Asymptotic Equipartition Property of Output when Rate is above Capacity

Xiugang Wu and Liang-Liang Xie
Department of Electrical and Computer Engineering
University of Waterloo, Waterloo, ON, Canada N2L 3G1
Email: x23wu@uwaterloo.ca, llxie@uwaterloo.ca

Abstract

The output distribution, when rate is above capacity, is investigated. It is shown that there is an asymptotic equipartition property (AEP) of the typical output sequences, independently of the specific codebook used, as long as the codebook is typical according to the standard random codebook generation. This equipartition of the typical output sequences is caused by the mixup of input sequences when there are too many of them, namely, when the rate is above capacity. This discovery sheds some light on the optimal design of the compress-and-forward relay schemes.

I. Introduction

A fundamental observation of Shannon's channel coding theorem is that using a randomly generated codebook (i.i.d. generated according to some $p_0(x)$) at a rate below capacity will lead to a distribution pattern of the output sequences, by which, a decoding scheme with arbitrarily low probability of error can be devised.

In this paper, we are interested in the case when the rate is above capacity. We will show that such a pattern that can be used for decoding will disappear when there are too many input sequences, i.e., when the rate is above capacity. Instead, in this case, the output will have an asymptotic equipartition property on the set of typical output sequences (typical with respect to $p_0(y) = \sum_x p_0(x)p(y|x)$). Interestingly, this set is independent of the specific codebook used, as long as the codebook is typical according to the random codebook generation. The reason for this equipartition is that the input sequences are too dense, so that different input sequences can contribute to the same output sequence and get mixed up.

Part of the work [1] was presented at CWIT 2009.

Investigating the optimal compress-and-forward relay scheme has motivated this study of output distribution when rate is above capacity. The optimality of the compress-and-forward schemes is arguably one of the most critical problems in the development of network information theory, where ambiguity always arises when decoding cannot be done correctly. In the classical approach of [2], the compression scheme at the relay was only based on the distribution used for generating the codebook at the source, instead of the specific codebook generated. While many different codebooks can be generated according to the same distribution, can the knowledge of the specific codebook be helpful? There have been some discussions on this issue (e.g., [3]). Here, in this paper, we show that the observations at the relay are somehow independent of the specific codebook used at the source, and only depend on the distribution by which the codebook is generated.

To further explore the optimality of the compress-and-forward schemes, we compare the rates needed to losslessly compress the relay's observation in two different scenarios: i) the relay uses the knowledge of the source's codebook to do the compression; ii) the relay simply ignores this knowledge. It is shown that the minimum required rates in both scenarios are the same when the rate of the source's codebook is above the capacity of the source-to-relay link.

The remainder of the paper is organized as the following. In Section II, we first introduce some standard definitions of strongly typical sequences, and then give a definition of typical codebooks. Then, we summarize our main results in Section III, followed by the proof of these results in Section IV, V and VI. Finally, as an application of the results, the optimality of the compress-and-forward schemes is discussed in Section VII.

II. Preliminaries

Consider a discrete memoryless channel $(\mathcal{X}, p(y|x), \mathcal{Y})$ with capacity $C := \max_{p(x)} I(X;Y)$. Under the random coding framework, a random codebook $\mathbf{C}$ with respect to $p_0(x)$ with rate R and block length n is defined as

$$\mathbf{C} := \left\{X^n(w) \in \mathcal{X}^n, w = 1, \ldots, 2^{nR}\right\}, \tag{1}$$

where each codeword in $\mathbf{C}$ is an i.i.d. random sequence generated according to a fixed input distribution $p_0(x)$.

It is well known that information can be transmitted with arbitrarily small probability of error for sufficiently large n if $R < C$. In this paper, however, we are interested in the case where the rate is above capacity.

A. Strong Typicality

We begin with some standard definitions on strong typicality [3, Ch.13].

Definition 2.1: The ϵ-strongly typical set with respect to $p_0(x)$, denoted by $A_{\epsilon,0}^{(n)}(X)$, is the set of sequences $x^n \in \mathcal{X}^n$ satisfying:

1. For all $a \in \mathcal{X}$ with $p_0(a) > 0$,

$$\left|\frac{1}{n}N(a|x^n) - p_0(a)\right| < \frac{\epsilon}{|\mathcal{X}|},$$

2. For all $a \in \mathcal{X}$ with $p_0(a) = 0$, $N(a|x^n) = 0$.

$N(a|x^n)$ is the number of occurrences of a in x^n.

Similarly, we can define the ϵ-strongly typical set with respect to $p_0(y)$ and denote it by $A_{\epsilon,0}^{(n)}(Y)$.

Definition 2.2: The ϵ-strongly typical set with respect to $p_0(x,y)$, denoted by $A_{\epsilon,0}^{(n)}(X,Y)$, is the set of sequences $(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$ satisfying:

1. For all $(a,b) \in \mathcal{X} \times \mathcal{Y}$ with $p_0(a,b) > 0$,

$$\left|\frac{1}{n}N(a,b|x^n,y^n) - p_0(a,b)\right| < \frac{\epsilon}{|\mathcal{X}||\mathcal{Y}|},$$

2. For all $(a,b) \in \mathcal{X} \times \mathcal{Y}$ with $p_0(a,b) = 0$,

$$N(a,b|x^n,y^n) = 0.$$

$N(a,b|x^n,y^n)$ is the number of occurrences of the pair (a,b) in the pair of sequences (x^n,y^n).

Definition 2.3: The ϵ-strongly conditionally typical set with the sequence x^n with respect to the conditional distribution $p(y|x)$, denoted by $A_{\epsilon}^{(n)}(Y|x^n)$, is the set of sequences $y^n \in \mathcal{Y}^n$ satisfying:

1. For all $(a,b) \in \mathcal{X} \times \mathcal{Y}$ with $p(b|a) > 0$,

$$\frac{1}{n}\left|N(a,b|x^n,y^n) - p(b|a)N(a|x^n)\right| \leq \epsilon(1 + \frac{1}{|\mathcal{Y}|}), \tag{2}$$

2. For all $(a,b) \in \mathcal{X} \times \mathcal{Y}$ with $p(b|a) = 0$,

$$N(a,b|x^n,y^n) = 0. \tag{3}$$

B. Typical Codebooks

Definition 2.4: For the discrete memoryless channel $(\mathcal{X}, p(y|x), \mathcal{Y})$, the channel noise is said to be ϵ-typical if for any given input x^n, the output Y^n is ϵ-strongly conditionally typical with x^n with respect to the channel transition function $p(y|x)$, i.e., $Y^n \in A_\epsilon^{(n)}(Y|x^n)$.

Due to the Law of Large Numbers, the channel noise is "typical" with high probability.

Index the sequences in $A_{\epsilon,0}^{(n)}(Y)$ as $y_{\epsilon,0}^n(i), i = 1, \ldots, M_{\epsilon,0}^{(n)}$, where $M_{\epsilon,0}^{(n)} = |A_{\epsilon,0}^{(n)}(Y)|$. Consider the set $F_{\epsilon,0}(i) \subseteq \mathcal{X}^n$, where each sequence in $F_{\epsilon,0}(i)$ is strongly typical and can reach $y_{\epsilon,0}^n(i)$ over a channel with typical noise, i.e.,

$$F_{\epsilon,0}(i) := \left\{ x^n \in A_{\epsilon,0}^{(n)}(X) : y_{\epsilon,0}^n(i) \in A_\epsilon^{(n)}(Y|x^n) \right\}.$$

The following notation is useful for defining the typical codebooks.

$$P_{\epsilon,0}(i) := \Pr(\tilde{X}^n \in F_{\epsilon,0}(i) | \tilde{X}^n \in A_{\epsilon,0}^{(n)}(X)),$$
$$N_{\epsilon,0}(i|\mathcal{C}) := \sum_{w=1}^{2^{nR}} \mathbb{I}(x^n(w) \in F_{\epsilon,0}(i)),$$

where $\tilde{X}^n$ is drawn i.i.d. according to $p_0(x)$ and $\mathbb{I}(A)$ is the indicator function:

$$\mathbb{I}(A) = \begin{cases} 1 & \text{if } A \text{ holds,} \\ 0 & \text{otherwise.} \end{cases}$$

Definition 2.5: A codebook

$$\mathcal{C} = \left\{ x^n(w) \in \mathcal{X}^n, w = 1, \ldots, 2^{nR} \right\}$$

is said to be ϵ-typical with respect to $p_0(x)$ if

1) $x^n(w) \in A_{\epsilon,0}^{(n)}(X), \forall w \in \{1, \ldots, 2^{nR}\}$,

2) $\sup_{i \in \{1,\ldots,M_{\epsilon,0}^{(n)}\}} \left| \frac{N_{\epsilon,0}(i|\mathcal{C})}{2^{nR}} - P_{\epsilon,0}(i) \right| \leq \frac{n^3 R}{2^{nR}}$.

III. Main Results

The main results of this paper are summarized by the following three theorems. Their proofs are presented in Sections IV, V and VI respectively. The application of these results to the relay channel will be discussed in Section VII.

Theorem 3.1: Given that an ϵ-typical codebook $\mathcal{C}$ is used and the channel noise is also ϵ-typical, then,[1]

$$\Pr(Y^n = y^n_{\epsilon,0}(i)|\mathcal{C}) \doteq 2^{-nH_0(Y)}, \forall i \in \{1, \ldots, M^{(n)}_{\epsilon,0}\},$$

when $R > I_0(X;Y)$, where both $H_0(Y)$ and $I_0(X;Y)$ are calculated according to $p_0(x,y) = p_0(x)p(y|x)$.

Throughout this paper, we generate the codebook $\mathbf{C}$ at random according to $p_0(x)$ and reserve only the ϵ-strongly typical codewords. Then we have Theorem 3.2 and 3.3.

Theorem 3.2: For any $\epsilon > 0$,

$$\Pr(\mathbf{C} \text{ is } \epsilon\text{-typical}) \to 1 \text{ as } n \to \infty. \tag{4}$$

Theorem 3.3: Consider the conditional entropy of the channel output given the source's codebook information, namely $H(Y^n|\mathbf{C})$. We have

$$\lim_{n\to\infty} \frac{1}{n} H(Y^n|\mathbf{C}) = \begin{cases} H_0(Y) & \text{when } R > I_0(X;Y), \\ R + H_0(Y|X) & \text{when } R < I_0(X;Y), \end{cases}$$

where $H_0(Y)$, $I_0(X;Y)$ and $H_0(Y|X)$ are all calculated according to $p_0(x,y) = p_0(x)p(y|x)$.

In contrast, without the codebook information, we have

$$\lim_{n\to\infty} \frac{1}{n} H(Y^n) = H_0(Y) \text{ for any } R > 0.$$

IV. AEP OF TYPICAL OUTPUT SEQUENCES

Essentially, Theorem 3.1 states that there exists an asymptotic equipartition property of the typical output sequences, irrespective of the specific codebook used, as long as the codebook is a typical codebook. To prove this theorem, we first introduce two lemmas.

Lemma 4.1: Let E_ϵ denote the event that the output $Y^n \in A^{(n)}_\epsilon(Y|x^n)$ for any given input x^n. For any $x^n \in F_{\epsilon,0}(i)$,

$$\begin{aligned} &\Pr(Y^n = y^n_{\epsilon,0}(i)|E_\epsilon, X^n = x^n) \geq 2^{-n(H_0(Y|X)+\epsilon_0)} \\ \text{and } &\Pr(Y^n = y^n_{\epsilon,0}(i)|E_\epsilon, X^n = x^n) \leq 2^{-n(H_0(Y|X)-\epsilon_0)}, \end{aligned}$$

where $H_0(Y|X)$ is calculated according to $p_0(x,y) = p_0(x)p(y|x)$ and ϵ_0 goes to 0 as $\epsilon \to 0$ and $n \to \infty$.

[1]Same as the notation in [4], we say $a_n \doteq b_n$ if $\lim_{n\to\infty} \frac{1}{n} \log \frac{a_n}{b_n} = 0$. "$\dot{\geq}$" and "$\dot{\leq}$" have similar interpretations.

Proof: By the definition of $F_{\epsilon,0}(i)$, we have for any x^n in $F_{\epsilon,0}(i)$, $x^n \in A_{\epsilon,0}^{(n)}(X)$ and $y_{\epsilon,0}^n(i) \in A_{\epsilon}^{(n)}(Y|x^n)$. Then, it follows from the definition of strong typicality that $(x^n, y_{\epsilon,0}^n(i)) \in A_{\epsilon',0}^{(n)}(X,Y)$, where $\epsilon' \to 0$ as $\epsilon \to 0$. Since strong typicality implies weak typicality, for any x^n in $F_{\epsilon,0}(i)$, we have

$$\left|-\frac{1}{n}\log p(x^n) - H_0(X)\right| < \epsilon'',$$
$$\left|-\frac{1}{n}\log p(x^n, y_{\epsilon,0}^n(i)) - H_0(X,Y)\right| < \epsilon'',$$

where $\epsilon'' \to 0$ as $\epsilon \to 0$. Thus,

$$\left|-\frac{1}{n}\log p(y_{\epsilon,0}^n(i)|x^n) - H_0(Y|X)\right| < 2\epsilon'',$$

and

$$2^{-n(H_0(Y|X)+2\epsilon'')} \le p(y_{\epsilon,0}^n(i)|x^n) \le 2^{-n(H_0(Y|X)-2\epsilon'')}.$$

Therefore, for any $x^n \in F_{\epsilon,0}(i)$, we have

$$\begin{aligned}
&\mathrm{Pr}(Y^n = y_{\epsilon,0}^n(i)|E_\epsilon, X^n = x^n)\\
=&\frac{\mathrm{Pr}(Y^n = y_{\epsilon,0}^n(i), E_\epsilon, X^n = x^n)}{\mathrm{Pr}(E_\epsilon, X^n = x^n)}\\
=&\frac{\mathrm{Pr}(Y^n = y_{\epsilon,0}^n(i), X^n = x^n)}{\mathrm{Pr}(E_\epsilon|X^n = x^n)\mathrm{Pr}(X^n = x^n)}\\
=&\frac{p(y_{\epsilon,0}^n(i)|x^n)}{\mathrm{Pr}(E_\epsilon|X^n = x^n)}\\
=&(1+o(1))p(y_{\epsilon,0}^n(i)|x^n)\\
\le&(1+o(1))2^{-n(H_0(Y|X)-2\epsilon'')}\\
=&2^{-n(H_0(Y|X)-\epsilon_0)},
\end{aligned}$$

where $\epsilon_0 := 2\epsilon'' + \frac{\log(1+o(1))}{n}$ and $\epsilon_0 \to 0$ as $\epsilon \to 0$ and $n \to \infty$. Similarly, for any $x^n \in F_{\epsilon,0}(i)$, we have

$$\begin{aligned}
&\mathrm{Pr}(Y^n = y_{\epsilon,0}^n(i)|E_\epsilon, X^n = x^n)\\
=&(1+o(1))p(y_{\epsilon,0}^n(i)|x^n)\\
\ge&2^{-n(H_0(Y|X)+2\epsilon''-\frac{\log(1+o(1))}{n})}\\
\ge&2^{-n(H_0(Y|X)+\epsilon_0)},
\end{aligned}$$

which finishes the proof of Lemma 4.1. ■

Lemma 4.2: If $\mathcal{C}$ is a typical codebook, then for any $i \in \{1, \ldots, M_{\epsilon,0}^{(n)}\}$,

$$N_{\epsilon,0}(i|\mathcal{C}) \geq 2^{nR} \cdot 2^{-n(I_0(X;Y)+\epsilon_0')} - n^3 R$$
$$\text{and} \quad N_{\epsilon,0}(i|\mathcal{C}) \leq 2^{nR} \cdot 2^{-n(I_0(X;Y)-\epsilon_0')} + n^3 R,$$

where $I_0(X;Y)$ is calculated according to $p_0(x)p(y|x)$ and ϵ_0' goes to 0 as $\epsilon \to 0$ and $n \to \infty$.

Proof: To prove Lemma 4.2, we need the following standard result on strong typicality (see Lemma 13.6.2 in [4]):

Let $\tilde{X}^n$ be drawn i.i.d. according to $p_0(x) = \sum_y p_0(x,y)$. For $y^n \in A_{\epsilon,0}^{(n)}(Y)$,

$$\Pr((\tilde{X}^n, y^n) \in A_{\epsilon,0}^{(n)}(X,Y)) \geq 2^{-n(I_0(X;Y)+\epsilon_1)} \tag{5}$$
$$\text{and} \quad \Pr((\tilde{X}^n, y^n) \in A_{\epsilon,0}^{(n)}(X,Y)) \leq 2^{-n(I_0(X;Y)-\epsilon_1)}, \tag{6}$$

where $I_0(X;Y)$ is calculated according to $p_0(x,y)$ and ϵ_1 goes to 0 as $\epsilon \to 0$ and $n \to \infty$.

According to the definition of $P_{\epsilon,0}(i)$,

$$P_{\epsilon,0}(i) = \Pr(y_{\epsilon,0}^n(i) \in A_\epsilon^{(n)}(Y|\tilde{X}^n)|\tilde{X}^n \in A_{\epsilon,0}^{(n)}(X)),$$

where $\tilde{X}^n$ is drawn i.i.d. according to $p_0(x)$.

Since $\tilde{X}^n \in A_{\epsilon,0}^{(n)}(X)$ and $y_{\epsilon,0}^n(i) \in A_\epsilon^{(n)}(Y|\tilde{X}^n)$ imply that $(\tilde{X}^n, y_{\epsilon,0}^n(i)) \in A_{\epsilon',0}^{(n)}(X,Y)$, where ϵ' goes to 0 as $\epsilon \to 0$, we have

$$\begin{aligned}
P_{\epsilon,0}(i) \leq & \Pr((\tilde{X}^n, y_{\epsilon,0}^n(i)) \in A_{\epsilon',0}^{(n)}(X,Y)|\tilde{X}^n \in A_{\epsilon,0}^{(n)}(X)) \\
= & \frac{\Pr((\tilde{X}^n, y_{\epsilon,0}^n(i)) \in A_{\epsilon',0}^{(n)}(X,Y), \tilde{X}^n \in A_{\epsilon,0}^{(n)}(X))}{\Pr(\tilde{X}^n \in A_{\epsilon,0}^{(n)}(X))} \\
\leq & (1+o(1))\Pr((\tilde{X}^n, y_{\epsilon,0}^n(i)) \in A_{\epsilon',0}^{(n)}(X,Y)) \\
\leq & (1+o(1))2^{-n(I_0(X;Y)-\epsilon_1')} \\
= & 2^{-n(I_0(X;Y)-\epsilon_1'-\frac{\log(1+o(1))}{n})} \\
= & 2^{-n(I_0(X;Y)-\epsilon_2')}
\end{aligned} \tag{7}$$

where $\epsilon_2' := \epsilon_1' + \frac{\log(1+o(1))}{n}$ and $\epsilon_2' \to 0$ as $\epsilon \to 0$ and $n \to \infty$.

Furthermore, by the standard definitions of strong typicality, it follows that $(x^n, y_{\epsilon,0}^n(i)) \in A_{\epsilon,0}^{(n)}(X,Y)$ implies $x^n \in A_{\epsilon,0}^{(n)}(X)$. Now, we show $(x^n, y_{\epsilon,0}^n(i)) \in A_{\epsilon,0}^{(n)}(X,Y)$ also implies $y_{\epsilon,0}^n(i) \in A_\epsilon^{(n)}(Y|x^n)$. Suppose $(x^n, y_{\epsilon,0}^n(i)) \in A_{\epsilon,0}^{(n)}(X,Y)$. Then, we have

1) For all $(a,b) \in \mathcal{X} \times \mathcal{Y}$ with $p(b|a) = 0$, $p_0(a,b) = 0$ and $N(a,b|x^n, y^n_{\epsilon,0}(i)) = 0$.
2) For all $(a,b) \in \mathcal{X} \times \mathcal{Y}$ with $p(b|a) > 0$ and $p_0(a) = 0$, $p_0(a,b) = 0$ and $N(a,b|x^n, y^n_{\epsilon,0}(i)) = 0$, as well as $N(a|x^n) = 0$.
3) For all $(a,b) \in \mathcal{X} \times \mathcal{Y}$ with $p(b|a) > 0$ and $p_0(a) > 0$, $p_0(a,b) > 0$ and

$$\left|\frac{1}{n}N(a|x^n) - p_0(a)\right| < \frac{\epsilon}{|\mathcal{X}|},$$
$$\left|\frac{1}{n}N(a,b|x^n, y^n_{\epsilon,0}(i)) - p_0(a,b)\right| < \frac{\epsilon}{|\mathcal{X}||\mathcal{Y}|}.$$

Thus,

$$\begin{aligned}
&\left|\frac{1}{n}N(a,b|x^n, y^n_{\epsilon,0}(i)) - \frac{1}{n}N(a|x^n)p(b|a)\right| \\
<&p_0(a,b) + \frac{\epsilon}{|\mathcal{X}||\mathcal{Y}|} - p(b|a)(p_0(a) - \frac{\epsilon}{|\mathcal{X}|}) \\
=&\frac{\epsilon}{|\mathcal{X}||\mathcal{Y}|} + p(b|a)\frac{\epsilon}{|\mathcal{X}|} \\
\leq&\frac{\epsilon}{|\mathcal{Y}|} + \epsilon \\
=&\epsilon(1 + \frac{1}{|\mathcal{Y}|}).
\end{aligned}$$

Therefore, $(x^n, y^n_{\epsilon,0}(i)) \in A^{(n)}_{\epsilon,0}(X,Y)$ implies that $y^n_{\epsilon,0}(i) \in A^{(n)}_{\epsilon}(Y|x^n)$, as well as $x^n \in A^{(n)}_{\epsilon,0}(X)$.

Then, we have

$$\begin{aligned}
P_{\epsilon,0}(i) =&\text{Pr}(y^n_{\epsilon,0}(i) \in A^{(n)}_{\epsilon}(Y|\tilde{X}^n)|\tilde{X}^n \in A^{(n)}_{\epsilon,0}(X)) \\
\geq&\text{Pr}((\tilde{X}^n, y^n_{\epsilon,0}(i)) \in A^{(n)}_{\epsilon,0}(X,Y)|\tilde{X}^n \in A^{(n)}_{\epsilon,0}(X)) \\
=&\frac{\text{Pr}((\tilde{X}^n, y^n_{\epsilon,0}(i)) \in A^{(n)}_{\epsilon,0}(X,Y), \tilde{X}^n \in A^{(n)}_{\epsilon,0}(X))}{\text{Pr}(\tilde{X}^n \in A^{(n)}_{\epsilon,0}(X))} \\
=&(1 + o(1))\text{Pr}((\tilde{X}^n, y^n_{\epsilon,0}(i)) \in A^{(n)}_{\epsilon,0}(X,Y)) \\
\geq&(1 + o(1))2^{-n(I_0(X;Y)+\epsilon_1)} \\
=&2^{-n(I_0(X;Y)+\epsilon_1 - \frac{\log(1+o(1))}{n})} \\
=&2^{-n(I_0(X;Y)+\epsilon_2)}
\end{aligned} \tag{8}$$

where $\epsilon_2 := \epsilon_1 - \frac{\log(1+o(1))}{n}$ and $\epsilon_2 \to 0$ as $\epsilon \to 0$ and $n \to \infty$.

Let $\epsilon'_0 = \max\{\epsilon_2, \epsilon'_2\}$. Combining (7) and (8), we have

$$2^{-n(I_0(X;Y)+\epsilon'_0)} \leq P_{\epsilon,0}(i) \leq 2^{-n(I_0(X;Y)-\epsilon'_0)}. \tag{9}$$

Therefore, if $\mathcal{C}$ is a typical codebook, by the definition of the typical codebooks and (9), for any $i \in \{1, \ldots, M_{\epsilon,0}^{(n)}\}$,

$$N_{\epsilon,0}(i|\mathcal{C}) \geq 2^{nR} \cdot 2^{-n(I_0(X;Y)+\epsilon_0')} - n^3 R$$
$$\text{and} \quad N_{\epsilon,0}(i|\mathcal{C}) \leq 2^{nR} \cdot 2^{-n(I_0(X;Y)-\epsilon_0')} + n^3 R,$$

where $I_0(X;Y)$ is calculated according to $p_0(x)p(y|x)$ and ϵ_0' goes to 0 as $\epsilon \to 0$ and $n \to \infty$. ■

Proof: [Proof of Theorem 3.1] Let E_ϵ denote the event $Y^n \in A_\epsilon^{(n)}(Y|x^n)$ for any given input x^n. Consider $\Pr(Y^n = y_{\epsilon,0}^n(i)|E_\epsilon, \mathcal{C} \text{ is typical})$ for any $i \in \{1, \ldots, M_{\epsilon,0}^{(n)}\}$. We lower bound this probability as follows:

$$\begin{aligned}
&\Pr(Y^n = y_{\epsilon,0}^n(i)|E_\epsilon, \mathcal{C} \text{ is typical}) \\
=&\sum_{w=1}^{2^{nR}} \Pr(Y^n = y_{\epsilon,0}^n(i)|E_\epsilon, \mathcal{C} \text{ is typical}, X^n = x^n(w)) \\
&\cdot \Pr(X^n = x^n(w)|E_\epsilon, \mathcal{C} \text{ is typical}) && (10)\\
=&\frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \Pr(Y^n = y_{\epsilon,0}^n(i)|E_\epsilon, \mathcal{C} \text{ is typical}, X^n = x^n(w)) && (11)\\
=&\frac{1}{2^{nR}} \sum_{x^n(w) \in F_{\epsilon,0}(i)} \Pr(Y^n = y_{\epsilon,0}^n(i)|E_\epsilon, \mathcal{C} \text{ is typical}, X^n = x^n(w)) && (12)\\
\geq&\frac{1}{2^{nR}} N_{\epsilon,0}(i|\mathcal{C}) \cdot 2^{-n(H_0(Y|X)+\epsilon_0)} && (13)\\
\geq&\frac{1}{2^{nR}} (2^{nR} \cdot 2^{-n(I_0(X;Y)+\epsilon_0')} - n^3 R) \cdot 2^{-n(H_0(Y|X)+\epsilon_0)} && (14)\\
=&2^{-n(H_0(Y)+\epsilon_0+\epsilon_0')} \cdot \left[1 - \frac{n^3 R}{2^{nR}} \cdot 2^{n(I_0(X;Y)+\epsilon_0')}\right].
\end{aligned}$$

(10) follows from the Law of Total Probability and accumulates the contributions from all the codewords in the codebook to the probability for $y_{\epsilon,0}^n(i)$ to be channel output.

(11) follows from the uniform distribution of message index W.

(12) follows from the the condition E_ϵ and the fact that $\mathcal{C}$ contains only strongly typical codewords.

(13) follows from Lemma 4.1.

(14) follows from Lemma 4.2.

Let $\epsilon \to 0$ as $n \to \infty$. Then for any $i \in \{1, \ldots, M_{\epsilon,0}^{(n)}\}$,

$$\Pr(Y^n = y_\epsilon^n(i)|E_\epsilon, \mathcal{C} \text{ is typical}) \dot{\geq} 2^{-nH_0(Y)}, \tag{15}$$

when $R > I_0(X;Y)$.

Similarly, following (12), by Lemmas 4.1 and 4.2, we have

$$\begin{aligned}
&\Pr(Y^n = y^n_{\epsilon,0}(i)|E_\epsilon, \mathcal{C} \text{ is typical}) \\
\leq &\frac{1}{2^{nR}} N_{\epsilon,0}(i|\mathcal{C}) \cdot 2^{-n(H_0(Y|X)-\epsilon_0)} \\
\leq &\frac{1}{2^{nR}} (2^{nR} \cdot 2^{-n(I_0(X;Y)-\epsilon_0')} + n^3 R) \cdot 2^{-n(H_0(Y|X)-\epsilon_0)} \\
= &2^{-n(H_0(Y)-\epsilon_0-\epsilon_0')} \cdot \left[1 + \frac{n^3 R}{2^{nR}} \cdot 2^{n(I_0(X;Y)-\epsilon_0')}\right].
\end{aligned}$$

Therefore, for any $i \in \{1, \ldots, M^{(n)}_{\epsilon,0}\}$,

$$\Pr(Y^n = y^n_{\epsilon,0}(i)|E_\epsilon, \mathcal{C} \text{ is typical}) \dot{\leq} 2^{-nH_0(Y)}, \tag{16}$$

when $R > I_0(X;Y)$. Combining (15) and (16), we establish Theorem 3.1. ■

V. The Probability that A Typical Codebook Appears

In this section, we will show that with high probability, a typical codebook will be generated by the random codebook generation. We begin with some relevant definitions and the Vapnik-Chervonenkis Theorem [5], [6]:

A Range Space is a pair $(X, \mathcal{F})$, where X is a set and $\mathcal{F}$ is a family of subsets of X. For any $A \subseteq X$, we define $P_{\mathcal{F}}(A)$, the projection of $\mathcal{F}$ on A, as $\{F \cap A : F \in \mathcal{F}\}$. We say that A is *shattered* by $\mathcal{F}$ if $P_{\mathcal{F}}(A) = 2^A$, i.e., if the projection of $\mathcal{F}$ on A includes all possible subsets of A. The VC-dimension of $\mathcal{F}$, denoted by VC-d($\mathcal{F}$) is the cardinality of the largest set A that $\mathcal{F}$ shatters. If arbitrarily large finite sets are shattered, the VC dimension of $\mathcal{F}$ is infinite.

The Vapnik-Chervonenkis Theorem: If $\mathcal{F}$ is a set of finite VC-dimension and $\{Y_j\}$ is a sequence of n i.i.d. random variables with common probability distribution P, then for every ϵ, $\delta > 0$

$$\Pr\left\{\sup_{F\in\mathcal{F}} \left|\frac{1}{n}\sum_{j=1}^{n} \mathbb{I}(Y_j \in F) - P(F)\right| \leq \epsilon\right\} > 1 - \delta \tag{17}$$

whenever

$$n > \max\left\{\frac{8\text{VC-d}(\mathcal{F})}{\epsilon}\log_2\frac{16e}{\epsilon}, \frac{4}{\epsilon}\log_2\frac{2}{\delta}\right\}. \tag{18}$$

Let $\mathcal{F}_{\epsilon,0} = \{F_{\epsilon,0}(i), i = 1, \ldots, M^{(n)}_{\epsilon,0}\}$. To show Theorem 3.2, a finite VC dimension of $\mathcal{F}_{\epsilon,0}$ is desired in order to employ the Vapnik-Chervonenkis Theorem. For this reason, we introduce Lemma 5.1.

Lemma 5.1: For a fixed block length n, VC-d$(\mathcal{F}_{\epsilon,0}) \leq n(H_0(Y)+\epsilon')$, where $\epsilon' \to 0$ as $\epsilon \to 0$.

Proof: By the Asymptotic Equipartition Property, $|\mathcal{F}_{\epsilon,0}| = M_{\epsilon,0}^{(n)} \leq 2^{n(H_0(Y)+\epsilon')}$, where $\epsilon' \to 0$ as $\epsilon \to 0$. Thus, for any $A \subseteq \mathcal{X}^n$,

$$|\{F_{\epsilon,0}(i) \cap A : F_{\epsilon,0}(i) \in \mathcal{F}_{\epsilon,0}\}| \leq 2^{n(H_0(Y)+\epsilon')},$$

and hence VC-d$(\mathcal{F}_{\epsilon,0}) \leq n(H_0(Y)+\epsilon')$. ■

Proof: [Proof of Theorem 3.2] Since we reserve only the ϵ-strongly typical codewords when generating the codebook, for any random codebook, the first condition in Definition 2.5 is obviously satisfied. Below, we focus on showing that a random codebook satisfies the second condition in Definition 2.5 with high probability.

For the given $p_0(x)$, consider all the codewords in a random codebook, $X^n(w)$, $w = 1, \ldots, 2^{nR}$. They are generated with the common distribution $p(x^n) = \Pr(\tilde{X}^n = x^n | \tilde{X}^n \in A_{\epsilon,0}^{(n)}(X))$, where $\tilde{X}^n$ is drawn i.i.d. according to $p_0(x)$. Since VC-d$(\mathcal{F}_{\epsilon,0})$ is finite for a fixed n, we employ the Vapnik-Chervonenkis Theorem under the range space $(\mathcal{X}^n, \mathcal{F}_{\epsilon,0})$. To satisfy (18), let both ϵ and δ in (17) be $\frac{\Delta_\epsilon nR}{2^{nR}}$, where $\Delta_\epsilon := \max\{8\text{VC-d}(\mathcal{F}_{\epsilon,0}), 16e\}$. Then the Vapnik-Chervonenkis Theorem states that

$$\begin{aligned}
&\Pr\left\{\sup_{F_{\epsilon,0}(i)\in\mathcal{F}_{\epsilon,0}} \left|\frac{N_{\epsilon,0}(i|\mathbf{C})}{2^{nR}} - P_{\epsilon,0}(i)\right| \leq \frac{\Delta_\epsilon nR}{2^{nR}}\right\} \\
\geq & 1 - \frac{\Delta_\epsilon nR}{2^{nR}} \\
\to & 1 \text{ as } n \to \infty, \qquad (19)
\end{aligned}$$

where $N_{\epsilon,0}(i|\mathbf{C}) = \sum_{w=1}^{2^{nR}} \mathbb{I}(X^n(w) \in F_{\epsilon,0}(i))$. Since $\frac{n^3 R}{2^{nR}} \geq \frac{\Delta_\epsilon nR}{2^{nR}}$ for sufficiently large n, (19) concludes the proof of Theorem 3.2. ■

VI. Proof of Theorem 3.3

Before proceeding to the proof of Theorem 3.3, we first introduce Lemma 6.1, which will facilitate the later discussions. The proof of Lemma 6.1 is given in Appendix I.

Lemma 6.1: For the channel $(\mathcal{X}, p(y|x), \mathcal{Y})$, generate the codebook at random according to $p_0(x)$ and reserve only the ϵ-strongly typical codewords. The channel input and output X^n and Y^n satisfy that

1) $\Pr((X^n, Y^n) \in A_{\epsilon,0}^{(n)}(X,Y)) \to 1$ as $n \to \infty$, for any $\epsilon > 0$;

2) $\lim_{n\to\infty}\frac{1}{n}H(X^n) = H_0(X)$, $\lim_{n\to\infty}\frac{1}{n}H(Y^n) = H_0(Y)$, and $\lim_{n\to\infty}\frac{1}{n}H(X^n,Y^n) = H_0(X,Y)$.

Remark 6.1: Since we reserve only the ϵ-typical codewords when generating the codebook, generally, the channel input X^n is no longer an i.i.d. random process. However, Lemma 6.1 essentially states that the random process (X^n,Y^n) still satisfies the joint asymptotic equipartition property and furthermore, the entropy rates of the random processes X^n, Y^n and (X^n,Y^n) can still be simply expressed in the single letter form respectively. This observation will facilitate our later discussions.

Proof: [Proof of Theorem 3.3] We prove Theorem 3.3 by characterizing $\lim_{n\to\infty}\frac{1}{n}H(Y^n|\mathbf{C})$ in two different cases: when $R > I_0(X;Y)$ and when $R < I_0(X;Y)$, respectively.

A. When $R > I_0(X;Y)$

Define an indicator random variable E as

$$E := \mathbb{I}(E_\epsilon),$$

where E_ϵ denotes the event $Y^n \in A_\epsilon^{(n)}(Y|x^n)$ for any given input x^n.

When $R > I_0(X;Y)$, we have

$$
\begin{aligned}
& H(Y^n|\mathbf{C}) \\
\geq & H(Y^n|E,\mathbf{C}) && (20)\\
= & \Pr(E=1)H(Y^n|E=1,\mathbf{C}) + \Pr(E=0)H(Y^n|E=0,\mathbf{C}) \\
\geq & \Pr(E=1)\cdot H(Y^n|E=1,\mathbf{C}) \\
= & (1-o(1))\cdot H(Y^n|E=1,\mathbf{C}) && (21)\\
= & (1-o(1))\cdot \sum_{\mathcal{C}} p(\mathcal{C})\cdot H(Y^n|E=1,\mathbf{C}=\mathcal{C}) \\
\geq & (1-o(1))\cdot \sum_{\mathcal{C} \text{ is typical}} p(\mathcal{C})\cdot H(Y^n|E=1,\mathbf{C}=\mathcal{C}) \\
= & (1-o(1))\cdot \sum_{\mathcal{C} \text{ is typical}} p(\mathcal{C})\cdot \left(\sum_{y^n} p(y^n|E_\epsilon,\mathcal{C})\log\frac{1}{p(y^n|E_\epsilon,\mathcal{C})}\right) \\
\geq & (1-o(1))\cdot \sum_{\mathcal{C} \text{ is typical}} p(\mathcal{C})\cdot \left(\sum_{y^n\in A^{(n)}_{\epsilon,0}(Y)} p(y^n|E_\epsilon,\mathcal{C})\log\frac{1}{p(y^n|E_\epsilon,\mathcal{C})}\right) \\
\geq & (1-o(1))\cdot \sum_{\mathcal{C} \text{ is typical}} p(\mathcal{C})\cdot \left(\sum_{y^n\in A^{(n)}_{\epsilon,0}(Y)} p(y^n|E_\epsilon,\mathcal{C})\log 2^{n[H_0(Y)-\epsilon^*]}\right) && (22)\\
= & n[H_0(Y)-\epsilon^*]\cdot(1-o(1))\cdot \sum_{\mathcal{C} \text{ is typical}} p(\mathcal{C})\cdot \left(\sum_{y^n\in A^{(n)}_{\epsilon,0}(Y)} p(y^n|E_\epsilon,\mathcal{C})\right) \\
= & n[H_0(Y)-\epsilon^*]\cdot(1-o(1))\cdot \sum_{\mathcal{C} \text{ is typical}} p(\mathcal{C})\cdot \Pr(Y^n\in A^{(n)}_{\epsilon,0}(Y)|E_\epsilon,\mathcal{C}) \\
= & n[H_0(Y)-\epsilon^*]\cdot(1-o(1))\cdot \sum_{\mathcal{C} \text{ is typical}} p(\mathcal{C}|E_\epsilon)\cdot \Pr(Y^n\in A^{(n)}_{\epsilon,0}(Y)|E_\epsilon,\mathcal{C}) \\
= & n[H_0(Y)-\epsilon^*]\cdot(1-o(1))\cdot \Pr(Y^n\in A^{(n)}_{\epsilon,0}(Y),\mathbf{C}\text{ is typical}|E_\epsilon) \\
= & n[H_0(Y)-\epsilon^*]\cdot(1-o(1))\cdot(1-o(1)) && (23)\\
= & n[H_0(Y)-\epsilon^*]\cdot(1-o(1))
\end{aligned}
$$

(20) follows from the fact that conditioning reduces entropy.

(21) follows from the fact that $\Pr(E_\epsilon)\to 1$ as $n\to\infty$, for any $\epsilon>0$.

(22) follows from Theorem 3.1, which upper bounds $p(y^n|E_\epsilon,\mathcal{C})$ by $2^{-n[H_0(Y)-\epsilon^*]}$ for any $y^n\in A^{(n)}_{\epsilon,0}(Y)$ and typical $\mathcal{C}$, where $\epsilon^*\to 0$ as $n\to\infty$.

(23) follows from the fact that

$$\Pr(Y^n \in A_{\epsilon,0}^{(n)}(Y), \mathbf{C} \text{ is typical} | E_\epsilon) \to 1 \text{ as } n \to \infty.$$

This can be seen from the following.

$$\begin{aligned}
&\Pr(Y^n \in A_{\epsilon,0}^{(n)}(Y), \mathbf{C} \text{ is typical} | E_\epsilon) \\
=&\frac{\Pr(Y^n \in A_{\epsilon,0}^{(n)}(Y), E_\epsilon, \mathbf{C} \text{ is typical})}{\Pr(E_\epsilon)} \\
\geq&\frac{\Pr((X^n, Y^n) \in A_{\epsilon,0}^{(n)}(X,Y), E_\epsilon, \mathbf{C} \text{ is typical})}{\Pr(E_\epsilon)}.
\end{aligned} \tag{24}$$

Since $\Pr(E_\epsilon)$, $\Pr(\mathbf{C} \text{ is typical})$ and $\Pr((X^n, Y^n) \in A_{\epsilon,0}^{(n)}(X,Y))$ all go to 1, obviously both the numerator and denominator of (24) go to 1 as $n \to \infty$. Thus,

$$\Pr(Y^n \in A_{\epsilon,0}^{(n)}(Y), \mathbf{C} \text{ is typical} | E_\epsilon) \to 1 \text{ as } n \to \infty.$$

Therefore, when $R > I_0(X;Y)$,

$$\begin{aligned}
&\liminf_{n\to\infty} \frac{1}{n} H(Y^n|\mathbf{C}) \\
\geq &\liminf_{n\to\infty} \frac{1}{n} \left(n[H_0(Y) - \epsilon^*] \cdot (1 - o(1))\right) \\
= &\liminf_{n\to\infty} [H_0(Y) - \epsilon^*] \cdot (1 - o(1)) \\
= &H_0(Y).
\end{aligned} \tag{25}$$

Furthermore,

$$\limsup_{n\to\infty} \frac{1}{n} H(Y^n|\mathbf{C}) \leq \limsup_{n\to\infty} \frac{1}{n} H(Y^n) = H_0(Y), \tag{26}$$

where the last equality follows from Lemma 6.1.

Combining (25) and (26), we have that when $R > I_0(X;Y)$,

$$\lim_{n\to\infty} \frac{1}{n} H(Y^n|\mathbf{C}) = H_0(Y).$$

B. When $R < I_0(X;Y)$

To find $\lim_{n\to\infty} \frac{1}{n} H(Y^n|\mathbf{C})$ when $R < I_0(X;Y)$, we first introduce two lemmas. The proofs of these two lemmas are given at the end of this section.

Lemma 6.2: When $R < I_0(X;Y)$,

$$\frac{1}{n}H(X^n|\mathbf{C},Y^n) \to 0, \text{ as } n \to \infty.$$

Lemma 6.3:

$$\lim_{n\to\infty}\frac{1}{n}H(X^n|\mathbf{C}) = R.$$

Now, expanding $H(X^n,Y^n|\mathbf{C})$ in two different ways, we have

$$\begin{aligned}H(X^n,Y^n|\mathbf{C}) =& H(X^n|\mathbf{C}) + H(Y^n|X^n,\mathbf{C})\\ =& H(Y^n|\mathbf{C}) + H(X^n|\mathbf{C},Y^n),\end{aligned}$$

and thus

$$H(Y^n|\mathbf{C}) = H(X^n|\mathbf{C}) + H(Y^n|X^n,\mathbf{C}) - H(X^n|\mathbf{C},Y^n).$$

Therefore, when $R < I_0(X;Y)$,

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}H(Y^n|\mathbf{C}) &= \lim_{n\to\infty}\frac{1}{n}H(X^n|\mathbf{C}) + \lim_{n\to\infty}\frac{1}{n}H(Y^n|X^n,\mathbf{C}) - \lim_{n\to\infty}\frac{1}{n}H(X^n|\mathbf{C},Y^n)\\ &= R + \lim_{n\to\infty}\frac{1}{n}H(Y^n|X^n,\mathbf{C}) \qquad (27)\\ &= R + \lim_{n\to\infty}\frac{1}{n}H(Y^n|X^n) \qquad (28)\\ &= R + \lim_{n\to\infty}\frac{1}{n}[H(X^n,Y^n) - H(X^n)]\\ &= R + H_0(X,Y) - H_0(X) \qquad (29)\\ &= R + H_0(Y|X),\end{aligned}$$

where (27) follows from Lemma 6.2 and 6.3, (28) follows from the fact that $\mathbf{C} \to X^n \to Y^n$ forms a Markov Chain, and (29) follows from Lemma 6.1. This completes the proof of Theorem 3.3. ■

Proof: [Proof of Lemma 6.2] To prove Lemma 6.2, we begin with Fano's Inequality (see Theorem 2.11.1 in [4]):

Let $P_e = \Pr(g(Y) \neq X)$, where g is any function of Y. Then

$$1 + P_e \log|\mathcal{X}| \geq H(X|Y). \qquad (30)$$

For the channel $(\mathcal{X}, p(y|x), \mathcal{Y})$ with a codebook $\mathcal{C}$, we estimate the message index W from Y^n. Let the estimate be $\hat{W} = g(Y^n)$ and $P_e^{(n)}(\mathcal{C}) = \Pr(W \neq g(Y^n)|\mathcal{C})$. Then, applying Fano's Inequality, we have

$$H(W|Y^n,\mathcal{C}) \leq 1 + P_e^{(n)}(\mathcal{C})\log 2^{nR} = 1 + P_e^{(n)}(\mathcal{C})nR.$$

Since given $\mathcal{C}$, X^n is a function of W, say $X^n = X^n(W)$, we have

$$H(X^n|Y^n,\mathcal{C}) \le H(W|Y^n,\mathcal{C}) \le 1 + P_e^{(n)}(\mathcal{C})nR.$$

Then,

$$H(X^n|Y^n,\mathbf{C}) = \sum_{\mathcal{C}} p(\mathcal{C})H(W|Y^n,\mathcal{C}) \le \sum_{\mathcal{C}} p(\mathcal{C})(1 + P_e^{(n)}(\mathcal{C})nR).$$

Recall the channel coding theorem, which states that if we randomly generate the codebook according to $p_0(x)$, then when $R < I_0(X;Y)$,

$$\sum_{\mathcal{C}} p(\mathcal{C})P_e^{(n)}(\mathcal{C}) \to 0. \tag{31}$$

Therefore, when $R < I_0(X;Y)$,

$$\begin{aligned}
\limsup_{n\to\infty} \frac{1}{n} H(X^n|Y^n,\mathbf{C}) \le & \limsup_{n\to\infty} \frac{1}{n} \sum_{\mathcal{C}} p(\mathcal{C})[1 + P_e^{(n)}(\mathcal{C})nR] \\
= & \limsup_{n\to\infty} \frac{1}{n}[1 + nR \sum_{\mathcal{C}} p(\mathcal{C})P_e^{(n)}(\mathcal{C})] \\
= & \limsup_{n\to\infty} \frac{1}{n} + \limsup_{n\to\infty} R \sum_{\mathcal{C}} p(\mathcal{C})P_e^{(n)}(\mathcal{C}) \\
= & 0.
\end{aligned}$$

Furthermore, it is obvious that $\frac{1}{n}H(X^n|Y^n,\mathbf{C}) \ge 0$ and hence

$$\frac{1}{n}H(X^n|\mathbf{C},Y^n) \to 0, \text{ as } n \to \infty,$$

when $R < I_0(X;Y)$. ■

Proof: [Proof of Lemma 6.3] Given any $\mathcal{C}$, X^n is a function of W. Thus, $H(X^n|\mathcal{C}) \le H(W|\mathcal{C}) = nR,$ and

$$\frac{1}{n}H(X^n|\mathbf{C}) = \frac{1}{n} \sum_{\mathcal{C}} p(\mathcal{C})H(X^n|\mathcal{C}) \le R. \tag{32}$$

Therefore, to show Lemma 6.3, it suffices to show that $\lim_{n\to\infty} \frac{1}{n}H(X^n|\mathbf{C}) \ge R$. For this purpose, we first define a class of codebooks as regular codebooks and focus on characterizing $H(X^n|\mathcal{C})$ for a regular codebook $\mathcal{C}$. Then, we show that a regular codebook appears with high probability when we randomly generate the codebook, and conclude that $\lim_{n\to\infty} \frac{1}{n}H(X^n|\mathbf{C}) \ge R$.

We say a codebook $\mathcal{C}$ is regular if

$$\sup_{x^n \in A_{\epsilon,0}^{(n)}(X)} \left| \frac{N(x^n|\mathcal{C})}{2^{nR}} - p(x^n) \right| \leq \frac{n^3 R}{2^{nR}},$$

where $N(x^n|\mathcal{C})$ is the number of occurrences of x^n in $\mathcal{C}$, defined by

$$N(x^n|\mathcal{C}) = \sum_{w=1}^{2^{nR}} \mathbb{I}(x^n(w) = x^n),$$

and $p(x^n) = \Pr(\tilde{X}^n = x^n | \tilde{X}^n \in A_{\epsilon,0}^{(n)}(X))$ where $\tilde{X}^n$ is drawn i.i.d. according to $p_0(x)$.

Given a regular $\mathcal{C}$, for any $x^n \in A_{\epsilon,0}^{(n)}(X)$, we have

$$\begin{aligned}
N(x^n|\mathcal{C}) \leq & 2^{nR} p(x^n) + n^3 R \\
= & 2^{nR} \Pr(\tilde{X}^n = x^n | \tilde{X}^n \in A_{\epsilon,0}^{(n)}(X)) + n^3 R \\
\leq & 2^{nR}(1 + o(1)) 2^{-n(H_0(X)-\epsilon')} + n^3 R \qquad (33) \\
= & n^3 R + o(1), \qquad (34)
\end{aligned}$$

where the ϵ' in (33) goes to 0 as $\epsilon \to 0$ and (34) follows from the general assumption that $R < H_0(X)$. Note that the message index W is uniformly distributed, we have for a given $\mathcal{C}$ and any $x^n \in A_{\epsilon,0}^{(n)}(X)$,

$$\begin{aligned}
p(x^n|\mathcal{C}) = & \frac{\sum_{w=1}^{2^{nR}} \mathbb{I}(x^n(w) = x^n)}{2^{nR}} \\
= & \frac{N(x^n|\mathcal{C})}{2^{nR}} \\
\leq & \frac{n^3 R + o(1)}{2^{nR}} \\
=: & 2^{-n(R-\epsilon'')}
\end{aligned}$$

where ϵ'' goes to 0 as $n \to \infty$. Therefore,

$$\begin{aligned}
H(X^n|\mathcal{C}) = & \sum_{x^n \in A_{\epsilon,0}^{(n)}(X)} p(x^n|\mathcal{C}) \log \frac{1}{p(x^n|\mathcal{C})} \\
\geq & \sum_{x^n \in A_{\epsilon,0}^{(n)}(X)} p(x^n|\mathcal{C}) \log 2^{n(R-\epsilon'')} \\
= & [n(R-\epsilon'')] \sum_{x^n \in A_{\epsilon,0}^{(n)}(X)} p(x^n|\mathcal{C}) \\
= & [n(R-\epsilon'')].
\end{aligned}$$

Below, We use the Vapnik-Chervonenkis Theorem to show that a regular codebook appears with high probability.

Let $\mathcal{B} = \{\{x^n\}, x^n \in A_{\epsilon,0}^{(n)}(X)\}$. Since $|\mathcal{B}| = |A_{\epsilon,0}^{(n)}(X)| \leq 2^{n(H_0(X)+\epsilon)}$, for any $A \subseteq \mathcal{X}^n$,

$$|\{\{x^n\} \cap A : x^n \in A_{\epsilon,0}^{(n)}(X)\}| \leq 2^{n(H_0(Y)+\epsilon)},$$

and hence VC-d$(\mathcal{B}) \leq n(H_0(X) + \epsilon)$.

Since VC-d$(\mathcal{B})$ is finite for a fixed n, we employ the Vapnik-Chervonenkis Theorem under the range space $(\mathcal{X}^n, \mathcal{B})$. To satisfy (18), let both ϵ and δ in (17) be $\frac{\Delta_\epsilon nR}{2^{nR}}$, where $\Delta_\epsilon := \max\{8\text{VC-d}(\mathcal{B}), 16e\}$. Then the Vapnik-Chervonenkis Theorem states that

$$\begin{aligned}
&\Pr\left\{\sup_{x^n \in A_{\epsilon,0}^{(n)}(X)} \left|\frac{N(x^n|\mathbf{C})}{2^{nR}} - p(x^n)\right| \leq \frac{\Delta_\epsilon nR}{2^{nR}}\right\} \\
\geq & 1 - \frac{\Delta_\epsilon nR}{2^{nR}} \\
\to & 1 \text{ as } n \to \infty. \qquad (35)
\end{aligned}$$

Since $\frac{n^3R}{2^{nR}} \geq \frac{\Delta_\epsilon nR}{2^{nR}}$ for sufficiently large n, (35) concludes that $\Pr(\mathbf{C} \text{ is regular}) \to 1$ as $n \to \infty$.

Therefore,

$$\begin{aligned}
H(X^n|\mathbf{C}) = & \sum_{\mathcal{C}} p(\mathcal{C}) H(X^n|\mathcal{C}) \\
\geq & \sum_{\mathcal{C} \text{ is regular}} p(\mathcal{C}) H(X^n|\mathcal{C}) \\
\geq & [n(R - \epsilon'')] \sum_{\mathcal{C} \text{ is regular}} p(\mathcal{C}) \\
= & [n(R - \epsilon'')](1 - o(1)),
\end{aligned}$$

and

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n} H(X^n|\mathbf{C}) \geq & \lim_{n\to\infty} \frac{1}{n} [n(R - \epsilon'')](1 - o(1)) \\
= & \lim_{n\to\infty} (R - \epsilon'')(1 - o(1)) \\
= & R. \qquad (36)
\end{aligned}$$

Combining (32) and (36), we finish the proof of Lemma 6.3. ■

VII. Rate Needed to Compress Relay's Observation

To study the optimality of the compress-and-forward strategy, in this section, we investigate the rate needed for the relay to losslessly compress its observation. In the classical approach of [2], the compression scheme at the relay was only based on the distribution used for generating the codebook at the source, without being specific on the codebook generated. However, since both the relay and destination have the knowledge of the exact codebook used at the source, it is natural to ask whether it is beneficial for the relay to compress its observation based on this codebook information. This question motivates us to compare the rates needed to compress the relay's observation in two different scenarios: when the relay uses the knowledge of the source's codebook and when the relay simply ignores this knowledge.

Specifically, we consider the two compression problems shown in Figure 1, where Y^n is generated from X^n through the channel $(\mathcal{X}, p(y|x), \mathcal{Y})$, and $\mathbf{C}$ in (b) is the source's codebook information available to both the encoder and decoder. Interestingly, we will show that to perfectly recover Y^n, the minimum required rates in both scenarios are the same when the rate R associated with $\mathbf{C}$ is greater than the channel capacity.

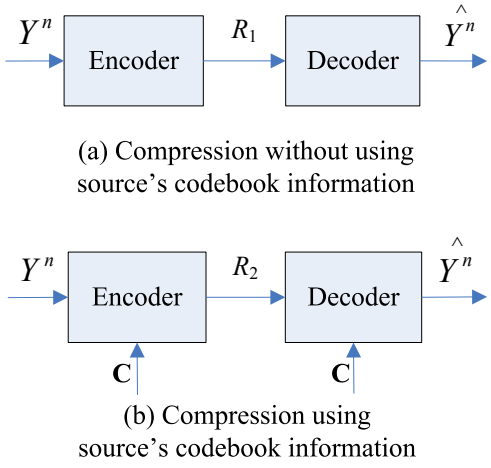

Fig. 1. Two scenarios where the relay compresses its observation.

Formally, we have the following theorem:

Theorem 7.1: For the discrete memoryless channel $(\mathcal{X}, p(y|x), \mathcal{Y})$, generate the codebook at random according to $p_0(x)$ and reserve only the ϵ-strongly typical codewords. Let $\mathbf{C}$ be the source's codebook with rate R, and X^n and Y^n be the input and output of the channel respectively. When $R > I_0(X;Y)$, to compress the channel output Y^n, we have

1) Y^n can be encoded at rate R_1 and recovered with arbitrarily low probability of error if $R_1 > H_0(Y)$.

2) Given that the source's codebook information $\mathbf{C}$ is available to both the encoder and decoder and Y^n is encoded at rate R_2, the decoding probability of error will be bounded away from zero if $R_2 < H_0(Y)$, which implies that we cannot compress the channel output better even if the source's codebook information is employed.

To show Theorem 7.1, we need the following lemma.

Lemma 7.1: For the compression problem in Figure 1-(b), we can encode Y^n at rate R_2 and recover it with the probability of error $P_e^{(n)} \to 0$ only if

$$R_2 \geq \lim_{n\to\infty} \frac{1}{n} H(Y^n|\mathbf{C}). \tag{37}$$

Proof: [Proof of Lemma 7.1] The source code for Figure 1-(b) consists of an encoder mapping $f(Y^n, \mathbf{C})$ and a decoder mapping $g(f(Y^n, \mathbf{C}), \mathbf{C})$. Let $I = f(Y^n, \mathbf{C})$, then $P_e^{(n)} = \Pr(g(I, \mathbf{C}) \neq Y^n)$. By Fano's Inequality, for any source code with $P_e^{(n)} \to 0$, we have

$$H(Y^n|I, \mathbf{C}) \leq P_e^{(n)} \log |\mathcal{Y}^n| + 1 = P_e^{(n)} n \log |\mathcal{Y}| + 1 = n\epsilon_n, \tag{38}$$

where $\epsilon_n \to 0$ as $n \to \infty$.

Therefore, for any source code with rate R_2 and $P_e^{(n)} \to 0$, we have the following chain of inequalities

$$\begin{aligned}
nR_2 &\geq H(I) && (39)\\
&\geq H(I|\mathbf{C}) \\
&= H(Y^n, I|\mathbf{C}) - H(Y^n|I, \mathbf{C}) \\
&= H(Y^n|\mathbf{C}) + H(I|Y^n, \mathbf{C}) - H(Y^n|I, \mathbf{C}) \\
&= H(Y^n|\mathbf{C}) - H(Y^n|I, \mathbf{C}) && (40)\\
&\geq H(Y^n|\mathbf{C}) - n\epsilon_n && (41)
\end{aligned}$$

where (39) follows from the fact that $I \in \{1, 2, \ldots, 2^{nR_2}\}$, (40) follows from the fact that I is a function of Y^n and $\mathbf{C}$, and (41) follows from (38). Dividing the inequality $nR_2 \geq H(Y^n|\mathbf{C}) - n\epsilon_n$ by n and taking the limit as $n \to \infty$, we establish Lemma 7.1. ■

Proof: [Proof of Theorem 7.1]

Proof of Part 1): To show Part 1), we only need to show that the sequence Y^n satisfies the Asymptotic Equipartition Property, i.e., $\Pr(Y^n \in A_{\epsilon,0}^{(n)}(Y)) \to 1$, as $n \to \infty$. Then, following the

classical approach to show the source coding theorem, we can conclude that the rate $R_1 > H_0(Y)$ is achievable. By Lemma 6.1, $\Pr((X^n, Y^n) \in A_{\epsilon,0}^{(n)}(X,Y)) \to 1$ as $n \to \infty$. Thus, the sequence Y^n satisfies the Asymptotic Equipartition Property and the rate $R_1 > H_0(Y)$ is achievable.

Proof of Part 2): By Lemma 7.1, given that the codebook information $\mathbf{C}$ is available to both the encoder and decoder and Y^n is encoded at rate R_2, $P_e^{(n)} \to 0$ only if $R_2 \geq \lim_{n\to\infty} \frac{1}{n} H(Y^n|\mathbf{C})$. By Theorem 3.3, $\lim_{n\to\infty} \frac{1}{n} H(Y^n|\mathbf{C}) = H_0(Y)$ when $R > I_0(X;Y)$. Therefore, when $R > I_0(X;Y)$, $P_e^{(n)} \to 0$ only if $R_2 \geq H_0(Y)$, which establishes Part 2). ■

APPENDIX I
PROOF OF LEMMA 6.1

Proof of Part 1): Let $\tilde{X}^n$ be drawn i.i.d. according to $p_0(x)$ and $\tilde{Y}^n$ be generated from $\tilde{X}^n$ through the channel $(\mathcal{X}, p(y|x), \mathcal{Y})$. Then, we have

$$
\begin{aligned}
&\Pr((X^n, Y^n) \in A_{\epsilon,0}^{(n)}(X,Y)) \\
=&\sum_{(x^n,y^n)\in A_{\epsilon,0}^{(n)}(X,Y)} p(x^n)p(y^n|x^n) \\
=&\sum_{(x^n,y^n)\in A_{\epsilon,0}^{(n)}(X,Y)} \Pr(\tilde{X}^n = x^n|\tilde{X}^n \in A_{\epsilon,0}^{(n)}(X)) \cdot \Pr(Y^n = y^n|X^n = x^n) \\
=&\sum_{(x^n,y^n)\in A_{\epsilon,0}^{(n)}(X,Y)} \Pr(\tilde{X}^n = x^n|\tilde{X}^n \in A_{\epsilon,0}^{(n)}(X)) \cdot \Pr(\tilde{Y}^n = y^n|\tilde{X}^n = x^n) \\
=&\sum_{(x^n,y^n)\in A_{\epsilon,0}^{(n)}(X,Y)} \Pr((\tilde{X}^n, \tilde{Y}^n) = (x^n, y^n)|\tilde{X}^n \in A_{\epsilon,0}^{(n)}(X)) \\
=&\Pr((\tilde{X}^n, \tilde{Y}^n) \in A_{\epsilon,0}^{(n)}(X,Y)|\tilde{X}^n \in A_{\epsilon,0}^{(n)}(X)) \\
=&\frac{\Pr((\tilde{X}^n, \tilde{Y}^n) \in A_{\epsilon,0}^{(n)}(X,Y))}{\Pr(\tilde{X}^n \in A_{\epsilon,0}^{(n)}(X))} \\
\to&1, \text{as } n \to \infty.
\end{aligned}
$$

Proof of Part 2): Denote the ϵ-weakly typical sets with respect to $p_0(x)$, $p_0(y)$ and $p_0(x,y)$ by $W_{\epsilon,0}^{(n)}(X)$, $W_{\epsilon,0}^{(n)}(Y)$ and $W_{\epsilon,0}^{(n)}(X,Y)$ respectively. Along the same line as in the proof of part 1), we can prove that $\Pr((X^n, Y^n) \in W_{\epsilon,0}^{(n)}(X,Y)) \to 1$ as $n \to \infty$, and hence $\Pr(X^n \in W_{\epsilon,0}^{(n)}(X))$ and $\Pr(Y^n \in W_{\epsilon,0}^{(n)}(Y))$ both go to 1 as $n \to \infty$.

Now, consider $H(Y^n)$. We have

$$\begin{aligned} H(Y^n) =& \sum_{y^n} p(y^n) \log \frac{1}{p(y^n)} \\ =& \sum_{y^n \in W_{\epsilon,0}^{(n)}(Y)} p(y^n) \log \frac{1}{p(y^n)} + \sum_{y^n \notin W_{\epsilon,0}^{(n)}(Y)} p(y^n) \log \frac{1}{p(y^n)} \\ =&: \phi_1 + \phi_2 \end{aligned}$$

For ϕ_1, we have

$$\begin{aligned} \phi_1 =& \sum_{y^n \in W_{\epsilon,0}^{(n)}(Y)} p(y^n) \log \frac{1}{p(y^n)} \\ \leq& \sum_{y^n \in W_{\epsilon,0}^{(n)}(Y)} p(y^n) \log 2^{n(H_0(Y)+\epsilon)} \\ =& n(H_0(Y)+\epsilon) \mathrm{Pr}(Y^n \in W_{\epsilon,0}^{(n)}(Y)) \\ =& n(H_0(Y)+\epsilon)(1-o(1)), \end{aligned}$$

where the inequality follows from the fact that $p(y^n) \geq 2^{-n(H_0(Y)+\epsilon)}$ for any $y^n \in W_{\epsilon,0}^{(n)}(Y)$.

For ϕ_2, we have

$$\begin{aligned} \phi_2 =& \sum_{y^n \notin W_{\epsilon,0}^{(n)}(Y)} p(y^n) \log \frac{1}{p(y^n)} \\ =& - \sum_{y^n \in W_{\epsilon,0}^{(n)c}(Y)} p(y^n) \log p(y^n) \\ \leq& - \left(\sum_{y^n \in W_{\epsilon,0}^{(n)c}(Y)} p(y^n) \right) \log \frac{\sum_{y^n \in W_{\epsilon,0}^{(n)c}(Y)} p(y^n)}{|W_{\epsilon,0}^{(n)c}(Y)|} \qquad (42) \\ =& - \mathrm{Pr}(Y^n \notin W_{\epsilon,0}^{(n)}(Y)) \log \frac{\mathrm{Pr}(Y^n \notin W_{\epsilon,0}^{(n)}(Y))}{|W_{\epsilon,0}^{(n)c}(Y)|} \\ =& - \mathrm{Pr}(Y^n \notin W_{\epsilon,0}^{(n)}(Y)) \log \mathrm{Pr}(Y^n \notin W_{\epsilon,0}^{(n)}(Y)) + \mathrm{Pr}(Y^n \notin W_{\epsilon,0}^{(n)}(Y)) \log |W_{\epsilon,0}^{(n)c}(Y)| \\ =& o(1) + \mathrm{Pr}(Y^n \notin W_{\epsilon,0}^{(n)}(Y)) \log |W_{\epsilon,0}^{(n)c}(Y)| \qquad (43) \\ \leq& o(1) + \mathrm{Pr}(Y^n \notin W_{\epsilon,0}^{(n)}(Y)) \log |\mathcal{Y}|^n \\ =& o(1) + n \cdot \mathrm{Pr}(Y^n \notin W_{\epsilon,0}^{(n)}(Y)) \log |\mathcal{Y}| \\ =& n \cdot o(1). \qquad (44) \end{aligned}$$

(42) follows from the the log sum inequality (see Theorem 2.7.1 in [4]), which states that for non-negative numbers, $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$,

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i \right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$$

with equality if and only if $\frac{a_i}{b_i}$ are equal for all i.

(43) and (44) both follow from the fact that $\Pr(Y^n \in W_{\epsilon,0}^{(n)}(Y)) \to 1$ as $n \to \infty$.

Therefore,

$$\begin{aligned} H(Y^n) =& \phi_1 + \phi_2 \\ \leq& n(H_0(Y) + \epsilon)(1 - o(1)) + n \cdot o(1) \\ =& n(H_0(Y) + \epsilon)(1 - o(1)). \end{aligned} \tag{45}$$

Similarly, we have

$$\begin{aligned} H(Y^n) \geq& \sum_{y^n \in W_{\epsilon,0}^{(n)}(Y)} p(y^n) \log \frac{1}{p(y^n)} \\ \geq& \sum_{y^n \in W_{\epsilon,0}^{(n)}(Y)} p(y^n) \log 2^{n(H_0(Y)-\epsilon)} \\ =& n(H_0(Y) - \epsilon) \Pr(Y^n \in W_{\epsilon,0}^{(n)}(Y)) \\ =& n(H_0(Y) - \epsilon)(1 - o(1)). \end{aligned} \tag{46}$$

Combining (45) and (46), we have $\lim_{n\to\infty} \frac{1}{n} H(Y^n) = H_0(Y)$.

Along the same line as above, we can also prove that $\lim_{n\to\infty} \frac{1}{n} H(X^n) = H_0(X)$ and $\lim_{n\to\infty} \frac{1}{n} H(X^n, Y^n) = H_0(X, Y)$, which concludes the proof of Lemma 6.1.

REFERENCES

[1] X. Wu and L.-L. Xie, "AEP of output when rate is above capacity," in *Proc. of the 11th Canadian Workshop on Information Theory*, Ottawa, Canada, May 13-15, 2009.

[2] T. Cover and A. El Gamal, "Capacity theorems for the relay channel," *IEEE Trans. Inform. Theory*, vol. 25, pp. 572–584, 1979.

[3] F. Xue, P. R. Kumar and L.-L. Xie, "The conditional entropy of the jointly typical set when coding rate is larger than Shannon Capacity," Manuscript, 2006.

[4] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[5] V. N. Vapnik and A. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory of Probability and its Applications*, vol. 16, no. 2, pp. 264–280, Jan. 1971.

[6] V. N. Vapnik, *Estimation of dependences based on empirical data*. New York: Springer-Verlag, 1982.