}}\). We make this precise in the proofs. See the steps leading to (142).

2) Other Channels and Distributions: For any two channels \(Q_{Y_2|X_2}, \tilde{Q}_{Y_2|X_2} : \mathcal{X}_2 \rightarrow \mathcal{Y}_2\), define two \(\hat{Y}_2\)-modified channels as follows:

\[
Q_{Y_2|X_2}(\hat{y}_2|x_2) := \sum_{y_2} Q_{Y_2|X_2}(\hat{y}_2|y_2, x_2)Q_{Y_2|X_2}(y_2|x_2),
\]

(68)

\[
\tilde{Q}_{Y_2|X_2}(\hat{y}_2|x_2) := \sum_{y_2} \tilde{Q}_{Y_2|X_2}(\hat{y}_2|y_2, x_2)\tilde{Q}_{Y_2|X_2}(y_2|x_2).
\]

(69)
Implicit in these definitions are $Q_{Y_2|X_2}$, $Q_{Y_2|X_2}$ and $Q_{Y_2|X_2}$ but these dependencies are suppressed for the sake of brevity. For any $V: X_1 \times X_2 \times \hat{Y}_2 \rightarrow Y_3$, let the induced conditional distributions $V_{Q_{X_1}}: X_2 \times \hat{Y}_2 \rightarrow Y_3$ and $Q_{Y_2|X_2} \times V : X_1 \times X_2 \rightarrow \hat{Y}_2 \times Y_3$ be defined as

$$V_{Q_{X_1}}(y_3|x_2, \hat{y}_2) := \sum_{x_1} V(y_3|x_1, x_2, \hat{y}_2) Q_{X_1}(x_1), \quad (70)$$

$$(\hat{Q}_{Y_2|X_2} \times V)(\hat{y}_2, y_3|x_1, x_2) := V(y_3|x_1, x_2, \hat{y}_2) \hat{Q}_{Y_2|X_2}(\hat{y}_2|x_2). \quad (71)$$

3) Sets of Distributions and α-Decoder: Define the set of joint types $P_{X_1 X_2 Y_2}$ with marginals consistent with $Q_{X_1}, Q_{X_2}$ and $Q_{Y_2|X_2}$ as

$$\mathcal{P}_n(Q_{X_1}, Q_{X_2}, Q_{Y_2|X_2}) := \{P_{X_1 X_2 Y_2} \in \mathcal{P}(\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{Y}_2) : (P_{X_1}, P_{X_2}, P_{Y_2|X_2}) = (Q_{X_1}, Q_{X_2}, Q_{Y_2|X_2}) \}. \quad (72)$$

We will use the notation $\mathcal{P}(Q_{X_1}, Q_{X_2}, Q_{Y_2|X_2})$ (without subscript $n$) to mean the same set as in (72) without the restriction to types but all distributions in $\mathcal{P}(\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{Y}_2)$ satisfying the constraints in (72). For any four sequences $(x_1^n, x_2^n, \hat{y}_2^n, y_3^n)$, define the function $\alpha$ as

$$\alpha(x_1^n, \hat{y}_2^n, y_3^n|x_2^n) = \alpha(P, V) := D(V\|W_{Q_{Y_2|X_2}, Q_{Y_2|X_2}}|P) + H(V|P), \quad (73)$$

where $P$ is the joint type of $(x_1^n, x_2^n, \hat{y}_2^n)$, $V: \mathcal{X}_1 \times \mathcal{X}_2 \times \hat{Y}_2 \rightarrow \mathcal{Y}_3$ is the conditional type of $y_3^n$ given $(x_1^n, x_2^n, \hat{y}_2^n)$, and $W_{Q_{Y_2|X_2}, Q_{Y_2|X_2}}$ is the channel defined in (67). Roughly speaking, to decode the bin index, we will maximize $\alpha$ over bin indices, messages and conditional types $Q_{Y_2|X_2}$. This is exactly ML decoding as discussed in Section II.D. Define the set of conditional types

$$\mathcal{K}_n(Q_{Y_2|X_2}, Q_{Y_2|X_2}) := \{V \in \mathcal{P}_n(\mathcal{Y}_3 : Q_{X_1}, Q_{X_2}, Q_{Y_2|X_2}) : \alpha(Q_{X_1}, Q_{X_2}, Q_{Y_2|X_2}, V) \geq \alpha(Q_{X_1}, Q_{X_2}, Q_{Y_2|X_2}, W_{Q_{Y_2|X_2}, Q_{Y_2|X_2}}) \}. \quad (74)$$

Note that $Q_{Y_2|X_2}$ is given in (68) and is induced by the two arguments of $\mathcal{K}_n$. Intuitively, the conditional types contained in $\mathcal{K}_n(Q_{Y_2|X_2}, Q_{Y_2|X_2})$ are those corresponding to sequences $y_3^n$ that lead to an error as the likelihood computed with respect to $V$ is larger than (or equal to) that for the true averaged channel $W_{Q_{Y_2|X_2}, Q_{Y_2|X_2}}$. The marginal types $Q_{X_1}, Q_{X_2}$ are fixed in the notation $\mathcal{K}_n$ but we omit them for brevity. We will use the notation $\mathcal{K}(Q_{Y_2|X_2}, Q_{Y_2|X_2})$ (without subscript $n$) to mean the same set as in (74) without the restriction to conditional types but all conditional distributions from $\mathcal{X}_1 \times \mathcal{X}_2 \times \hat{Y}_2$ to $\mathcal{Y}_3$ satisfying the constraints in (74) so $\mathcal{K} = \text{cl}(\lim_{n \rightarrow \infty} \mathcal{K}_n)$ where $\text{cl}(\cdot)$ is the closure operation.

B. Error Exponent for Compress-Forward

Theorem 2 (Error Exponent for Compress-Forward). Fix $b \in \mathbb{N}$ and “Wyner-Ziv rate” $R_2 \geq 0$, distributions $Q_{X_1} \in \mathcal{P}(\mathcal{X}_1)$ and $Q_{X_2} \in \mathcal{P}(\mathcal{X}_2)$ and auxiliary alphabet $\hat{Y}_2$. The following is a lower bound on the reliability function:

$$E(R_{\text{eff}}) \geq E_{\text{cl}}^{(b)}(R_{\text{eff}}) := \frac{1}{b} \min \{ G_1(R, R_2), G_2(R, R_2) \} \quad (75)$$

where the constituent exponents are defined as

$$G_1(R, R_2) := \min_{V, X_2 \rightarrow Y_3} D(V\|W_{Q_{X_1}}|Q_{X_2}) + |I(Q_{X_2}, V) - R_2|^{+} \quad (76)$$

where $W_{Q_{X_1}}$ is defined in (65) and

$$G_2(R, R_2) := \min_{Q_{Y_2|X_2}, \hat{Y}_2} D(Q_{Y_2|X_2}|W_{Q_{X_1}}|Q_{X_2}) + \max_{Q_{Y_2|X_2}, \hat{Y}_2} |J(R, R_2, Q_{Y_2|X_2}, Q_{Y_2|X_2}, Q_{Y_2|X_2})). \quad (77)$$
The quantity $J(R, R_2, Q_{Y_2}|X_2, Q_{Y_2}|X_2)$ that constitutes $G_2(R, R_2)$ is defined as

$$J(R, R_2, Q_{Y_2}|X_2, Q_{Y_2}|X_2)$$

$$:= \min_{P_{X_1, X_2, Y_2} \in \mathcal{P}(Q_{X_1}, Q_{X_2}, Q_{Y_2}|X_2)} \left\{ D(P_{Y_2|X_2, Q_{Y_2}|X_2, X_2} \parallel W_{Q_{Y_2|x_2}, Q_{Y_2|x_2}} | Q_{X_1}, Q_{X_2}) ight. $$

$$+ \left. \min_{Q_{Y_2|x_2} : X_2 \rightarrow Y_2} \min_{V \in \mathcal{V}(Q_{Y_2|x_2}, Q_{Y_2|x_2})} \left[ \min_{l=1,2} \psi_l(V, \bar{Q}_{Y_2|x_2}, R, R_2, P_{X_1, X_2, \bar{Y}_2}) \right] \right\}$$

(78)

where $W_{Q_{Y_2|x_2}, Q_{Y_2|x_2}}$, $Q_{Y_2|x_2}$, and $\bar{Q}_{Y_2|x_2}$ are defined in (67), (68) and (69) respectively and the functions $\psi_l, l = 1, 2$ are defined as

$$\psi_1(V, \bar{Q}_{Y_2|x_2}, R, R_2, P_{X_1, X_2, \bar{Y}_2}) := |I(Q_{X_1}, \bar{Q}_{Y_2|x_2} \times V | Q_{X_2}) - R|^+$$

(79)

and

$$\psi_2(V, \bar{Q}_{Y_2|x_2}, R, R_2, P_{X_1, X_2, \bar{Y}_2}) :=$$

$$\begin{cases} 
|I(\bar{Q}_{Y_2|x_2}, V_{Q_{X_1}} | Q_{X_2}) + |I(Q_{X_1}, \bar{Q}_{Y_2|x_2} \times V | Q_{X_2}) - R|^+ \\
R_2 - I(\bar{Q}_{Y_2|x_2}, Q_{Y_2|x_2}) | Q_{X_2}) |^+ 
\end{cases}$$

if $R_2 \leq I(\bar{Q}_{Y_2|x_2}, Q_{Y_2|x_2}) | Q_{X_2})$

$$I(\bar{Q}_{Y_2|x_2}, V_{Q_{X_1}} | Q_{X_2}) + |I(Q_{X_1}, \bar{Q}_{Y_2|x_2} \times V | Q_{X_2}) - R|^+$$

if $R_2 > I(\bar{Q}_{Y_2|x_2}, Q_{Y_2|x_2}) | Q_{X_2})$

(80)

where $V_{Q_{X_1}}$ and $\bar{Q}_{Y_2|x_2} \times V$ are defined in (70) and (71) respectively.

Note that (80) can be written more succinctly as

$$\psi_2(V, \bar{Q}_{Y_2|x_2}, R, R_2, P_{X_1, X_2, \bar{Y}_2})$$

$$:= \left| |I(Q_{X_1}, \bar{Q}_{Y_2|x_2} \times V) - R|^+ + I(\bar{Q}_{Y_2|x_2}, V_{Q_{X_1}} | Q_{X_2}) - |I(\bar{Q}_{Y_2|x_2}, Q_{Y_2|x_2} | Q_{X_2}) - R_2|^+ \right|^+.$$  

(81)

C. Remarks on the Error Exponent for Compress-Forward

In this Section, we dissect the main features of the CF error exponent presented in Theorem[2] To help the reader follow the proof, the point at which the various terms in the CF exponent are derived are summarized in Table II.

1) We are free to choose the independent input distributions $Q_{X_1}$ and $Q_{X_2}$, though these will be $n$-types for finite $n \in \mathbb{N}$. We also have the freedom to choose any “Wyner-Ziv rate” $R_2 \geq 0$. Thus, we can optimize over $Q_{X_1}$, $Q_{X_2}$ and $R_2$. The $X_1$- and $X_2$-codewords are uniformly distributed in $T_{Q_{X_1}}$ and $T_{Q_{X_2}}$ respectively.

2) As is well known in CF [2], the relay transmits a description $\tilde{y}_2^n(j)$ of its received sequence $y_2^n(j)$ (conditioned on $x_2^n(j)$ which is known to both relay and decoder) via a covering step. This explains the final mutual
information term in \(80\) which can be written as the rate loss \(I(Y_2; \hat{Y}_2|X_2)\), where \((X_2, Y_2, \hat{Y}_2)\) is distributed as \(Q_{X_2} \times Q_{Y_2|X_2} \times Q_{\hat{Y}_2|Y_2,X_2}\). Since covering results in super-exponential decay in the error probability, this does not affect the overall exponent since the smallest one dominates. See the steps leading to \(104\) in the proof.

3) The exponent \(G_1(R, R_2)\) in \(75\) is analogous to \(G(R')\) in \(26\). This represents the error exponent in the estimation of \(X_2\)'s index given \(Y_2^n\) using MMI decoding. However, in the CF proof, we do not use the packing lemma \([8\text{, Lem. 10.1}]\). Rather we construct a random code and show that on expectation (over the random code), the error probability decays exponentially fast with exponent \(G_1(R, R_2)\). This is the same as in Remark \(6\text{ in Section III-B on the discussion of the PDF exponent}. In fact, other authors such as Kelly-Wagner \([18]\) and Moulin-Wang \([28]\) also derive their error exponents in this way.

4) In the exponent \(G_2(R, R_2)\) in \(77\), \(Q_{Y_2|X_2}\) is the realization of the conditional type of the received signal at the relay \(y_2^n(j)\) given \(x_2^n(j)\). The divergence term \(D(Q_{Y_2|X_2}||W_{Q_{X_1}|Q_{X_2}})\) represents the deviation from the true channel behavior \(W_{Q_{X_1}}\) similarly to the interpretation of the random coding error exponent for point-to-point channel coding in \(12\). We can optimize for the conditional distribution (test channel) \(Q_{Y_2\mid Y_2,X_2}\) compatible with \(Q_{Y_2|X_2}Q_{X_2}\). This explains the outer minimization over \(Q_{Y_2|X_2}\) and inner maximization over \(Q_{Y_2|Y_2,X_2}\) in \(77\). This is a game-theoretic-type result along the same lines as in \([13\text{, 28, 29}].

5) The first part of \(J\) given by \(\psi_1\) in \(79\) represents the incorrect decoding of the index of \(X_1^n\) (message \(M_j\)) as well as the conditional type \(Q_{Y_2|X_2}\) given that the bin index of the description \(\hat{Y}_2^n\) is decoded correctly. The second part of \(J\) given by \(\psi_2\) in \(80\) represents the incorrect decoding the bin index of \(\hat{Y}_2^n\), the index of \(X_1^n\) (message \(M_j\)) as well as the conditional type \(Q_{Y_2|X_2}\). We see the different sources of \(\text{“errors”}\) in \(78\): There is a minimization over the different types of channel behavior represented by \(P_{X_1,X_2\hat{Y}_2Y_3}\) and also a minimization over estimated conditional types \(Q_{Y_2|X_2}\). Subsequently, the error involved in \(\alpha\)-decoding of the message and the bin index of the description sequence is represented by the minimization over \(V \in \mathcal{K}(Q_{Y_2|X_2}, Q_{\hat{Y}_2|Y_2,X_2})\).

6) We see that the freedom of choice of the “Wyner-Ziv rate” \(R_2 \geq 0\) allows us to operate in one of two distinct regimes. This can be seen from the two different cases in \(80\). The number of description codewords in \(Y_2^n\) is designed to be \(\exp(nI(\hat{Q}_{Y_2|X_2}, Q_{Y_2|Y_2,X_2}|Q_{X_2})) = \exp(nI(\hat{Y}_2; Y_2|X_2))\) to first order in the exponent, where the choice of \(Q_{Y_2|Y_2,X_2}\) depends on the realized conditional type \(Q_{Y_2|X_2}\). The number of Wyner-Ziv bins is \(\exp(nR_2)\). When \(R_2 \leq I(\hat{Q}_{Y_2|X_2}, Q_{\hat{Y}_2|Y_2,X_2}|Q_{X_2})\), we do additional Wyner-Ziv binning as there are more description sequences than bins. If \(R_2\) is larger than \(I(\hat{Q}_{Y_2|X_2}, Q_{\hat{Y}_2|Y_2,X_2}|Q_{X_2})\), no additional binning is required. The excess Wyner-Ziv rate is thus

\[
\Delta R_2 := I(\hat{Q}_{Y_2|X_2}, Q_{Y_2|Y_2,X_2}|Q_{X_2}) - R_2.
\]

This explains the presence of this term in \(80\) or equivalently, \(81\).

7) For the analysis of the error in decoding the bin and message indices, if we simply apply the packing lemmas in \([8\text{, 13–17, 21}]\), this would result in a suboptimal rate vis-à-vis CF. This is because the conditional correlation of \(X_1\) and \(\hat{Y}_2\) given \(X_2\) would not be taken into account. Thus, we need to analyze this error exponent more carefully using the one-at-a-time union bound \(19\) used originally for the multiple-access channel under mismatched decoding and stated in \(19\). Note that the sum of the first two mutual informations (ignoring the \(| \cdot |^{\perp}\) in \(80\), which represents the bound on the sum of the message rate and the description sequence rate (cf. \([8\text{ Eq. (16.11)}]\)), can be written as

\[
I(\hat{Y}_2; Y_3|X_2) + I(X_1; \hat{Y}_2, Y_3|X_2)
= H(X_1|X_2) + H(\hat{Y}_2|X_2) + H(Y_3|X_2) - H(X_1, \hat{Y}_2, Y_3|X_2),
\]

where \((X_1, X_2, \hat{Y}_2, Y_3) \sim Q_{X_1} \times Q_{X_2} \times Q_{Y_2|X_2} \times V\). The entropies in \(83\) demonstrate the symmetry between \(X_1\) and \(\hat{Y}_2\) when they are decoded jointly at the receiver \(Y_3\). The expressions in \(83\)–\(84\) are in fact the \textit{conditional mutual information among three variables} \(I(X_1; Y_2; Y_3|X_2)\) defined, for example, in Liu and Hughes \([38\text{ Sec. III-B}]. This quantity is also called the \textit{multi-information} by Studený and Vejnarová \([42\). In addition, the proof shows that by modifying the order of applying union bounds in \(19\), we can get another achievable exponent. Indeed, \(\psi_2\) in \(81\) can be strengthened to be the maximum of the expression on its
right-hand-side and
\[
\psi_2(V, \tilde{Q}_{Y_2|X_2}, R, R_2, P_{X_1X_2Y_2Y_3}) := \left| I(Q_{X_1}, \tilde{Q}_{Y_2|X_2} \times V) - R + \left| I(\tilde{Q}_{Y_2|X_2}, V_{Q_{X_1}}|Q_{X_2}) - \Delta R_2 \right| + \right|, \tag{85}
\]
where \(\Delta R_2\) is defined in (82).

8) From the exponents in Theorem 2, it is clear upon eliminating \(R_2\) (if \(R_2\) is chosen small enough so that Wyner-Ziv binning is necessary) that we recover the CF lower bound in (2). Indeed, if \(\psi_1\) is active in the minimization in (78), the first term in (2) is positive if and only if the error exponent \(G_2\) is positive for some choice of distributions \(Q_{X_1}, Q_{X_2}\) and \(Q_{Y_2|X_2}\). Also, if \(\psi_2\) is active in the minimization in (78) and \(R_2\) is chosen sufficiently small (so that the first clause of (80) is active), \(G_2\) is positive if
\[
R < R_2 + I(X_1; \hat{Y}_2Y_3|X_2) + I(\hat{Y}_2; Y_2|X_2) - I(\hat{Y}_2; Y_2|X_2) \tag{86}
\]
\[
< I(X_2; Y_3) + I(X_1; \hat{Y}_2Y_3|X_2) + I(\hat{Y}_2; Y_2|X_2) - I(\hat{Y}_2; Y_2|X_2) \tag{87}
\]
\[
= I(X_1X_2; Y_3) + I(X_1; \hat{Y}_2X_2Y_3) + I(\hat{Y}_2; Y_2|X_2) - I(X_1X_2; \hat{Y}_2X_2Y_2) \tag{88}
\]
\[
= I(X_1X_2; Y_3) + I(X_1Y_3; \hat{Y}_2|X_2) - I(X_1Y_3; \hat{Y}_2X_2Y_2) \tag{89}
\]
\[
= I(X_1X_2; Y_3) - I(Y_2; \hat{Y}_2X_1X_2Y_3). \tag{90}
\]
for some \(Q_{X_1}, Q_{X_2}\) and \(Q_{Y_2|X_2}\). In (87), we used the fact that \(G_1\) in (76) is positive if and only if \(R_2 < I(X_2; Y_3)\) and in (88) we also used the chain rule for mutual information (twice) and the Markov chain \(\hat{Y}_2 - (X_2, Y_2) - (X_1, Y_3)\) [3] pp. 402 so the final mutual information term \(I(X_1X_3; \hat{Y}_2X_2Y_2) = 0\). Equations (89) and (90) follow by repeated applications of the chain rule for mutual information. Equation (90) matches the second term in (2).

9) Lastly, the evaluation of the CF exponent appears to be extremely difficult because of (i) non-convexity of the optimization problem and (ii) multiple nested optimizations. It is, however, not apparent how to simplify the CF exponent (to the Gallager form, for example) to make it amenable to evaluation for a given DM-RC \(W\).

D. Proof of Theorem 2

Proof: Random Codebook Generation: We again use block-Markov coding [2]. Fix types \(Q_{X_1} \in \mathcal{P}_n(X_1)\) and \(Q_{X_2} \in \mathcal{P}_n(X_2)\) as well as rates \(R, R_2 \geq 0\). For each \(j \in [b]\), generate a random codebook in the following manner. Randomly and independently generate \(\exp(nR)\) codewords \(x^n_i(m_j) \sim \text{Unif}[\mathcal{T}_Q X_1]\), where \(\text{Unif}[\mathcal{A}]\) is the uniform distribution over the finite set \(\mathcal{A}\). Randomly and independently generate \(\exp(nR_2)\) codewords \(x^n_2((l-1)) \sim \text{Unif}[\mathcal{T}_Q X_2]\). Now for every \(Q_{Y_2|X_2} \in \mathcal{Y}_{S_2} (Y_2; Q_{X_2})\) fix a different test channel \(Q_{Y_2|X_2} (Q_{Y_2|X_2}) \in \mathcal{Y}_n (Q_2; Q_{Y_2|X_2})\). For every \(Q_{Y_2|X_2} \in \mathcal{Y}_{S_2} (Q_{Y_2|X_2})\) and every \(x^n_2((l-1))\) construct a conditional type-dependent codebook \(B(Q_{Y_2|X_2}, X_1); l_1\) \(B(Q_{Y_2|X_2}, X_1); l_1\) whose rate, which we call the inflated rate, satisfies
\[
\tilde{R}_2(Q_{Y_2|X_2}) := \frac{1}{n} \log |B(Q_{Y_2|X_2}, X_1); l_1| = I(Q_{Y_2|X_2}, Q_{Y_2|X_2} Q_{Y_2|X_2} (Q_{Y_2|X_2}) + \nu_n, \tag{91}
\]
where \(\nu_n \in \Theta(\frac{\log n}{n})\) and more precisely,
\[
\frac{(|X_2| |Y_2| \hat{Y}_2) + 2) \log (n + 1)}{n} \leq \nu_n \leq \frac{(|X_2| |Y_2| \hat{Y}_2 + 3) \log (n + 1)}{n}. \tag{92}
\]
Each sequence in \(B(Q_{Y_2|X_2}, X_1); l_1\) is indexed as \(y^n_2(k_j l_1)\) and is drawn independently according to the uniform distribution \(\text{Unif}[\mathcal{T}_Q Y_2|X_2 (x^n_2((l-1)))\)] where \(Q_{Y_3|X_2}\) is the marginal induced by \(Q_{Y_2|X_2}\) and \(Q_{Y_2|X_2} Q_{Y_2|X_2} (Q_{Y_2|X_2})\). See the definition in (68). Depending on the choice of \(R_2\), do one of the following:

- If \(R_2 \leq \tilde{R}_2(Q_{Y_2|X_2})\), partition the conditional type-dependent codebook \(B(Q_{Y_2|X_2}, X_1); l_1\) into \(\exp(nR_2)\) equal-sized bins \(B_j (Q_{Y_2|X_2}, X_1); l_1, l_j \in [\exp(nR_2)]\).
- If \(R_2 > \tilde{R}_2(Q_{Y_2|X_2})\), assign each element of \(B(Q_{Y_2|X_2}, X_1); l_1\) a unique index in \([\exp(nR_2)]\).

Transmitter Encoding: The encoder transmits \(x^n_1(m_j)\) at block \(j \in [b]\).
Relay Encoding: At the end of block \( j \in [b] \), the relay encoder has \( x_2^n(l_{j-1}) \) (by convention \( l_0 := 1 \)) and its input sequence \( y_2^n(j) \). It computes the conditional type \( Q_{Y_2|X_2} \in \mathcal{Y}_n(Y_2; Q_{X_2}) \). Then it searches in \( B(Q_{Y_2|X_2}, l_{j-1}) \) for a description sequence \( \hat{y}_2^n(\hat{k}_j|l_{j-1}) \in T_{Q_{Y_2|X_2}}(Q_{Y_2|X_2})(y_2^n(j), x_2^n(l_{j-1})). \) (93)

If more than one such sequence exists, choose one uniformly at random in \( B(Q_{Y_2|X_2}, l_{j-1}) \) from those satisfying (93). If none exists, choose uniformly at random from \( B(Q_{Y_2|X_2}, l_{j-1}) \). Identify the bin index \( \hat{l}_j \) of \( \hat{y}_2^n(\hat{k}_j|l_{j-1}) \) and send \( x_2^n(\hat{l}_j) \) in block \( j + 1 \).

Decoding: At the end of block \( j + 1 \), the receiver has the channel output \( y_2^n(j + 1) \). It does MMI decoding [7] by finding \( \hat{l}_j \) satisfying

\[
\hat{l}_j := \arg \max_{l_j \in [\exp(nR_2)]} \hat{I}(x_2^n(l_j) \cap y_2^n(j + 1)).
\] (94)

Having identified \( \hat{l}_{j-1}, \hat{l}_j \) from (94), find message \( \hat{m}_j \), index \( \hat{k}_j \) and conditional type \( \hat{Q}^{(j)}_{Y_2|X_2} \in \mathcal{Y}_n(Y_2; Q_{X_2}) \) satisfying

\[
(\hat{m}_j, \hat{k}_j, \hat{Q}^{(j)}_{Y_2|X_2}) = \arg \max_{(m_j, k_j, Q_{Y_2|X_2}); \hat{y}_2^n(k_j, \hat{l}_{j-1}) \in B_{l_j}(Q_{Y_2|X_2}, \hat{l}_{j-1})} \alpha \left( x_2^n(m_j), \hat{y}_2^n(k_j, \hat{l}_{j-1}), y_2^n(j) \mid x_2^n(\hat{l}_{j-1}) \right),
\] (95)

where the function \( \alpha \) was defined in (73). This is an \( \alpha \)-decoder [21] which finds the \((\hat{m}_j, \hat{k}_j, \hat{Q}^{(j)}_{Y_2|X_2}) \) maximizing \( \alpha \) subject to \( \hat{y}_2^n(k_j, \hat{l}_{j-1}) \in B_{l_j}(\hat{Q}^{(j)}_{Y_2|X_2}, \hat{l}_{j-1}) \), where \( \hat{l}_{j-1}, \hat{l}_j \) were found in (94). We decode \( Q_{Y_2|X_2} \) so as to know which bin \( \hat{y}_2^n(k_j, \hat{l}_{j-1}) \) lies in. Since there are only polynomially many conditional types \( Q_{Y_2|X_2} \) this does not degrade our CF error exponent. The decoding of the conditional type is inspired partly by Moulin-Wang’s derivation of the error exponent for Gel’fand-Pinsker coding [28] pp. 138]. Declare that \( \hat{m}_j \) was sent.

Let us pause to understand why we used two different rules (MMI and ML-decoding) in (94) and (95). In the former, we are simply decoding a single index \( l_j \) given \( y_2^n(j + 1) \) hence MMI suffices. In the latter, we are decoding two indices \( m_j \) and \( k_j \) and as mentioned in Section [A], we need to take into account the conditional correlation between \( X_1 \) and \( Y_2 \). If we had simply used an MMI decoder, it appears that we would have a strictly smaller quantity \( I(\hat{Y}_2; Y_3|X_2) + I(X_1; Y_3|X_2 \hat{Y}_2) = I(X_1 \hat{Y}_2; Y_3|X_2) \) in the analogue of (55)–(56), which represents the upper bound on the sum of \( R \) and the excess Wyner-Ziv rate \( \Delta R_2 \) defined in (82). This would not recover the CF lower bound in the steps from (86)–(90). Hence, we used the ML-decoder in (95).

Analysis of Error Probability: We now analyze the error probability. Assume as usual that \( M_j = 1 \) for all \( j \in [b - 1] \) and let \( L_{j-1}, L_j \) and \( K_j \) be indices chosen by the relay in block \( j \). First, note that as in (38),

\[
\mathbb{P}(\hat{M} \neq M) \leq (b - 1) (\epsilon_R + 2\epsilon_{D,1} + \epsilon_{D,2}),
\] (96)

where \( \epsilon_R \) is the error event that there is no description sequence \( y_2^n(\hat{k}_j|l_{j-1}) \) in the bin \( B(Q_{Y_2|X_2}, l_{j-1}) \) that satisfies (93) (covering error),

\[
\epsilon_{D,1} := \mathbb{P}(\hat{L}_{j} \neq L_{j})
\] (97)

is the error probability in decoding the wrong \( L_j \) bin index, and

\[
\epsilon_{D,2} := \mathbb{P}(\hat{M}_j \neq 1|L_j, L_{j-1} \text{ decoded correctly})
\] (98)

is the error probability in decoding the message incorrectly. See the proof of compress-forward in [3] Thm. 16.4) for details of the calculation in (96). Again, since \( b \) is a constant, it does not affect the exponential dependence on the error probability in (96). We bound each error probability separately. Note that the error probability is an average over the random codebook generation so, by the usual random coding argument, as long as this average is small, there must exist at least one code with such a small error probability.

Covering Error \( \epsilon_R \): For \( \epsilon_R \), we follow the proof idea in [18] Lem. 2]. In the following, we let the conditional type of \( y_2^n \) given \( x_2^n \) be \( Q_{Y_2|X_2} \). Consider the conditional covering error conditioned on \( X_2^n = x_2^n \) and \( Y_2^n = y_2^n \), namely

\[
\epsilon_R(x_2^n, y_2^n) := \mathbb{P}(F|Y_2^n = y_2^n, X_2^n = x_2^n),
\] (99)
where \( \mathcal{F} \) is the event that every sequence \( \hat{y}_2^n(k_j|l_{j-1}) \in \mathcal{B}(Q_{Y_2|X_2}, l_{j-1}) \) does not satisfy (93). Let \( \exp_e(t) := e^t \). (\( e \) is the base of the natural logarithm.) Now we use the mutual independence of the codewords in \( \mathcal{B}(Q_{Y_2|X_2}, l_{j-1}) \) and basic properties of types (Lemmas 1 and 2) to assert that

\[
\epsilon_R(x_2^n, y_2^n) = \prod_{k_j} \mathbb{P} \left( \hat{Y}_2^n(k_j|l_{j-1}) \notin \mathcal{T}_{Q_{Y_2|X_2}, y_2^n}(Q_{Y_2|X_2}) \left( y_2^n, x_2^n \right) \right) 
\]

\[
= \left[ 1 - \mathbb{P} \left( \hat{Y}_2^n(1|l_{j-1}) \in \mathcal{T}_{Q_{Y_2|X_2}, y_2^n}(Q_{Y_2|X_2}) \left( y_2^n, x_2^n \right) \right) \right]^{|\mathcal{B}(Q_{Y_2|X_2}, l_{j-1})|} 
\]

\[
\leq \left[ 1 - (n+1)^{-|X_2||Y_2||\mathcal{B}_2|} \exp(-nI(Q_{Y_2|X_2}, Q_{Y_2|X_2}(Q_{Y_2|X_2})|Q_{X_2}) \right)^{|\mathcal{B}(Q_{Y_2|X_2}, l_{j-1})|} 
\]

\[
\leq \exp_{e} \left[ -(n+1)^{-|X_2||Y_2||\mathcal{B}_2|} \exp(-nI(Q_{Y_2|X_2}, Q_{Y_2|X_2}(Q_{Y_2|X_2})|Q_{X_2}) - \tilde{R}_2(Q_{Y_2|X_2}) \right] 
\]

\[
\leq \exp_{e} \left[ -(n+1)^2 \right], 
\]

where the product in (100) extends over all indices \( k_j \) for which \( \hat{y}_2^n(k_j|l_{j-1}) \in \mathcal{B}(Q_{Y_2|X_2}, l_{j-1}) \) for some fixed realization of \( l_{j-1} \) which we can condition on. Inequality (103) follows from the inequality \((1-x)^k \leq \exp_e(-kx)\), (104) follows from the choice of \( \tilde{R}_2(Q_{Y_2|X_2}) \) and \( \nu_n \) in (91) and (92) respectively. This derivation is similar to the type covering lemma for source coding with a fidelity criterion (rate-distortion) with excess distortion probability by Marton [35]. Now, let \( \mathcal{E}_C \) be the set of all pairs \((x_2^n, y_2^n)\) that lead to a covering error according to the decoding rule (93). We follow the argument proposed by Kelly-Wagner [18, pp. 5100] to assert that

\[
\epsilon_R = \sum_{(x_2^n, y_2^n) \in \mathcal{E}_C} \mathbb{P}(X_2^n = x_2^n, Y_2^n = y_2^n) 
\]

\[
= \sum_{(x_2^n, y_2^n) \in \mathcal{E}_C} \mathbb{P}(X_2^n = x_2^n, Y_2^n = y_2^n, \mathcal{F}) 
\]

\[
\leq \sum_{(x_2^n, y_2^n) \in \mathcal{E}_C} \mathbb{P}(\mathcal{F}|X_2^n = x_2^n, Y_2^n = y_2^n) 
\]

\[
= \sum_{(x_2^n, y_2^n) \in \mathcal{E}_C} \epsilon_R(x_2^n, y_2^n) 
\]

\[
\leq \sum_{(x_2^n, y_2^n) \in \mathcal{E}_C} \exp_{e} \left[ -(n+1)^2 \right] = 0. 
\]

The punchline is that \( \epsilon_R \) decays super-exponentially (i.e., \( \epsilon_R = 0 \) or the exponent is infinity) and thus it does not affect the overall error exponent since the smallest one dominates.

**First Packing Error** \( \epsilon_{D,1} \): We assume \( L_j = 1 \) here. The calculation here is very similar to that in Section III-C but we provide the details for completeness. We evaluate \( \epsilon_{D,1} \) by partitioning the sample space into subsets where \( X_2^n(1) \) takes on various values \( x_2^n \in \mathcal{T}_{Q_{X_2}} \). Thus, we have

\[
\epsilon_{D,1} \leq \mathbb{P} \left( \exists \tilde{l}_j \neq 1 : \tilde{I}(X_2^n(\tilde{l}_j) \wedge Y_3^n(j+1)) \geq I(X_2^n(1) \wedge Y_3^n(j+1)) \right) 
\]

\[
= \sum_{x_2^n \in \mathcal{T}_{Q_{X_2}}} \frac{1}{|\mathcal{T}_{Q_{X_2}}|} \beta_n(x_2^n) 
\]

where

\[
\beta_n(x_2^n) := \mathbb{P} \left( \exists \tilde{l}_j \neq 1 : \tilde{I}(X_2^n(\tilde{l}_j) \wedge Y_3^n(j+1)) \geq I(X_2^n(1) \wedge Y_3^n(j+1)) \mid X_2^n(1) = x_2^n \right). 
\]

It can easily be seen that \( \beta_n(x_2^n) \) is independent of \( x_2^n \in \mathcal{T}_{Q_{X_2}} \) so we abbreviate \( \beta_n(x_2^n) \) as \( \beta_n \). Because \( X_2^n(1) \) is generated uniformly at random from \( \mathcal{T}_{Q_{X_1}} \), we have

\[
\beta_n = \sum_{x_1^n \in \mathcal{T}_{Q_{X_1}}} \frac{1}{|\mathcal{T}_{Q_{X_1}}|} \sum_{y_3^n} W^n(y_3^n|x_1^n, x_2^n) \mu_n(y_3^n) 
\]
where
\[
\mu_n(y_3^n) := \mathbb{P}\left( \exists \tilde{I}_j \neq 1 : \tilde{I}(X_2^n | Y_3^n(j + 1)) \geq \tilde{I}(X_2^n(1) \wedge Y_3^n(j + 1)) \mid Y_3^n(j + 1) = y_3^n, X_2^n(1) = x_2^0 \right),
\]
(114)

Note that the event before the conditioning in \(\mu_n(y_3^n)\) does not depend on the event \(\{X_1^n(1) = x_1^n\}\) so we drop the dependence on \(x_1^n\) from the notation \(\mu_n(y_3^n)\). Also the only source of randomness in the probability in (114) is \(X_2^n(\tilde{I}_j), \tilde{I}_j \neq 1\) which is independent of \(X_1^n(1)\). We continue to bound \(\beta_n\) as follows:
\[
\begin{align*}
\beta_n & \leq \sum_{x_1^n \in \mathcal{T}_{Q_x}} (n + 1)^{|X_1|} \exp(-nH(Q_{X_1})) \sum_{y_3^n} W_n(y_3^n|x_1^n, x_2^n) \mu_n(y_3^n) \\
& = \sum_{x_1^n \in \mathcal{T}_{Q_x}} (n + 1)^{|X_1|} Q_{X_1}(x_1^n) \sum_{y_3^n} W_n(y_3^n|x_1^n, x_2^n) \mu_n(y_3^n) \\
& \leq (n + 1)^{|X_1|} \sum_{y_3^n} \mu_n(y_3^n) \sum_{x_1^n} Q^{\tilde{I}}_{X_1}(x_1^n) W_n(y_3^n|x_1^n, x_2^n) \\
& = (n + 1)^{|X_1|} \sum_{y_3^n} W_n^\tilde{I}(y_3^n|x_2^n) \mu_n(y_3^n).
\end{align*}
\]
(115)
(116)
(117)
(118)

where (115) follows from the lower bound on the size of a type class (Lemma 1), (116) follows from the fact that the \(Q_{X_1}^n\)-probability of a sequence \(x_1^n\) of type \(Q_{X_1}\) is exactly \(\exp(-nH(Q_{X_1}))\) and (118) is an application of the definition of \(W_{Q_x^n}\) in (65). It remains to bound \(\mu_n(y_3^n)\) in (114). We do so by first applying the union bound
\[
\mu_n(y_3^n) \leq \min \{1, \exp(nR_2) \tau_n(y_3^n)\},
\]
(119)

where
\[
\tau_n(y_3^n) := \mathbb{P}\left( \tilde{I}(X_2^n(2) \wedge Y_3^n(j + 1)) \geq \tilde{I}(X_2^n(1) \wedge Y_3^n(j + 1)) \mid Y_3^n(j + 1) = y_3^n, X_2^n(1) = x_2^0 \right).
\]
(120)

We now use notation \(\tilde{V} : \mathcal{X}_3 \rightarrow \mathcal{X}_2\) to denote a reverse channel. Also let \(P_{y_3^n}^\tilde{V}\) be the type of \(y_3^n\). Let \(\mathcal{R}(Q_{X_2})\) be the class of reverse channels satisfying \(\sum_{y_3^n} \tilde{V}(x_2^n | y_3^n) P_{y_3^n}^\tilde{V} = Q_{X_2}(x_2)\). Then, we have
\[
\tau_n(y_3^n) = \sum_{\tilde{V} \in \mathcal{R}(Q_{X_2})} \mathbb{P}\left( X_2^n \in \mathcal{T}_{\tilde{V}}(y_3^n) \right),
\]
(121)

where \(X_2^n\) is uniformly distributed over the type class \(\mathcal{T}_{Q_{X_2}}\). From Lemma 3 (with the identifications \(X_1 \rightarrow \emptyset\), \(V \leftarrow Q_{X_2}\), \(V' \leftarrow P_{y_3^n}^\tilde{V}\), \(W \leftarrow \tilde{V}\)), we have that for every \(\tilde{V} \in \mathcal{R}(Q_{X_2})\),
\[
\mathbb{P}\left( X_2^n \in \mathcal{T}_{\tilde{V}}(y_3^n) \right) \leq \exp(-nI(P_{y_3^n}^\tilde{V}, \tilde{V})).
\]
(122)

Hence using the clause in (121) yields
\[
\tau_n(y_3^n) \leq \exp(-nI(x_2^n \wedge y_3^n)).
\]
(123)

Substituting this into the the bound for \(\mu_n(y_3^n)\) in (119) yields
\[
\mu_n(y_3^n) \leq \exp\left[ -n(\tilde{I}(x_2^n \wedge y_3^n) - R_2 \mid^+) \right].
\]
(124)

Plugging this back into the bound for \(\beta_n\) in (118) yields
\[
\begin{align*}
\beta_n & \leq \sum_{y_3^n} W_{Q_{X_1}}^n(y_3^n | x_2^n) \exp\left[ -n(\tilde{I}(x_2^n \wedge y_3^n) - R_2 \mid^+) \right] \\
& = \sum_{\tilde{V} \in \mathcal{R}(Q_{X_2})} W_{Q_{X_1}}^n(\mathcal{T}_{\tilde{V}}(x_2^n) | x_2^n) \exp\left[ -nI(Q_{X_2}, \tilde{V}) - R_2 \mid^+ \right] \\
& \leq \sum_{\tilde{V} \in \mathcal{R}(Q_{X_2})} \exp\left[ -n(D(V\|W_{Q_{X_1}} | Q_{X_2}) + I(Q_{X_2}, \tilde{V}) - R_2 \mid^+) \right].
\end{align*}
\]
(125)
(126)
(127)

This gives the exponent \(G_1(R, R_2)\) in (76) upon minimizing over all \(\tilde{V} \in \mathcal{R}(Q_{X_2})\).
Second Packing Error $\epsilon_{D,2}$: We evaluate $\epsilon_{D,2}$ by partitioning the sample space into subsets where the conditional type of relay input $y^n_2$ given relay output $x^n_2$ is $Q_{Y_2|X_2} \in \mathcal{V}(Y_2; Q_{X_2})$. That is,

$$
\epsilon_{D,2} = \sum_{Q_{Y_2|X_2} \in \mathcal{V}(Y_2; Q_{X_2})} \mathbb{P}(Y^n_2 \in \mathcal{T}_{Q_{Y_2|X_2}}(X^n_2)) \varphi_n(Q_{Y_2|X_2})
$$

(128)

where $\varphi_n(Q_{Y_2|X_2})$ is defined as

$$
\varphi_n(Q_{Y_2|X_2}) := \mathbb{P}\left(\hat{M}_j \neq 1 \mid L_j, L_{j-1} \text{ decoded correctly}, Y^n_2 \in \mathcal{T}_{Q_{Y_2|X_2}}(X^n_2)\right).
$$

(129)

We bound the probability in (128) and $\varphi_n(Q_{Y_2|X_2})$ in the following. Then we optimize over all conditional types $Q_{Y_2|X_2} \in \mathcal{V}(Y_2; Q_{X_2})$. This corresponds to the minimization in (77).

The probability in (128) can be first bounded using the same steps in (115) to (118) as

$$
\mathbb{P}(Y^n_2 \in \mathcal{T}_{Q_{Y_2|X_2}}(X^n_2)) \leq (n + 1)^{|X_1|} \sum_{x^n_2 \in \mathcal{T}_{Q_{X_2}}} \frac{1}{|Q_{X_2}|} \sum_{y^n_2 \in \mathcal{T}_{Q_{Y_2|X_2}}(x^n_2)} W^n_{Q_{X_1}}(y^n_2 | x^n_2)
$$

(130)

Now by using Lemma [1]

$$
\mathbb{P}(Y^n_2 \in \mathcal{T}_{Q_{Y_2|X_2}}(X^n_2)) \leq \exp\left[-nD(Q_{Y_2|X_2} \parallel W_{Q_{X_1}} | Q_{X_2})\right].
$$

(131)

This gives the first part of the exponent $G_2(R, R_2)$ in (77).

Recall the notations $\mathcal{P}_n(Q_{X_1}, Q_{X_2}, Q_{Y_2|X_2})$ and $\mathcal{X}_n(Q_{Y_2|X_2}, Q_{Y_2|Y_2,X_2})$ in (72) and (74) respectively. These sets will be used in the subsequent calculation. Implicit in the calculations below is the fact that $Y^n_2 \in \mathcal{T}_{Q_{Y_2|X_2}}(X^n_2)$ and also for the fixed $Q_{Y_2|X_2}$ we have the fixed test channel $Q_{Y_2|X_2}(Y_2|X_2)$ which we will denote more succinctly as $Q_{Y_2|Y_2,X_2}$. See (77) and the code generation. In the following steps, we simply use the notation $\mathcal{T}_{P_{X_1,X_2,Y_2,Y_3}}$ as an abbreviation for the event that the random quadruple of sequences $(X^n_1(1), X^n_2(L_{j-1}), Y^n_2(K_j | L_{j-1}), Y^n_3(j))$ belongs to the type class $\mathcal{T}_{P_{X_1,X_2,Y_2,Y_3}}$. Following the one-at-a-time union bound strategy in [19] (see [19]), we now bound $\varphi_n(Q_{Y_2|X_2})$ in (129) by conditioning on various joint types $P_{X_1,X_2,Y_2,Y_3} \in \mathcal{P}_n(Q_{X_1}, Q_{X_2}, Q_{Y_2|X_2})$

$$
\varphi_n(Q_{Y_2|X_2}) \leq \sum_{P_{X_1,X_2,Y_2,Y_3} \in \mathcal{P}_n(Q_{X_1}, Q_{X_2}, Q_{Y_2|X_2})} \mathbb{P}\left(\mathcal{T}_{P_{X_1,X_2,Y_2,Y_3}}\right) \mathbb{P}\left(\bigcup_{V \in \mathcal{X}_n(Q_{Y_2|X_2}, Q_{Y_2|Y_2,X_2})} \mathcal{E}_V \middle| \mathcal{T}_{P_{X_1,X_2,Y_2,Y_3}}\right),
$$

(132)

where $Q_{Y_2|X_2}$ is specified in (68) based on $Q_{Y_2|Y_2,X_2}$ and $Q_{Y_2|X_2}$ and the event $\mathcal{E}_V$ is defined as

$$
\mathcal{E}_V := \bigcup_{\tilde{Q}_{Y_2|X_2} \in \mathcal{V}(Y_2; Q_{X_2})} \mathcal{E}_V(\tilde{Q}_{Y_2|X_2}),
$$

(133)

with the constituent events defined as

$$
\mathcal{E}_V(\tilde{Q}_{Y_2|X_2}) := \bigcup_{\tilde{m}_j \in \exp(nR)} \bigcup_{\tilde{k}_j \in B_{L_j}(\tilde{Q}_{Y_2|X_2}, \tilde{L}_{j-1})} \mathcal{E}_V(\tilde{Q}_{Y_2|X_2}, \tilde{m}_j, \tilde{k}_j),
$$

(134)

and

$$
\mathcal{E}_V(\tilde{Q}_{Y_2|X_2}, \tilde{m}_j, \tilde{k}_j) := \left\{ (X^n_1(\tilde{m}_j), X^n_2(L_{j-1}), \tilde{Y}_2^n(\tilde{k}_j | L_{j-1}), Y^n_3(j)) \in \mathcal{T}_{Q_{X_1}Q_{X_2}Q_{Y_2|X_2}} \right\}.
$$

(135)

Recall the definition of $\tilde{Q}_{Y_2|X_2}$ in (69). This is a function of the decoded conditional type $\tilde{Q}_{Y_2|X_2}$. Note that $\tilde{Q}_{Y_2|X_2}$ indexes a decoded conditional type (of which there are only polynomially many), $\tilde{m}_j$ indexes an incorrect decoded message and $\tilde{k}_j$ indexes a correctly ($\tilde{k}_j = K_j$) or incorrectly decoded bin index ($\tilde{k}_j \neq K_j$). The union over $\tilde{k}_j$ extends over the entire bin $B_{L_j}(\tilde{Q}_{Y_2|X_2}, \tilde{L}_{j-1})$ and not only over incorrect bin indices. This is because an error is declared only if $\tilde{m}_j \neq 1$. Essentially, in the crucial step in (132), we have conditioned on the channel behavior (from $X_1 \times X_2$ to $\tilde{Y}_2 \times \tilde{Y}_3$) and identified the set of conditional types (indexed by $V$) that leads to an error based on the $\alpha$-decoding step in (95).

Now we will bound the constituent elements in (132). Recall that $X^n_1$ is drawn uniformly at random from $\mathcal{T}_{Q_{X_1}}$, $X^n_2$ is drawn uniformly at random from $\mathcal{T}_{Q_{X_2}}$ and $Y^n_2$ is drawn uniformly at random from $\mathcal{T}_{Q_{Y_2|X_2,Y_2}}(y^n_2 | x^n_2)$ where
\( y_2^n \in T_{Q_{Y_2|X_2}}(x_2^n) \) and \( Q_{Y_2|X_2} \) is the conditional type fixed in (128). Note that if we are given that \( Y_2^n \in T_{Q_{Y_2|X_2}}(x_2^n) \), it must be uniformly distributed in \( T_{Q_{Y_2|X_2}}(x_2^n) \) given \( X_2^n = x_2^n \). Finally, \( Y_3^n \) is drawn from the relay channel \( W^n(\cdot|y_3^n,x_1^n,x_2^n) \). Using these facts, we can establish that the first probability in (132) can be expressed as

\[
P \left( T_{P_{X_1|X_2}Z_{Y_3}} \right) = \sum_{x_1^n \in T_{Q_{X_1}}} \sum_{x_2^n \in T_{Q_{X_2}}} \frac{\partial(x_1^n,x_2^n)}{|T_{Q_{X_1}}||T_{Q_{X_2}}|},
\]

where

\[
\partial(x_1^n,x_2^n) := \sum_{y_2^n \in T_{Q_{Y_2|X_2}}(x_2^n)} \frac{1}{|T_{Q_{Y_2|X_2}}(x_2^n)|} \sum_{(\tilde{y}_2^n,y_2^n) \in T_{P_{Y_2|Y_3}|X_1|X_2}(x_1^n,x_2^n)} \frac{W^n(y_2^n,y_2^n,x_1^n,x_2^n)}{|T_{Q_{Y_2|X_2}}(x_2^n)|}.
\]

Notice that the term \(|T_{Q_{Y_2|X_2}}(x_2^n)|^{-1} \{ y_2^n \in T_{Q_{Y_2|X_2}}(x_2^n) \} \) in (137) indicates that \( Y_2^n \) is uniformly distributed in \( T_{Q_{Y_2|X_2}}(x_2^n) \) given \( X_2^n = x_2^n \). We now bound \( \partial(x_1^n,x_2^n) \) by the same logic as the steps from (115) to (118) (relating size of shells to probabilities of sequences). More precisely,

\[
\partial(x_1^n,x_2^n) \leq (n+1)^{|X_2||Y_3|^2} \sum_{y_2^n \in T_{Q_{Y_2|X_2}}} Q_{Y_2|X_2}(\tilde{y}_2^n,y_2^n,x_1^n,x_2^n) \times \sum_{(\tilde{y}_2^n,y_2^n) \in T_{P_{Y_2|Y_3}|X_1|X_2}(x_1^n,x_2^n)} W^n(y_2^n,y_2^n,x_1^n,x_2^n) \leq \sum_{(\tilde{y}_2^n,y_2^n) \in T_{P_{Y_2|Y_3}|X_1|X_2}(x_1^n,x_2^n)} W^n_{Q_{Y_2|X_2},Q_{Y_2|Y_3|X_2}}(\tilde{y}_2^n,y_2^n,x_1^n,x_2^n)
\]

\[
= \exp \left[ -nD(P_{Y_2|X_2,Z_{Y_3}} \| W_{Q_{Y_2|X_2},Q_{Y_2|Y_3|X_2}} | Q_{X_1}Q_{X_2} \right] \]
This corresponds to the case involving $\psi_1$ in (79).

For the other case (i.e., $\psi_2$ in (80)) where both the message and bin index are incorrect ($\tilde{m}_j \neq 1$ and $\tilde{k}_j \neq K_j$), slightly more intricate analysis is required. For any conditional type $\hat{Q}_{Y_2|X_2}$, define the excess Wyner-Ziv rate given $\hat{Q}_{Y_2|X_2}$ as

$$\Delta R_2(\hat{Q}_{Y_2|X_2}) := \hat{R}_2(\hat{Q}_{Y_2|X_2}) - R_2$$

(146)

where the inflated rate $\hat{R}_2(\hat{Q}_{Y_2|X_2})$ is defined in (91). This is exactly (82) but we make the dependence on $\hat{Q}_{Y_2|X_2}$ explicit in (146). Assume for the moment that $\Delta R_2(\hat{Q}_{Y_2|X_2}) \geq 0$. Equivalently, this means that $R_2 \leq I(\hat{Q}_{Y_2|X_2}, Q_{Y_2|Y_2|X_2}) + \nu_n$, which is, up to the $\nu_n \in \Theta(\log_2 n)$ term, the first clause in (80). Again, using bars to denote random variables generated uniformly from their respective marginal type classes and arbitrary sequences in their respective marginal type classes, define as in (19)

$$\xi_n(V, \hat{Q}_{Y_2|X_2}) := \exp(n \Delta R_2(\hat{Q}_{Y_2|X_2})) \cdot \mathbb{P}\left[ \tilde{x}_2^n, \tilde{y}_2^n, \tilde{y}_3^n \in T_{Q_2X_2}(\hat{Q}_{Y_2|X_2}) \right]$$

(147)

$$\zeta_n(V, \hat{Q}_{Y_2|X_2}) := \exp(n R) \cdot \mathbb{P}\left[ \tilde{x}_2^n, \tilde{y}_2^n, \tilde{x}_3^n, \tilde{y}_3^n \in T_{Q_2X_2}(\hat{Q}_{Y_2|X_2} \times V) \right]$$

(148)

where the conditional distributions $V_{Q_{X_2}}$ and $\hat{Q}_{Y_2|X_2} \times V$ are defined in (70) and (71) respectively. By applying the one-at-a-time union to the two unions in (134) as was done in (19) (see (19) in the Introduction), we obtain

$$\mathbb{P}\left[ \mathcal{E}_V(\hat{Q}_{Y_2|X_2}) \right] \leq \gamma_n(V, \hat{Q}_{Y_2|X_2})$$

(149)

where

$$\gamma_n(V, \hat{Q}_{Y_2|X_2}) := \min \{ 1, \xi_n(V, \hat{Q}_{Y_2|X_2}) \cdot \min \{ 1, \zeta_n(V, \hat{Q}_{Y_2|X_2}) \} \}.$$  

(150)

Now by using the same reasoning that led to (145) (i.e., Lemma 2), we see that (147) and (148) evaluate to

$$\xi_n(V, \hat{Q}_{Y_2|X_2}) \approx \exp \left[ -n \left( I(\hat{Q}_{Y_2|X_2}, V_{Q_{X_2}}|Q_{X_2}) - \Delta R_2(\hat{Q}_{Y_2|X_2}) \right) \right]$$

(151)

$$\zeta_n(V, \hat{Q}_{Y_2|X_2}) \approx \exp \left[ -n \left( I(Q_{X_1}, \hat{Q}_{Y_2|X_2} \times V|Q_{X_2}) - R \right) \right]$$

(152)

Hence, $\gamma_n(V, \hat{Q}_{Y_2|X_2})$ in (150) has the following exponential behavior:

$$\gamma_n(V, \hat{Q}_{Y_2|X_2}) \approx \exp \left[ -n \left( I(\hat{Q}_{Y_2|X_2}, V_{Q_{X_2}}|Q_{X_2}) - \Delta R_2(\hat{Q}_{Y_2|X_2}) \right) + \left| I(Q_{X_1}, \hat{Q}_{Y_2|X_2} \times V|Q_{X_2}) - R \right| \right]$$

(153)

Note that we can swap the order of the union bounds in the bounding of the probability in (149). As such, the probability in of the event $\mathcal{E}_V(\hat{Q}_{Y_2|X_2})$ can also be upper bounded by

$$\gamma'_n(V, \hat{Q}_{Y_2|X_2}) := \min \{ 1, \zeta_n(V, \hat{Q}_{Y_2|X_2}) \cdot \min \{ 1, \xi_n(V, \hat{Q}_{Y_2|X_2}) \} \},$$

(154)

which, in view of (151) and (152), has the exponential behavior

$$\gamma'_n(V, \hat{Q}_{Y_2|X_2}) \approx \exp \left[ -n \left( I(Q_{X_1}, \hat{Q}_{Y_2|X_2} \times V|Q_{X_2}) - R + I(\hat{Q}_{Y_2|X_2}, V_{Q_{X_2}}|Q_{X_2}) - \Delta R_2(\hat{Q}_{Y_2|X_2}) \right) \right].$$

(155)

Compare and contrast (155) to (153).

Now consider $\Delta R_2(\hat{Q}_{Y_2|X_2}) < 0$. Equivalently, this means that $R_2 > I(\hat{Q}_{Y_2|X_2}, Q_{Y_2|Y_2|X_2}) + \nu_n$, which is, up to the $\nu_n \in \Theta(\log_2 n)$ term in (82), the second clause in (80). In this case, we simply upper bound $\exp(n \Delta R_2(\hat{Q}_{Y_2|X_2}))$ by unity and hence, $\xi_n(V, \hat{Q}_{Y_2|X_2})$ as

$$\xi_n(V, \hat{Q}_{Y_2|X_2}) \leq \exp \left[ -n I(\hat{Q}_{Y_2|X_2}, V_{Q_{X_2}}|Q_{X_2}) \right]$$

(156)

and this yields

$$\gamma_n(V, \hat{Q}_{Y_2|X_2}) \leq \exp \left[ -n \left( I(\hat{Q}_{Y_2|X_2}, V_{Q_{X_2}}|Q_{X_2}) + \left| I(Q_{X_1}, \hat{Q}_{Y_2|X_2} \times V|Q_{X_2}) - R \right| \right) \right].$$

(157)

Uniting the definition of $\hat{R}_2(\hat{Q}_{Y_2|X_2})$ in (91), the probabilities in (131) and (142), the case where only the message is incorrect in (145), the definition of the excess Wyner-Ziv rate $\Delta R_2(\hat{Q}_{Y_2|X_2})$ in (146) and the case where both message and bin index are incorrect in (153) and (157) yields the exponent $G_2(R, R_2)$ in (77) as desired.

Finally, we remark that the alternative exponent given in (85) comes from using (155) instead of (133).
V. AN UPPER BOUND FOR THE RELIABILITY FUNCTION

In this section, we state and prove an upper bound on the reliability function per Definition 4. This bound is inspired by Haroutunian’s exponent for channels with feedback [23]. Also see [8, Ex. 10.36]. The upper bound on the reliability function is stated in Section V-A; discussions are provided in Section V-B, and the proof is detailed in Section V-C.

A. The Upper Bound on the Reliability Function

Before we state the upper bound, define the function

\[ C_{cs}(V) \) := \max_{P_{X_1 X_2} \in \mathcal{P}(X_1 \times X_2)} \min \{ I(P_{X_1 X_2}, V_{Y_3 | X_1 X_2}), I(P_{X_1 | X_2}, V|P_{X_2}) \}, \]

(158)

where \( V \) represents a transition matrix from \( X_1 \times X_2 \) to \( Y_2 \times Y_3 \) and \( V_{Y_3 | X_1 X_2} \) is its \( Y_3 \)-marginal. We recognize that (158) is the cutset bound on all achievable rates for the DM-RC \( V \) (introduced in (3)) but written in a different form in which the distributions are explicit. Note that the subscript \( cs \) stands for “cutset.”

**Theorem 3** (Upper Bound on the Reliability Function). We have the following upper bound on the reliability function:

\[ E(R) \leq E_{cs}(R) := \min_{V: X_1 \times X_2 \rightarrow Y_2 \times Y_3} \max_{C_{cs}(V) \leq R} D(V\|W|P_{X_1 X_2}). \]

(159)

The proof of this theorem is provided in Section V-C.

B. Remarks on the Upper Bound on the Reliability Function

1) Clearly, the upper bound \( E_{cs}(R) \) is positive if and only if \( R < C_{cs}(W) \). Furthermore, because \( P_{X_1 X_2} \to D(V\|W|P_{X_1 X_2}) \) is linear, the maximum is achieved at a particular symbol pair \((x_1, x_2)\), i.e.,

\[ E_{cs}(R) := \min_{V, X_1 \times X_2 \rightarrow Y_2 \times Y_3} \max_{x_1, x_2} D(V(\cdot, | x_1, x_2)\|W(\cdot, | x_1, x_2)). \]

(160)

The computation of the \( E_{cs}(R) \) appears to be less challenging than CF but is still difficult because \( C_{cs}(V) \) is not convex in general and so \( E_{cs}(R) \) in (159) or (160) are not convex optimization problems. Finding the joint distribution \( P_{X_1 X_2} \) that achieves \( C_{cs}(V) \) for any \( V \) is also not straightforward as one needs to develop an extension of the Blahut-Arimoto algorithm [8, Ch. 8]. We do not explore this further as developing efficient numerical algorithms is not the focus of the current work.

2) We expect that, even though the cutoff rate (rate at which the exponent transitions from being positive to being zero) is the cutset bound, the upper bound we derived in (159) is quite loose relative to the achievable bounds prescribed by Theorems 1 and 2. This is because the achievability theorems leverage on block-Markov coding and hence the achievable exponents are attenuated by the number of blocks \( b \) causing significant delay when \( b \) is large. (See Section VI for a numerical example.) This factor is not present in Theorem 3.

3) One method to strengthen the exponent is to consider a more restrictive class of codes, namely codes with finite memory. For this class of codes, there exists some integer \( l \geq 1 \) (that does not depend on \( n \)) such that \( g_i(y_{2,i-1}) = g_i(y_{2,i-1}^{l-1}) \) for all \( i \in [n] \). Under a similar assumption for the discrete memoryless channel (DMC), Como and Nakiboğlu [43] showed that the sphere-packing bound [25] is an upper bound for DMCs with feedback, thus improving on Haroutunian’s original result [23]. In our setting, this would mean that

\[ \bar{E}_{cs}(R) := \max_{P_{X_1 X_2} \in \mathcal{P}(X_1 \times X_2)} \min_{V: X_1 \times X_2 \rightarrow Y_2 \times Y_3} \frac{D(V\|W|P_{X_1 X_2})}{\min\{I(P_{X_1 X_2}, V_{Y_3 | X_1 X_2}), I(P_{X_1 | X_2}, V|P_{X_2})\}} \leq R \]

(161)

is also an upper bound on the reliability function. This bound, reminiscent of the sphere-packing exponent [25], is, in general, tighter (smaller) than (159) for general DM-RCs \( W \). We defer this extension to future work.

4) To prove an upper bound on the reliability function for channel coding problems, many authors first demonstrate a strong converse. See the proof that the sphere-packing exponent is an upper bound for the reliability function of a DMC without feedback in Csiszár-Körner [8, Thm. 10.3]; the proof of the sphere-packing
exponent for asymmetric broadcast channels in [15] Thm. 1(b) or Eq. (12)]; and the proof of the upper bound of the reliability function for Gel’fand-Pinsker coding [44] Thms. 2 and 3] for example. Haroutunian’s original proof of the former does not require the strong converse though [25, Eq. (26)]. For relay channels and relay networks, the strong converse above the cutset bound was recently proved by Behboodi and Piantanida [45], [46] using information spectrum methods [47] but we do not need the strong converse for the proof of Theorem 3. Instead we leverage on a more straightforward change-of-measure technique by Palaiyanur [24].

C. Proof of Theorem 3

Proof: Fix $\delta > 0$ and let a given DM-RC $V : \mathcal{X}_1 \times \mathcal{X}_2 \rightarrow \mathcal{Y}_2 \times \mathcal{Y}_3$ be such that $C_{cs}(V) \leq R - \delta$. Since the rate $R$ is larger than the cutset bound $C_{cs}(V)$, by the weak converse for DM-RCs [8] Thm. 16.1), the average error probability assuming the DM-RC is $V$ (defined in Definition 3) is bounded away from zero, i.e.,

$$P_e(V) \geq \eta,$$  \hspace{1cm} (162)

for some $\eta > 0$ that depends only on $R$. Because the signal $Y^i_{2-1}$ is provided to the relay encoder for each time $i \in [n]$ and we do not have a strong converse statement in (162), we cannot apply the change-of-measure technique in Csiszár-Körner [8] Thm. 10.3. We instead follow the proof strategy proposed in Palaiyanur’s thesis [24] Lem. 18] for channels with feedback.

First, an additional bit of notation: For any message $m \in \mathcal{M}$, joint input type $P \in \mathcal{P}_n(\mathcal{X}_1 \times \mathcal{X}_2)$, conditional type $U \in \mathcal{Y}_n(\mathcal{Y}_2 \times \mathcal{Y}_3; P)$ and code $(f, g^n, \varphi)$, define the relay shell as follows

$$A(m, P, U) := \{(y^n_2, y^n_3) : (f(m), g^n(y^n_2)) \in T_P, (y^n_2, y^n_3) \in T_U : (f(m), g^n(y^n_2))\}.$$  \hspace{1cm} (163)

Note that $(f(m), g^n(y^n_2)) \in \mathcal{X}_1^n \times \mathcal{X}_2^n$ can be regarded as the channel input when the input to the relay node is $y^n_2$. So $A(m, P, U)$ is the set of all $(y^n_2, y^n_3)$ that lie in the $U$-shell of the channel inputs $(x^n_1, x^n_2)$ which are of joint type $P$ and these channel inputs result from sending message $m$.

From the definition of $P_e(W)$ in [9], we have

$$P_e(W) = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \sum_{P \in \mathcal{P}_n(\mathcal{X}_1 \times \mathcal{X}_2)} \sum_{U \in \mathcal{Y}_n(\mathcal{Y}_2 \times \mathcal{Y}_3; P)} \sum_{y^n_2, y^n_3 \in A(m, P, U)} \mathbb{P}_W((Y^n_2, Y^n_3) = (y^n_2, y^n_3) | M = m),$$  \hspace{1cm} (164)

where we partitioned $\mathcal{X}_1^n \times \mathcal{X}_2^n$ into sequences of the same type $P$ and we partitioned $\mathcal{Y}_2^n \times \mathcal{Y}_3^n$ into conditional types $U$ compatible with $P$. Now, we change the measure in the inner probability to $V$ as follows:

$$P_e(W) = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \sum_{P \in \mathcal{P}_n(\mathcal{X}_1 \times \mathcal{X}_2)} \sum_{U \in \mathcal{Y}_n(\mathcal{Y}_2 \times \mathcal{Y}_3; P)} \sum_{y^n_2, y^n_3 \in A(m, P, U)} \mathbb{P}_V((Y^n_2, Y^n_3) = (y^n_2, y^n_3) | M = m)$$  \hspace{1cm} (165)

$$= \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \sum_{P \in \mathcal{P}_n(\mathcal{X}_1 \times \mathcal{X}_2)} \sum_{U \in \mathcal{Y}_n(\mathcal{Y}_2 \times \mathcal{Y}_3; P)} \sum_{y^n_2, y^n_3 \in A(m, P, U)} \mathbb{P}_V((Y^n_2, Y^n_3) = (y^n_2, y^n_3) | M = m)$$  \hspace{1cm} (166)

where (165) is the key step in this whole proof where we changed the conditional measure (channel) from $W$ to
$V$ and (166) follows from the definition of the set $\mathcal{A}(m, P, U)$ in (163). Continuing, we have

$$\frac{P_e(W)}{P_e(V)} = \frac{1}{|M|} \sum_{m \in M} \sum_{P \in \mathcal{P}(\mathcal{X}_1 \times \mathcal{X}_2)} \sum_{U \in \mathcal{Y} \mathcal{U}_n(\mathcal{Y}_2 \times \mathcal{Y}_3; P)} \frac{P_V((Y_2^n, Y_3^n) = (y_2^n, y_3^n)|M = m)}{P_e(V)}$$

$$\times \exp \left( -n \sum_{x_1, x_2, y_2, y_3} P(x_1, x_2)U(y_2, y_3|x_1, x_2) \log \frac{V(y_2, y_3|x_1, x_2)}{W(y_2, y_3|x_1, x_2)} \right)$$

$$\geq \exp \left[ -n \left( \frac{1}{|M|} \sum_{m \in M} \sum_{P \in \mathcal{P}(\mathcal{X}_1 \times \mathcal{X}_2)} \sum_{U \in \mathcal{Y} \mathcal{U}_n(\mathcal{Y}_2 \times \mathcal{Y}_3; P)} \frac{P_V((Y_2^n, Y_3^n) = (y_2^n, y_3^n)|M = m)}{P_e(V)} \right) \right. \sum_{x_1, x_2, y_2, y_3} P(x_1, x_2)U(y_2, y_3|x_1, x_2) \log \frac{V(y_2, y_3|x_1, x_2)}{W(y_2, y_3|x_1, x_2)} \right].$$

The last step (168) follows from the convexity of $t \mapsto \exp(-t)$. Now the idea is to approximate $U$ with $V$. For this purpose, define the following “typical” set

$$\mathcal{G}_\gamma(V) := \left\{ (P, U) \in \mathcal{P}(\mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{Y}_2 \times \mathcal{Y}_3) : \sum_{x_1, x_2, y_2, y_3} P(x_1, x_2)U(y_2, y_3|x_1, x_2) - V(y_2, y_3|x_1, x_2) \leq \gamma \right\}.$$ (169)

Also define the finite constant

$$\kappa_V := \max_{x_1, x_2, y_2, y_3} -\log V(y_2, y_3|x_1, x_2).$$ (170)

For $(P, U) \in \mathcal{G}_\gamma(V)$, it can be verified by using the definition of $D(V||W|P)$ [24 Prop. 13] that

$$\sum_{x_1, x_2, y_2, y_3} P(x_1, x_2)U(y_2, y_3|x_1, x_2) \log \frac{V(y_2, y_3|x_1, x_2)}{W(y_2, y_3|x_1, x_2)} \leq D(V||W|P) + \gamma \max\{\kappa_V, \kappa_W\}.$$ (171)

For the typical part of the exponent in (168), we have

$$T := \frac{1}{|M|} \sum_{m \in M} \sum_{P \in \mathcal{P}(\mathcal{X}_1 \times \mathcal{X}_2)} \sum_{U \in \mathcal{Y} \mathcal{U}_n(\mathcal{Y}_2 \times \mathcal{Y}_3; P)} \frac{P_V((Y_2^n, Y_3^n) = (y_2^n, y_3^n)|M = m)}{P_e(V)}$$

$$\times \sum_{x_1, x_2, y_2, y_3} P(x_1, x_2)U(y_2, y_3|x_1, x_2) \log \frac{V(y_2, y_3|x_1, x_2)}{W(y_2, y_3|x_1, x_2)}$$

$$\leq \max_P D(V||W|P) + \gamma \max\{\kappa_V, \kappa_W\}$$

$$\times \frac{1}{|M|} \sum_{m \in M} \sum_{P \in \mathcal{P}(\mathcal{X}_1 \times \mathcal{X}_2)} \sum_{U \in \mathcal{Y} \mathcal{U}_n(\mathcal{Y}_2 \times \mathcal{Y}_3; P)} \frac{P_V((Y_2^n, Y_3^n) = (y_2^n, y_3^n)|M = m)}{P_e(V)}$$

$$\times \sum_{x_1, x_2, y_2, y_3} P(x_1, x_2)U(y_2, y_3|x_1, x_2) \log \frac{V(y_2, y_3|x_1, x_2)}{W(y_2, y_3|x_1, x_2)}$$

(172)

(173)

Now we drop the condition $(P, U) \in \mathcal{G}_\gamma(V)$ and continue bounding $T$ as follows

$$T \leq \max_P D(V||W|P) + \gamma \max\{\kappa_V, \kappa_W\}$$

$$\times \frac{1}{|M|} \sum_{m \in M} \sum_{P \in \mathcal{P}(\mathcal{X}_1 \times \mathcal{X}_2)} \sum_{U \in \mathcal{Y} \mathcal{U}_n(\mathcal{Y}_2 \times \mathcal{Y}_3; P)} \frac{P_V((Y_2^n, Y_3^n) \in \mathcal{A}(m, P, U), Y_3^n \in \mathcal{D}_m|M = m)}{P_e(V)}$$

(174)

(175)

$$= \max_P D(V||W|P) + \gamma \max\{\kappa_V, \kappa_W\} \frac{1}{|M|} \sum_{m \in M} \frac{P_V(Y_3^n \in \mathcal{D}_m|M = m)}{P_e(V)}$$

(176)
The last step follows from the definition of the average error probability $P_e(V)$ in (9).
For $(P, U) \not\in \mathcal{G}(V)$, by Pinsker’s inequality [8, Ex. 3.18] and Jensen’s inequality (see [24, Lem. 19]),
\begin{equation}
D(U||V|P) \geq \mathbb{E}_P \left[ \frac{1}{2 \ln 2} \|U(\cdot, \cdot|X_1, X_2) - V(\cdot, \cdot|X_1, X_2)\|^2 \right]
\end{equation}
\begin{equation}
\geq \frac{1}{2 \ln 2} \mathbb{E}_P \left[ \|U(\cdot, \cdot|X_1, X_2) - V(\cdot, \cdot|X_1, X_2)\|^2 \right] \geq \frac{\gamma^2}{2 \ln 2}.
\end{equation}
Furthermore, for any $(y^n_2, y^n_3) \in \mathcal{A}(m, P, U)$,
\begin{equation}
\log \mathbb{P}_V((Y^n_2, Y^n_3) = (y^n_2, y^n_3)|M = m)
= \sum_{i=1}^{n} \log V(y_{2i}, y_{3i}|f_i(m), g_i(y_2^{i-1}))
\end{equation}
\begin{equation}
= n \sum_{x_1, x_2, y_2, y_3} \left( \frac{1}{n} \sum_{i=1}^{n} 1\{f_i(m) = x_1, g_i(y_2^{i-1}) = x_2, y_{2i} = y_2, y_{3i} = y_3 \} \right) \log V(y_2, y_3|x_1, x_2)
\end{equation}
\begin{equation}
= n \sum_{x_1, x_2, y_2, y_3} P(x_1, x_2) U(y_2, y_3|x_1, x_2) \log V(y_2, y_3|x_1, x_2)
\end{equation}
\begin{equation}
= -n(D(U||V|P) + H(U|P)),
\end{equation}
where (181) follows from the definition of the set $\mathcal{A}(m, P, U)$. So in a similar manner as in [24, Prop. 14(c)], we have $|\mathcal{A}(m, P, U)| \leq \exp(nH(U|P))$. Thus,
\begin{equation}
\sum_{P \in \mathcal{P}_n(X_1 \times X_2)} \sum_{U \in \mathcal{U}_n(Y_2 \times Y_3; P)} \sum_{(P, U) \not\in \mathcal{G}(V)} \mathbb{P}_V((Y^n_2, Y^n_3) = (y^n_2, y^n_3)|M = m)
\end{equation}
\begin{equation}
= \sum_{P \in \mathcal{P}_n(X_1 \times X_2)} \sum_{U \in \mathcal{U}_n(Y_2 \times Y_3; P)} \mathbb{P}_V((Y^n_2, Y^n_3) \in \mathcal{A}(m, P, U)|M = m) \exp[-n(D(U||V|P) + H(U|P))]
\end{equation}
\begin{equation}
\leq \sum_{P \in \mathcal{P}_n(X_1 \times X_2)} \sum_{U \in \mathcal{U}_n(Y_2 \times Y_3; P)} \mathbb{P}_V((Y^n_2, Y^n_3) \in \mathcal{A}(m, P, U)|M = m) \exp\left(-n \frac{\gamma^2}{2 \ln 2}\right),
\end{equation}
where the final step follows from (178) and Lemma [1]. As a result, for the atypical part of the exponent in (168),
\begin{equation}
S := \frac{1}{|M|} \sum_{m \in \mathcal{M}} \sum_{P \in \mathcal{P}_n(X_1 \times X_2)} \sum_{(y^n_2, y^n_3) \in \mathcal{A}(m, P, U)} \mathbb{P}_V((Y^n_2, Y^n_3) = (y^n_2, y^n_3)|M = m)
\times \sum_{x_1, x_2, y_2, y_3} P(x_1, x_2) U(y_2, y_3|x_1, x_2) \log \frac{V(y_2, y_3|x_1, x_2)}{W(y_2, y_3|x_1, x_2)}
\leq \frac{\kappa_W}{P_e(V)} \cdot \frac{1}{|M|} \sum_{m \in \mathcal{M}} \sum_{P \in \mathcal{P}_n(X_1 \times X_2)} \mathbb{P}_V((Y^n_2, Y^n_3) \in \mathcal{A}(m, P, U)|M = m)
\end{equation}
\begin{equation}
\leq \frac{\kappa_W}{P_e(V)} (n + 1)|X_1||X_2||Y_2||Y_3| \exp\left(-n \frac{\gamma^2}{2 \ln 2}\right).
\end{equation}
In (187) we upper bounded $V(y_2, y_3|x_1, x_2)$ by 1 and $-\log W(y_2, y_3|x_1, x_2)$ by $\kappa_W$ for those $(x_1, x_2, y_2, y_3)$ such that $W(y_2, y_3|x_1, x_2) > 0$. If instead $W(y_2, y_3|x_1, x_2) = 0$ and $U(y_2, y_3|x_1, x_2) > 0$ for some $(y^n_2, y^n_3)$ in the sum.
in $S$, the probability $P_V((Y_2^n, Y_3^n) = (y_2^n, y_3^n)|M = m) = 0$ so these symbols can be ignored. Combining (168), (176) and (188), we conclude that

$$\frac{P_e(W)}{P_e(V)} \geq \exp[-n(T + S)] \geq \exp \left[-n \left(\max_P D(V||W) + \varrho_{n,\gamma}\right)\right]$$  \hspace{1cm} (189)

where

$$\varrho_{n,\gamma} := \gamma \max\{\kappa_V, \kappa_W\} + \frac{\kappa_W}{P_e(V)} (n + 1)|X_1||X_2||Y_2||Y_3| \exp \left(-n \frac{\gamma^2}{2\ln 2}\right).$$  \hspace{1cm} (190)

Now assume that the DM-RC $V$ is chosen to achieve the minimum in (159) evaluated at $R - \delta$, i.e., $V \in \arg\min_{V:C_n(V) \leq R - \delta} \max_{P_{X_1,X_2}} D(V||W|P_{X_1,X_2})$. Then uniting (162) and (189) yields

$$P_e(W) \geq \eta \exp \left[-n (E_{cs}(R - \delta) + \varrho_{n,\gamma})\right].$$  \hspace{1cm} (191)

We then obtain

$$\frac{1}{n} \log \frac{1}{P_e(W)} \leq E_{cs}(R - \delta) + \varrho_{n,\gamma} - \log \eta - \log n.$$  \hspace{1cm} (192)

Let $\gamma = n^{-1/4}$ so $\varrho_{n,\gamma} \to 0$ as $n \to \infty$. Note also that $\eta > 0$. As such by taking limits,

$$\lim_{n \to \infty} \frac{1}{n} \log \frac{1}{P_e(W)} \leq E_{cs}(R - \delta).$$  \hspace{1cm} (193)

Since the left-hand-side does not depend on $\delta$, we may now take the limit of the right-hand-side as $\delta \to 0$ and use the continuity of $E_{cs}(R)$ (which follows the continuity of $V \mapsto C_{cs}(V)$ and $V \mapsto \max_{P_{X_1,X_2}} D(V||W|P_{X_1,X_2})$) and obtain

$$\lim_{n \to \infty} \frac{1}{n} \log \frac{1}{P_e(W)} \leq E_{cs}(R)$$  \hspace{1cm} (194)

as desired.

\section{VI. Numerical Evaluation for the Sato Relay Channel}

In this section we evaluate the error exponent for the PDF (or, in this case, decode-forward) scheme presented in Theorem 1 for a canonical DM-RC, namely the Sato relay channel [6]. This is a physically degraded relay channel in which $X_1 = Y_2 = Y_3 = \{0, 1, 2\}$, $X_2 = \{0, 1\}$, $Y_2 = X_1$ (deterministically) and the transition matrices from $(X_1, X_2)$ to $Y_3$ are

$$W_{Y_3|X_1,X_2}(y_3|x_1,0) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.5 & 0.5 \\ 0.5 & 0 & 0.5 \end{bmatrix},$$  \hspace{1cm} (195)

$$W_{Y_3|X_1,X_2}(y_3|x_1,1) = \begin{bmatrix} 0 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0 \\ 1 & 0 & 0 \end{bmatrix}. $$  \hspace{1cm} (196)

It is known that the capacity of the Sato relay channel is $C_{Sato} = 1.161878$ bits per channel use [3, Eg. 16.1] and the capacity-achieving input distribution is

$$Q_{X_1,X_2}^* (x_1,x_2) = \begin{bmatrix} p \\ q \\ q \\ q \end{bmatrix}$$  \hspace{1cm} (197)

where $p = 0.35431$ and $q = 0.072845$ [2, Table I]. The auxiliary random variable $U$ is set to $X_1$ in this case. It is easy to check that $I(X_1;X_2;Y_3) = I(X_1,Y_2;X_2) = C_{Sato}$ ($I(X_1,X_2;Y_3)$ and $I(X_1;Y_2|X_2)$ are the only two relevant mutual information terms in the DF lower bound and the cutset upper bound) when the distribution of $(X_1, X_2)$ is capacity-achieving, i.e., the random variables $(X_1, X_2)$ have distribution $Q_{X_1,X_2}^*$ in (197). We study the effect of the number of blocks in block-Markov coding in the following. We note that this is the first numerical study of error exponents on a discrete relay channel. Other numerical studies were done for continuous-alphabet relay channels, e.g., [9], [30–34].
We set the number of blocks $b \in \{10, 50, 100\}$. In view that the capacity is $C_{\text{Sato}} = 1.161878$, we let effective rate $R_{\text{eff}}$, defined in (11), be in the range $[1.00, 1.20]$ bits per channel use. The per-block rate can thus be computed as

$$R_b := \frac{b}{b-1} R_{\text{eff}}$$

(198)

where we regard $R_{\text{eff}}$ as fixed and constant and made the dependence of the per-block rate on $b$ explicit. Since error exponents are monotonically decreasing in the rate (if the rate is smaller than $C_{\text{Sato}}$),

$$b_1 < b_2 \Rightarrow F(R_{b_1}) < F(R_{b_2})$$

(199)

because $R_{b_1} > R_{b_2}$ and similarly for $G(R_b)$. We reiterate that $\tilde{G}(R'_b)$, defined in (27), is not relevant for the numerical evaluation for this example. Note that there is a tradeoff concerning $b$ here. If $b$ is small, the degradation of the exponent due to the effect in (199) is significant but we divide $F(R_b)$ by a smaller factor in (24). Conversely, if $b$ is larger, the degradation due to (199) is negligible but we divide by a larger number of blocks to obtain the overall exponent. We evaluated the relay exponent $F(R_b)$ and the decoder exponent $G(R_b)$ in their Gallager forms in (32) and (33) using the capacity-achieving input distribution in (197). In Fig. 2 we plot the exponents $F(R_b)$ and $G(R_b)$ divided by $b$ as functions of the effective rate $R_{\text{eff}}$. For each $R_{\text{eff}}$, we also optimized for the largest exponents over $b$ in both cases. We make a three observations concerning Fig. 2:

1) First, the relay exponent $F(R_b)$ dominates because it is uniformly smaller than the decoder exponent $G(R_b)$. Hence, the overall exponent for the Sato channel using decode-forward is the relay exponent (the scales on the vertical axes are different). Since we made this observation, we also evaluated the PDF exponent with a non-trivial $U$ (i.e., not equal to $X_1$) whose alphabet $|U|$ was allowed to be as large as 10, non-trivial split of $R$ into $R'$ and $R''$, while preserving $Q_{X_1 X_2}$ in (197) as the input distribution. This was done to possibly increase the overall (minimum) exponent in (32)–(34). However, from our numerical studies, it appears that there was no advantage in using PDF in the error exponents sense for the Sato relay channel.

2) Second, the cutoff rates can be seen to be $C_{\text{Sato}}$ bits per channel use in both plots and this can only be achieved by letting $b$ become large. This is consistent with block-Markov coding [3, Sec. 16.4.1] in which

Fig. 2. Plots of the relay and decoder exponents divided by the number of blocks $b$ against effective rate $R_{\text{eff}}$ for the Sato relay channel. The optimized exponents $R_{\text{eff}} \mapsto \max_b F(R_b)/b$ and $R_{\text{eff}} \mapsto \max_b G(R_b)/b$ are also shown.

![Fig. 2](image_url)
to achieve the DF lower bound asymptotically, we need to let $b$ tend to infinity in addition to letting the per-block blocklength $n$ tend to infinity.

3) Finally, observe that if one would like to operate at an effective rate slightly below $R_{\text{eff}} = 1.10$ bits per channel use (say), we would choose the number of blocks $b \in \{10, 50, 100\}$ to be a moderate 50 instead of the larger 100 so as to attain a larger overall exponent. This is because the overall exponent in (24) is the ratio of the relay exponent $F(R_b)$ and the number of blocks $b$ and dividing by $b$ degrades the overall exponent more than the effect in (199). This implies that to achieve a small error probability at a fixed rate, we should use a block-Markov scheme with a carefully chosen number of blocks not tending to infinity assuming the number of channel uses within each block, i.e., the per-block blocklength, is sufficiently large (so that pre-exponential factors are negligible). However, if we want to operate at rates close to capacity, naturally, we need a larger number of blocks otherwise the exponents are zero. For example if we use only 10 blocks, both exponents are 0 for all $R_{\text{eff}} > 1.05$ bits per channel use. The functions of $b \mapsto F(R_b)/b$ and $b \mapsto G(R_b)/b$ are illustrated in Fig. 3 for $R_{\text{eff}} \in \{1.00, 1.05, 1.10\}$.

VII. CONCLUSION AND FUTURE WORK

In this paper, we derived achievable error exponents for the DM-RC by carefully analyzing PDF and CF. One of the take-home messages is that to achieve the best error decay for a fixed rate, we need to choose a moderate number of blocks in block-Markov coding. We also derived an upper bound for the reliability function by appealing to Haroutunian’s techniques for channels with feedback.

It is the author’s hope that the present paper will precipitate research in exponential error bounds for noisy networks with relays. We discuss a few avenues for further research here.

1) Most importantly, it is imperative to develop alternate forms of or approximations to the CF exponent (Theorem 2) and the upper bound on the reliability function (Theorem 3) which are difficult to compute in their current forms.
It would also be useful to show that the error exponents we derived in achievable theorems are ensemble tight, i.e., given the random codebook, the error exponents cannot be improved by using other decoding rules. This establishes some form of optimality of our coding schemes and analyses. This was done for the point-to-point channel in [48] and mismatched decoding in [49].

3) The results contained herein hinge on the method of types which is well-suited to analyze discrete-alphabet systems. It is also essential, for real-world wireless communication networks, to develop parallels of the main results for continuous-alphabet systems such as the AWGN relay channel. While there has been some work on this in [9], [30], [51], schemes such as CF have remained relatively unexplored. However, one needs a Marton-like exponent [35] for lossy source coding with uncountable alphabets. Such an exponent was derived by Ihara and Kubo [50] for Gaussian sources using geometrical arguments. Incorporating Ihara and Kubo’s derivations into a CF exponent analysis would be interesting and challenging.

4) Since noisy network coding [51] is a variant of CF that generalizes various network coding scenarios, in the future, we hope to also derive an achievable error exponent based on the noisy network coding strategy and compare that to the CF exponent we derived in Theorem 2. In particular, it would be useful to observe if the resulting noisy network coding exponent is easier to compute compared to the CF exponent.

5) In addition, a combination of DF and CF was used for relay networks with at least 4 nodes in Kramer-Gastpar-Gupta [11]. It may be insightful to derive the corresponding error exponents at least for DF and understand how the exponents scale with the number of nodes in a network.

6) It is natural to wonder whether the technique presented in Section [8] applies to discrete memoryless relay channels with various forms of feedback [3 Sec. 17.4] since techniques from channel coding with feedback [23, 24] were employed to derive the upper bound on the reliability function.

7) Finally, we also expect that the moments of type class enumerator method by Merhav [52, 53] and co-authors may yield an alternate forms of the random coding and expurgated exponents that may have a different interpretation (perhaps, from the statistical physics perspective) vis-à-vis the types-based random coding error exponent presented in Section [3].

APPENDIX A

PROOF OF LEMMA

Proof: Because $X_n^2$ is generated uniformly at random from $\mathcal{T}_V(x_1^n)$,

$$\Pr[\bar{y}^n \in \mathcal{T}_V(x_1^n, X_2^n)] = \frac{1}{|\mathcal{T}_V(x_1^n)|} \sum_{x_2^n \in \mathcal{T}_V(x_1^n)} \{\bar{y}^n \in \mathcal{T}_V(x_1^n, x_2^n)\}. \quad (200)$$

Consider reverse channels $\bar{V} : \mathcal{X}_1 \times \mathcal{Y} \rightarrow \mathcal{X}_2$ and let $\mathcal{R}(V)$ be the collection of reverse channels satisfying $\sum_y \bar{V}(x_2|x_1, y)W(y|x_1) = V(x_2|x_1)$ for all $x_1, x_2, y$. Note that $\bar{y}^n \in \mathcal{T}_V(x_1^n, x_2^n)$ holds if and only if there exists some $\bar{V} \in \mathcal{R}_n(\mathcal{X}_2; P \times W) \cap \mathcal{R}(V)$ such that $x_2^n \in \mathcal{T}_{\bar{V}}(x_1^n, \bar{y}^n)$. Then we may rewrite (200) as

$$\Pr[\bar{y}^n \in \mathcal{T}_V(x_1^n, X_2^n)] = \sum_{\bar{V} \in \mathcal{R}_n(\mathcal{X}_2; P \times W) \cap \mathcal{R}(V)} \exp(nH(\bar{V}|P \times W)) \exp(-|\mathcal{X}_1||\mathcal{X}_2|+1) \exp(nH(\bar{V}^*|P \times W)) \exp(nH(V|P))$$

$$\leq \sum_{\bar{V} \in \mathcal{R}_n(\mathcal{X}_2; P \times W) \cap \mathcal{R}(V)} \frac{\exp(nH(\bar{V}|P \times W))}{(n+1)|\mathcal{X}_1||\mathcal{X}_2|+1} \exp(nH(V|P)) \exp(nH(V|P))$$

$$\leq p_2(n) \exp[-nI(V, V'|P)], \quad (202)$$

where in (203), $\bar{V}^* \in \mathcal{R}_n(\mathcal{X}_2; P \times W)$ is the conditional type that maximizes the conditional entropy $H(\bar{V}|P \times W)$ subject to the constraint that it also belongs to $\mathcal{R}(V)$; and in (204), $p_2(n)$ is some polynomial function of $n$ given in the previous expression, and the equality follows from the fact that $I(X_2; Y|X_1) = H(X_2|X_1) - H(X_2|X_1Y)$ and marginal consistency. The lower bound proceeds similarly,

$$\Pr[\bar{y}^n \in \mathcal{T}_V(x_1^n, X_2^n)] \geq \frac{(n+1)^{-|\mathcal{X}_1||\mathcal{X}_2|+1}}{\exp(nH(\bar{V}^*|P \times W))} = \frac{1}{p_1(n)} \exp[-nI(V, V'|P)], \quad (205)$$
where \( p_1(n) \) is some polynomial. This proves the lemma.

**APPENDIX B**

**PROOF OF LEMMA [3]**

**Proof:** Because \( X_2^n \) is uniformly distributed in \( \mathcal{T}_W(x_1^n) \), we have

\[
\Pr[X_2^n \in \mathcal{T}_W(y^n, x_1^n)] = \sum_{x_2^n \in \mathcal{T}_W(x_1^n)} \frac{1}{|\mathcal{T}_W(x_1^n)|} \mathbb{1}\{x_2^n \in \mathcal{T}_W(y^n, x_1^n)\}.
\]  

(206)

As a result,

\[
\Pr[X_2^n \in \mathcal{T}_W(y^n, x_1^n)] = \frac{|\mathcal{T}_W(x_1^n) \cap \mathcal{T}_W(y^n, x_1^n)|}{|\mathcal{T}_W(x_1^n)|} \leq \frac{|\mathcal{T}_W(y^n, x_1^n)|}{|\mathcal{T}_W(x_1^n)|} \leq \exp(nH(W|P \times V'}} \exp(nH(V|P))^{-1}.
\]  

(207)

Thus, denoting \( p_3(n) \) as some polynomial function of \( n \), we have

\[
\Pr[X_2^n \in \mathcal{T}_W(y^n, x_1^n)] \leq p_3(n) \exp[-nI(V', W|P)]
\]  

(208)

because \( W \) satisfies the marginal consistency property in the statement of the lemma and \( I(X_2; Y|X_1) = H(X_2|X_1) - H(X_2|X_1 Y) \).

Acknowledgements: I am extremely grateful to Yeow-Khiang Chia and Jonathan Scarlett for many helpful discussions and comments that helped to improve the content and the presentation in this work. I would also like to sincerely acknowledge the Associate Editor Aaron Wagner and the two anonymous reviewers for their extensive and useful comments during the revision process.

**REFERENCES**

[1] E. C. van der Meulen, “Three-terminal communication channels,” *Advances in Applied Probability*, vol. 3, pp. 120–54, 1971.
[2] T. Cover and A. El Gamal, “Capacity theorems for the relay channel,” *IEEE Transactions on Information Theory*, vol. 25, no. 5, pp. 572–84, 1979.
[3] A. El Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge, U.K.: Cambridge University Press, 2012.
[4] A. D. Wyner and J. Ziv, “The rate-distortion function for source coding with side information at the decoder,” *IEEE Transactions on Information Theory*, vol. 22, no. 1, pp. 1–10, Jan 1976.
[5] A. El Gamal, M. Mohseni, and S. Zahedi, “Bounds on capacity and minimum energy-per-bit for AWGN relay channels,” *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1545–61, 2006.
[6] H. Sato, “Information transmission through a channel with relay,” The Aloha System, University of Hawaii, Honolulu, Tech. Rep. B76-7, Mar 1976.
[7] V. D. Goppi, “Nonprobabilistic mutual information without memory,” *Probl. Contr. and Inform. Theory*, vol. 4, pp. 97–102, 1975.
[8] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2011.
[9] G. J. Bradford and J. N. Laneman, “Error exponents for block Markov superposition encoding with varying decoding latency,” in *IEEE Information Theory Workshop (ITW)*, 2012.
[10] A. B. Carleial, “Multiple-access channels with different generalized feedback signals,” *IEEE Transactions on Information Theory*, vol. 28, no. 6, pp. 841–850, Nov 1982.
[11] G. Kramer, M. Gastpar, and P. Gupta, “Cooperative strategies and capacity theorems for relay networks,” *IEEE Transactions on Information Theory*, vol. 51, no. 9, pp. 3073–3063, Sep 2005.
[12] C.-M. Zeng, F. Kuhlmann, and A. Buzo, “Achievability proof of some multiuser channel coding theorems using backward decoding,” *IEEE Transactions on Information Theory*, vol. 35, no. 6, pp. 1160–1165, 1989.
[13] F. M. J. Willems and E. C. van der Meulen, “The discrete memoryless multiple-access channel with cribbing encoders,” *IEEE Transactions on Information Theory*, vol. 31, no. 3, pp. 313–327, 1985.
[14] I. Csiszár, “The method of types,” *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2505–23, Oct 1998.
[15] J. Körner and A. Sgarro, “Universally attainable error exponents for broadcast channels with degraded message sets,” *IEEE Transactions on Information Theory*, vol. 26, no. 6, pp. 670–79, 1980.
[16] I. Csiszár, J. Körner, and K. Marton, “A new look at the error exponent of a discrete memoryless channel,” in *IEEE International Symposium on Information Theory (ISIT)*, Cornell University, Ithaca, New York, 1977.
[17] E. A. Haroutunian, M. E. Haroutunian, and A. N. Harutyunyan, *Reliability Criteria in Information Theory and Statistical Hypothesis Testing*, ser. Foundations and Trends in Communications and Information Theory. Now Publishers Inc, 2008, vol. 4.
[18] G. Kelly and A. B. Wagner, “Reliability in source coding with side information,” *IEEE Transactions on Information Theory*, vol. 58, no. 8, pp. 5086–5111, 2012.
[19] J. Scarlett and A. Guillén i Fàbregas, “An achievable error exponent for the mismatched multiple-access channel,” in *50th Annual Allerton Conference on Communication, Control, and Computing*, 2012.
[20] J. Scarlett, A. Martinez, and A. Guillén i Fàbregas, “Multiuser coding techniques for mismatched decoding,” *arXiv:1311.6635*, Nov 2013.
Vincent Y. F. Tan (S’07-M’11) is an Assistant Professor in the Department of Electrical and Computer Engineering (ECE) and the Department of Mathematics at the National University of Singapore (NUS). He received the B.A. and M.Eng. degrees in Electrical and Information Sciences from Cambridge University in 2005. He received the Ph.D. degree in Electrical Engineering and Computer Science (EECS) from the Massachusetts Institute of Technology in 2011. He was a postdoctoral researcher in the Department of ECE at the University of Wisconsin-Madison and following that, a research scientist at the Institute for Infocomm (I²R) Research, A*STAR, Singapore. His research interests include information theory, machine learning and signal processing.

Dr. Tan received the MIT EECS Jin-Au Kong outstanding doctoral thesis prize in 2011 and the NUS Young Investigator Award in 2014. He has authored a research monograph on *Asymptotic Estimates in Information Theory with Non-Vanishing Error Probabilities* in the Foundations and Trends® in Communications and Information Theory Series (NOW Publishers).