The Likelihood Encoder for Lossy Compression

Eva C. Song, *Student Member, IEEE,* Paul Cuff, *Member, IEEE,* and H. Vincent Poor, *Fellow, IEEE*

Abstract

In this work, a likelihood encoder is studied in the context of lossy source compression. The analysis of the likelihood encoder is based on the soft-covering lemma. It is demonstrated that the use of a likelihood encoder together with the soft-covering lemma yields simple achievability proofs for classical source coding problems. The cases of the point-to-point rate-distortion function, the rate-distortion function with side information at the decoder (i.e. the Wyner-Ziv problem), and the multi-terminal source coding inner bound (i.e. the Berger-Tung problem) are examined in this paper. Furthermore, a non-asymptotic analysis is used for the point-to-point case to examine the upper bound on the excess distortion provided by this method. The likelihood encoder is also compared, both in concept and performance, to a recent alternative technique using properties of random binning.

Index Terms

Berger-Tung, likelihood encoder, rate-distortion theory, soft-covering, source coding, Wyner-Ziv

I. Introduction

Rate-distortion theory, founded by Shannon in [1] and [2], provides the fundamental limits of lossy source compression. The minimum rate required to represent an independent and identically

The authors are with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA e-mail: {csong, cuff, poor}@princeton.edu.

This paper was presented in part at 2013 IEEE Information Theory Workshop (ITW) and 2014 IEEE International Symposium on Information Theory (ISIT).

distributed (i.i.d.) source sequence under a given tolerance of distortion is given by the rate-distortion function. Related problems such as source coding with side information available at the decoder [3] and distributed source coding [4], [5], [6] have also been heavily studied in the past decades. Standard proofs [7], [8] of achievability for these rate-distortion problems often use joint-typicality encoding, i.e. the encoder looks for a codeword that is jointly typical with the source sequence. The distortion analysis involves bounding several "error" events which may come from either encoding or decoding. These bounds use the joint asymptotic equipartition principle (J-AEP) and its immediate consequences as the main tool. In the cases where there are multiple information sources, such as side information at the decoder, intricacies arise, such as the need for a Markov lemma [7], [8]. These subtleties also lead to error-prone proofs involving the analysis of error caused by random binning, which have been pointed out in several existing works [9], [10].

In this work, we propose using a likelihood encoder to achieve these source coding results. The likelihood encoder is a stochastic encoder. As stated in [11], for a chosen joint distribution P_{XY}, to encode a source sequence $x_1, ..., x_n$ (i.e. x^n) with codebook $y^n(m)$, the encoder stochastically chooses an index m proportional to the likelihood of $y^n(m)$ passed through the memoryless "test channel" $P_{X|Y}$.

The advantage of using such an encoder is that it naturally leads to an idealized distribution which is simple to analyze, based on the "test channel." The distortion performance of the idealized distribution carries over to the true system induced distribution because the two distributions are shown to be close in total variation. Unlike the proof using the joint-typicality encoder, we do not need to identify different kinds of "error" events – the distortion analysis of the idealized distribution is straightforward.

This proof technique of using an idealized approximating distribution gives a macroscopic analysis of the system. Precise behaviors of the system are illuminated through the approximating distributions. In other contexts, beyond the scope of this paper, this feature of the proof method can greatly simplify the analysis of secrecy and other objectives which demand comprehensive characterization of the behavior of the system. In this paper we demonstrate this technique in more basic settings of source coding, showing its effectiveness in simplifying and illuminating even those proofs.

Just as the joint-typicality encoder relies on the J-AEP, the likelihood encoder relies on the soft-covering lemma. The idea of soft-covering was first introduced in [12] and was later used in [13] for channel resolvability. The use of the likelihood encoder in conjunction with the soft-covering lemma appeared in [14] and [15] to achieve strong coordination and also in [16] for secrecy.

The application of the likelihood encoder together with the soft-covering lemma is not limited to only discrete alphabets. The proof for sources from continuous alphabets is readily included, since the soft-covering lemma imposes no restriction on alphabet size. Therefore, no extra work, i.e. quantization of the source, is needed to extend the standard proof for discrete sources to continuous sources as in [8]. This advantage becomes more pronounced for the multi-terminal case, since generalization of the type-covering lemma and the Markov lemma to continuous alphabets is non-trivial. Strong versions of the Markov lemma on finite alphabets that can prove the Berger-Tung inner bound can be found in [8] and [17]. However, generalization to continuous alphabets is still an ongoing research topic. Some works, such as [18] and [19], have been dedicated to making this transition, yet are not strong enough to be applied to the Berger-Tung case.

The rest of the paper is organized as follows. In Section II, we will introduce notation, some basic concepts and properties, define the likelihood encoder and give the soft-covering lemma. Sections III to V deal with the point-to-point rate-distortion, Wyner-Ziv, and Berger-Tung problems, respectively, with increasing complexity. Within each of these sections, we first review the problem setup along with the result, and then give the achievability proof using the likelihood encoder. In Section VI, we apply a non-asymptotic analysis to the excess distortion for the point-to-point case. In Section VII, we compare the performance of the likelihood encoder to a proportional-probability encoder [20], whose analysis is based on random-binning, in both the asymptotic and non-asymptotic senses. Finally, in Section VIII, we summarize the work.

II. Preliminaries

A. Notation

A sequence $X_1, ..., X_n$ is denoted by X^n. Limits taken with respect to "$n \to \infty$" are abbreviated as "$\to_n$". Inequalities with $\limsup_{n\to\infty} h_n \leq h$ and $\liminf_{n\to\infty} h_n \geq h$ are abbreviated as $h_n \leq_n h$ and $h_n \geq_n h$, respectively. When X denotes a random variable, x is used to denote a realization, $\mathcal{X}$ is used to denote the support of that random variable, and $\Delta_{\mathcal{X}}$ is used to denote the probability simplex of distributions with alphabet $\mathcal{X}$. A Markov relation is denoted by the symbol $-$. We use $\mathbb{E}_P$, $\mathbb{P}_P$, and $I_P(X;Y)$ to indicate expectation, probability, and mutual information taken with respect to a distribution P; however, when the distribution is clear from the context, the subscript will be omitted. To keep the notation uncluttered, the arguments of a distribution are sometimes omitted when the arguments' symbols match the subscripts of the distribution, e.g. $P_{X|Y}(x|y) = P_{X|Y}$. We use a bold capital letter $\mathbf{P}$ to denote that a distribution P is random. We use $\mathbb{R}$ to denote the set of real numbers and $\mathbb{R}^+$ to denote the nonnegative subset.

For a distortion measure $d : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}^+$, we use $\mathbb{E}[d(X,Y)]$ to measure the distortion of X incurred by representing it as Y. The maximum distortion is defined as

$$d_{max} = \max_{(x,y)\in\mathcal{X}\times\mathcal{Y}} d(x,y).$$

The distortion between two sequences is defined to be the per-letter average distortion

$$d(x^n, y^n) = \frac{1}{n}\sum_{t=1}^{n} d(x_t, y_t).$$

B. *Total Variation Distance*

The total variation distance between two probability measures P and Q on the same σ-algebra $\mathcal{F}$ of subsets of the sample space $\mathcal{X}$ is defined as

$$\|P - Q\|_{TV} \triangleq \sup_{\mathcal{A}\in\mathcal{F}} |P(\mathcal{A}) - Q(\mathcal{A})|.$$

Property 1 (Property 2 [16])**.** *Total variation distance satisfies the following properties:*

(a) If $\mathcal{X}$ is countable, then total variation can be rewritten as

$$\|P - Q\|_{TV} = \frac{1}{2}\sum_{x\in\mathcal{X}} |p(x) - q(x)|, \tag{1}$$

where $p(\cdot)$ and $q(\cdot)$ are the probability mass functions of X under P and Q, respectively.

(b) Let $\varepsilon > 0$ and let $f(x)$ be a function in a bounded range with width $b \in \mathbb{R}^+$. Then

$$\|P - Q\|_{TV} < \varepsilon \implies \left|\mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(X)]\right| < \varepsilon b. \tag{2}$$

(c) Total variation satisfies the triangle inequality. For any $S \in \Delta_{\mathcal{X}}$,

$$\|P - Q\|_{TV} \le \|P - S\|_{TV} + \|S - Q\|_{TV}. \tag{3}$$

(d) Let $P_X P_{Y|X}$ *and* $Q_X P_{Y|X}$ *be two joint distributions on* $\Delta_{\mathcal{X}\times\mathcal{Y}}$*. Then*

$$\|P_X P_{Y|X} - Q_X P_{Y|X}\|_{TV} = \|P_X - Q_X\|_{TV}. \tag{4}$$

(e) For any $P, Q \in \Delta_{\mathcal{X}\times\mathcal{Y}}$*,*

$$\|P_X - Q_X\|_{TV} \leq \|P_{XY} - Q_{XY}\|_{TV}. \tag{5}$$

C. The Likelihood Encoder

We now define the likelihood encoder, operating at rate R, which receives a sequence $x_1, ..., x_n$ and maps it to a message $M \in [1 : 2^{nR}]$. In normal usage, a decoder will then use M to form an approximate reconstruction of the $x_1, ..., x_n$ sequence.

The encoder is specified by a codebook of $y^n(m)$ sequences and a joint distribution P_{XY}. Consider the likelihood function for each codeword, with respect to a memoryless channel from Y to X, defined as follows:

$$\mathcal{L}(m|x^n) \triangleq P_{X^n|Y^n}(x^n|y^n(m)).$$

A likelihood encoder is a stochastic encoder that determines the message index with probability proportional to $\mathcal{L}(m|x^n)$, i.e.

$$P_{M|X^n}(m|x^n) = \frac{\mathcal{L}(m|x^n)}{\sum_{m' \in [1:2^{nR}]} \mathcal{L}(m'|x^n)} \propto \mathcal{L}(m|x^n).$$

D. Soft-Covering Lemma

Now we introduce the core lemma that serves as the foundation for this analysis. One can consider the role of the soft-covering lemma in analyzing the likelihood encoder as analogous to that of the J-AEP which is used for the analysis of joint-typicality encoders. The general idea of the soft-covering lemma is that the distribution induced by selecting uniformly from a random codebook and passing the codeword through a memoryless channel is close to an i.i.d. distribution as long as the codebook size is large enough.

Lemma 1 (Lemma IV.1 [15])**.** *Given a joint distribution* P_{XY}*, let* $\mathcal{C}^{(n)}$ *be a random collection of sequences* $Y^n(m)$*, with* $m = 1, ..., 2^{nR}$*, each drawn independently and i.i.d. according to* P_Y*.*

Denote by P_{X^n} the output distribution induced by selecting an index m uniformly at random and applying $Y^n(m)$ to the memoryless channel specified by $P_{X|Y}$. Then if $R > I(X;Y)$,

$$\mathbb{E}_{\mathcal{C}^n}\left[\left\|P_{X^n} - \prod_{t=1}^{n} P_X\right\|_{TV}\right] \to_n 0.$$

Next, we will use the soft-covering lemma to obtain simple achievability proofs for the rate-distortion function, the Wyner-Ziv problem, and the Berger-Tung inner bound for distributed source coding.

III. The Point-to-Point Rate-Distortion Problem

Let us first start with point-to-point lossy compression, which was presented also in [11]. This simple setting outlines the key steps in the analysis, which will be applied again for the more complex settings.

A. Problem Setup and Result Review

Rate-distortion theory determines the optimal compression rate R for an i.i.d. source sequence X^n distributed according to $X_t \sim \overline{P}_X$ with the following constraints:

- Encoder $f_n : \mathcal{X}^n \mapsto \mathcal{M}$ (possibly stochastic);
- Decoder $g_n : \mathcal{M} \mapsto \mathcal{Y}^n$ (possibly stochastic);
- Compression rate: R, i.e. $|\mathcal{M}| = 2^{nR}$.

The system performance is measured according to the time-averaged distortion (as defined in the notation section):

- Average distortion: $d(X^n, Y^n) = \frac{1}{n}\sum_{t=1}^{n} d(X_t, Y_t)$.

Definition 1. *A rate distortion pair (R, D) is achievable if there exists a sequence of rate R encoders and decoders (f_n, g_n), such that $\mathbb{E}[d(X^n, Y^n)] \leq_n D$.*

Definition 2. *The rate distortion function is $R(D) \triangleq \inf_{\{(R,D) \text{ is achievable}\}} R$.*

The above mathematical formulation is illustrated in Fig. 1. The characterization of this fundamental quantity in information theory is given in [7] as

$$R(D) = \min_{\overline{P}_{Y|X}:\mathbb{E}[d(X,Y)]\leq D} I_{\overline{P}}(X;Y), \tag{6}$$

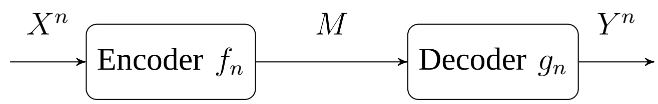

Fig. 1: Point-to-point lossy compression setup

where the mutual information is taken with respect to $\overline{P}_{XY} = \overline{P}_X \overline{P}_{Y|X}$. In other words, we are able to achieve distortion level D with any rate less than $R(D)$ given in (6).

B. Achievability Proof Using the Likelihood Encoder

To prove achievability, we will use the likelihood encoder and approximate the overall behavior of the system by a well-behaved distribution. The soft-covering lemma allows us to claim that the approximating distribution matches the system.

Let $R > R(D)$, where $R(D)$ is from (6). We prove that R is achievable for distortion D. By the rate-distortion formula stated in (6), we can fix $\overline{P}_{Y|X}$ such that $R > I_{\overline{P}}(X;Y)$ and $\mathbb{E}_{\overline{P}}[d(X,Y)] < D$. We will use the likelihood encoder derived from $\overline{P}_{XY}$ and a random codebook $\{y^n(m)\}$ generated according to $\overline{P}_Y$ to prove the result. The decoder will simply reproduce $y^n(M)$ upon receiving the message M.

The distribution induced by the encoder and decoder is

$$\begin{aligned} & \mathbf{P}_{X^n M Y^n}(x^n, m, y^n) \\ = \; & \overline{P}_{X^n}(x^n)\mathbf{P}_{M|X^n}(m|x^n)\mathbf{P}_{Y^n|M}(y^n|m) \qquad (7) \\ \triangleq \; & \overline{P}_{X^n}(x^n)\mathbf{P}_{LE}(m|x^n)\mathbf{P}_{D}(y^n|m) \qquad (8) \end{aligned}$$

where $\mathbf{P}_{LE}$ is the likelihood encoder and $\mathbf{P}_D$ is a codeword lookup decoder.

We now concisely restate the behavior of the encoder and decoder, as components of the induced distribution.

Codebook generation: We independently generate 2^{nR} sequences in $\mathcal{Y}^n$ according to $\prod_{i=1}^n \overline{P}_Y(y_i)$ and index them by $m \in [1:2^{nR}]$. We use $\mathcal{C}^{(n)}$ to denote the random codebook.

Encoder: The encoder $\mathbf{P}_{LE}(m|x^n)$ is the likelihood encoder that chooses M stochastically with probability proportional to the likelihood function given by

$$\mathcal{L}(m|x^n) = \overline{P}_{X^n|Y^n}(x^n|Y^n(m)).$$

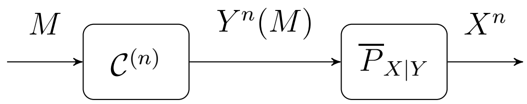

Fig. 2: Idealized distribution with test channel $\overline{P}_{Y|X}$

Decoder: The decoder $\mathbf{P}_D(y^n|m)$ is a codeword lookup decoder that simply reproduces $Y^n(m)$.

Analysis: We will consider two distributions for the analysis, the induced distribution $\mathbf{P}$ and an approximating distribution $\mathbf{Q}$, which is much easier to analyze. We will show that $\mathbf{P}$ and $\mathbf{Q}$ are close in total variation (on average over the random codebook). Hence, $\mathbf{P}$ achieves the performance of $\mathbf{Q}$.

Design the approximating distribution $\mathbf{Q}$ via a uniform distribution over a random codebook and a test channel $\overline{P}_{X|Y}$ as shown in Fig. 2. We will refer to a distribution of this structure as an idealized distribution. The joint distribution under the idealized distribution $\mathbf{Q}$ shown in Fig. 2 can be written as

$$\begin{aligned}
& \mathbf{Q}_{X^nY^nM}(x^n, y^n, m) \\
= \; & Q_M(m)\mathbf{Q}_{Y^n|M}(y^n|m)\mathbf{Q}_{X^n|M}(x^n|m) && (9) \\
= \; & \frac{1}{2^{nR}}\mathbb{1}\{y^n = Y^n(m)\}\prod_{t=1}^{n}\overline{P}_{X|Y}(x_t|Y_t(m)) && (10) \\
= \; & \frac{1}{2^{nR}}\mathbb{1}\{y^n = Y^n(m)\}\prod_{t=1}^{n}\overline{P}_{X|Y}(x_t|y_t). && (11)
\end{aligned}$$

The idealized distribution $\mathbf{Q}$ has the following property: for any $(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$,

$$\begin{aligned}
& \mathbb{E}_{\mathcal{C}^{(n)}}[\mathbf{Q}_{X^nY^n}(x^n, y^n)] \\
= \; & \mathbb{E}_{\mathcal{C}^{(n)}}\left[\frac{1}{2^{nR}}\sum_m \mathbb{1}\{y^n = Y^n(m)\}\right]\prod_{t=1}^{n}\overline{P}_{X|Y}(x_t|y_t) \\
= \; & \frac{1}{2^{nR}}\sum_m \mathbb{E}_{\mathcal{C}^{(n)}}[\mathbb{1}\{y^n = Y^n(m)\}]\prod_{t=1}^{n}\overline{P}_{X|Y}(x_t|y_t) \\
= \; & \frac{1}{2^{nR}}\sum_m \overline{P}_{Y^n}(y^n)\prod_{t=1}^{n}\overline{P}_{X|Y}(x_t|y_t) \\
= \; & \overline{P}_{X^nY^n}(x^n, y^n) && (12)
\end{aligned}$$

where $\overline{P}_{X^nY^n}$ denotes the i.i.d. distribution $\prod_{t=1}^{n} \overline{P}_{XY}$. This implies, in particular, that the distortion under the idealized distribution $\mathbf{Q}$ averaged over the random codebook, conveniently simplifies to $\mathbb{E}_{\overline{P}}[d(X,Y)]$. That is,

$$\begin{aligned}
& \mathbb{E}_{\mathcal{C}^{(n)}}\left[\mathbb{E}_{\mathbf{Q}}[d(X^n,Y^n)]\right] \\
= \; & \mathbb{E}_{\mathcal{C}^{(n)}}\left[\sum_{x^n,y^n} \mathbf{Q}(x^n,y^n) d(x^n,y^n)\right] && (13)\\
= \; & \sum_{x^n,y^n} \mathbb{E}_{\mathcal{C}^{(n)}}[\mathbf{Q}(x^n,y^n)] d(x^n,y^n) && (14)\\
= \; & \sum_{x^n,y^n} \overline{P}_{X^n,Y^n}(x^n,y^n) d(x^n,y^n) && (15)\\
= \; & \mathbb{E}_{\overline{P}}[d(X^n,Y^n)] && (16)\\
= \; & \mathbb{E}_{\overline{P}}[d(X,Y)], && (17)
\end{aligned}$$

where (15) follows from (12). It is worth emphasizing that although $\mathbf{Q}$ is very different from the i.i.d. distribution on (X^n, Y^n), it is exactly the i.i.d. distribution when averaged over codebooks and thus achieves the same expected distortion.

Our motivation for using the likelihood encoder comes from this construction of $\mathbf{Q}$. Notice that

$$\mathbf{Q}_{M|X^n}(m|x^n) = \mathbf{P}_{LE}(m|x^n), \tag{18}$$

and

$$\mathbf{Q}_{Y^n|M}(y^n|m) = \mathbf{P}_{D}(y^n|m). \tag{19}$$

Now invoking the soft-covering lemma, since $R > I_{\overline{P}}(X;Y)$, we have

$$\mathbb{E}_{\mathcal{C}^{(n)}}\left[\|\overline{P}_{X^n} - \mathbf{Q}_{X^n}\|_{TV}\right] \leq \epsilon_n,$$

where $\epsilon_n \to_n 0$. This gives us

$$\begin{aligned}
& \mathbb{E}_{\mathcal{C}^{(n)}}\left[\|\mathbf{P}_{X^nY^n} - \mathbf{Q}_{X^nY^n}\|_{TV}\right] \\
\leq \; & \mathbb{E}_{\mathcal{C}^{(n)}}\left[\|\mathbf{P}_{X^nY^nM} - \mathbf{Q}_{X^nY^nM}\|_{TV}\right] && (20)\\
\leq \; & \epsilon_n, && (21)
\end{aligned}$$

where (20) follows from Property $1(e)$ and (21) follows from (18),(19) and Property $1(d)$.

By Property $1(b)$,

$$|\mathbb{E}_{\mathbf{P}}[d(X^n,Y^n)] - \mathbb{E}_{\mathbf{Q}}[d(X^n,Y^n)]| \leq d_{max}\|\mathbf{P}-\mathbf{Q}\|_{TV}. \tag{22}$$

Now we apply the random coding argument.

$$\begin{aligned}
& \mathbb{E}_{\mathcal{C}^{(n)}}\left[\mathbb{E}_{\mathbf{P}}[d(X^n,Y^n)]\right] \\
\leq\ & \mathbb{E}_{\mathcal{C}^{(n)}}\left[\mathbb{E}_{\mathbf{Q}}[d(X^n,Y^n)]\right] + \mathbb{E}_{\mathcal{C}^{(n)}}\left[|\mathbb{E}_{\mathbf{P}}[d(X^n,Y^n)] - \mathbb{E}_{\mathbf{Q}}[d(X^n,Y^n)]|\right] && (23)\\
\leq\ & \mathbb{E}_{\overline{P}}[d(X,Y)] + d_{max}\mathbb{E}_{\mathcal{C}^{(n)}}\left[\|\mathbf{P}_{X^nY^n} - \mathbf{Q}_{X^nY^n}\|_{TV}\right] && (24)\\
\leq\ & \mathbb{E}_{\overline{P}}[d(X,Y)] + d_{max}\epsilon_n && (25)\\
\leq_n\ & D, && (26)
\end{aligned}$$

where (24) follows from (17) and (22); (25) follows from (21). Therefore, there exists a codebook satisfying the requirement. ■

Remark 1. As the proof emphasizes, the distribution $\mathbf{Q}$ serves as an accurate approximation to the true system behavior, and this is not unique to the likelihood encoder. In [21] a converse statement is shown. That is, any efficient source encoding satisfying a distortion constraint behaves like $\mathbf{Q}$ as measured by normalized divergence. However, a stochastic encoder is generally required for the approximation to hold in total variation. Furthermore, for the likelihood encoder, the accuracy of this approximation is easily verified using the soft-covering lemma. For other encoders, the proof requires more effort to establish.

IV. The Wyner-Ziv Problem

In this section, we will use the mechanism that was established in Section III and build up on it to solve a more complicated problem. The Wyner-Ziv problem, that is, the rate-distortion function with side information at the decoder, was solved in [3].

A. Problem Setup and Result Review

The source and side information pair (X^n, Z^n) is distributed i.i.d. according to $(X_t, Z_t) \sim \overline{P}_{XZ}$. The system has the following constraints:

- Encoder $f_n : \mathcal{X}^n \mapsto \mathcal{M}$ (possibly stochastic);
- Decoder $g_n : \mathcal{M} \times \mathcal{Z}^n \mapsto \mathcal{Y}^n$ (possibly stochastic);

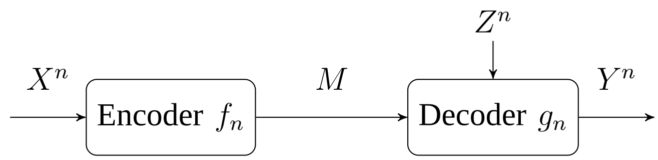

Fig. 3: Rate-distortion theory for source coding with side information at the decoder—the Wyner-Ziv problem

- Compression rate: R, i.e. $|\mathcal{M}| = 2^{nR}$.

The system performance is measured according to the time-averaged distortion (as defined in the notation section):

- Average distortion: $d(X^n, Y^n) = \frac{1}{n}\sum_{t=1}^{n} d(X_t, Y_t)$.

Definition 3. *A rate distortion pair (R, D) is achievable if there exists a sequence of rate R encoders and decoders (f_n, g_n), such that $\mathbb{E}\left[d(X^n, Y^n)\right] \leq_n D$.*

Definition 4. *The rate distortion function is $R(D) \triangleq \inf_{\{(R,D) \text{ is achievable}\}} R$.*

The above mathematical formulation is illustrated in Fig. 3.

As mentioned previously, the solution to this source coding problem is given in [3]. The rate-distortion function with side information at the decoder is

$$R(D) \;=\; \min_{\overline{P}_{V|XZ} \in \mathcal{M}(D)} I_{\overline{P}}(X; V|Z), \tag{27}$$

where

$$\mathcal{M}(D) = \Big\{ \overline{P}_{V|XZ} : \quad V - X - Z, \quad |\mathcal{V}| \leq |\mathcal{X}| + 1, \quad \text{and there exists a function } \phi \text{ s.t. } \mathbb{E}\left[d(X,Y)\right] \leq D, Y \triangleq \phi(V, Z) \Big\}. \tag{28}$$

B. Achievability Proof Using the Likelihood Encoder

Before going into the main proof, let us first establish a property of total variation that will be helpful for both the Wyner-Ziv problem and the Berger-Tung inner bound.

Lemma 2. *For a distribution P_{UVX} and $0 < \varepsilon < 1$, if $\mathbb{P}[U \neq V] \leq \varepsilon$, then*

$$\|P_{UX} - P_{VX}\|_{TV} \leq \varepsilon.$$

Proof: By definition,

$$\|P_{UX} - P_{VX}\|_{TV} = \sup_{\mathcal{A} \in \mathcal{F}} \left\{ \mathbb{P}[(U,X) \in \mathcal{A}] - \mathbb{P}[(V,X) \in \mathcal{A}] \right\}.$$

Since for every $\mathcal{A} \in \mathcal{F}$

$$\begin{aligned} & \mathbb{P}[(U,X) \in \mathcal{A}] - \mathbb{P}[(V,X) \in \mathcal{A}] \\ \leq \; & \mathbb{P}[(U,X) \in \mathcal{A}] - \mathbb{P}[(V,X) \in \mathcal{A}, (U,X) \in \mathcal{A}] && (29) \\ = \; & \mathbb{P}[(U,X) \in \mathcal{A}, (V,X) \neq \mathcal{A}] && (30) \\ \leq \; & \mathbb{P}[U \neq V] && (31) \\ \leq \; & \epsilon, && (32) \end{aligned}$$

we have

$$\sup_{\mathcal{A} \in \mathcal{F}} \left\{ \mathbb{P}[(U,X) \in \mathcal{A}] - \mathbb{P}[(V,X) \in \mathcal{A}] \right\} \leq \epsilon.$$

■

We are now ready to give the achievability proof of (27). We will introduce a virtual message which is produced by the encoder but not physically transmitted to the receiver so that this virtual message together with the actual message gives a high enough rate for applying the soft-covering lemma. Then we show that this virtual message can be reconstructed with vanishing error probability at the decoder by using the side information. This is analogous to the technique of random binning, where the index of the codeword within the bin is equivalent to the virtual message in our method.

Our proof technique again involves showing that the behavior of the system is approximated by a well-behaved distribution. The soft-covering lemma and channel decoding error bounds are used to analyze how well the approximating distribution matches the system.

Let $R > R(D)$, where $R(D)$ is from (27). We prove that R is achievable for distortion D. Let M' be a virtual message with rate R' which is not physically transmitted. By the rate-distortion formula in (27), we can fix R' and $\overline{P}_{V|XZ} \in \mathcal{M}(D)$ ($\overline{P}_{V|XZ} = \overline{P}_{V|X}$) such that $R + R' > I_{\overline{P}}(X;V)$ and $R' < I_{\overline{P}}(V;Z)$, and there exists a function $\phi(\cdot,\cdot)$ yielding $Y = \phi(V,Z)$

and $\mathbb{E}[d(X,Y)] \leq D$. We will use the likelihood encoder derived from $\overline{P}_{XV}$ and a random codebook $\{v^n(m,m')\}$ generated according to $\overline{P}_V$ to prove the result. The decoder will first use the transmitted message M and the side information Z^n to decode M' as $\hat{M}'$ and reproduce $v^n(M,\hat{M}')$. Then the reconstruction Y^n is produced as a symbol-by-symbol application of $\phi(\cdot,\cdot)$ to Z^n and V^n.

The distribution induced by the encoder and decoder is

$$\begin{aligned} & \mathbf{P}_{X^nZ^nMM'\hat{M}'Y^n}(x^n,z^n,m,m',\hat{m}',y^n) \\ = \; & \overline{P}_{X^nZ^n}(x^n,z^n)\mathbf{P}_{MM'|X^n}(m,m'|x^n)\mathbf{P}_{\hat{M}'|MZ^n}(\hat{m}'|m,z^n)\mathbf{P}_{Y^n|M\hat{M}'Z^n}(y^n|m,\hat{m}',z^n) \quad (33) \\ \triangleq \; & \overline{P}_{X^nZ^n}(x^n,z^n)\mathbf{P}_{LE}(m,m'|x^n)\mathbf{P}_D(\hat{m}'|m,z^n)\mathbf{P}_\Phi(y^n|m,\hat{m}',z^n), \quad (34) \end{aligned}$$

where $\mathbf{P}_{LE}(m,m'|x^n)$ is the likelihood encoder; $\mathbf{P}_D(\hat{m}'|m,z^n)$ is the first part of the decoder that decodes m' as $\hat{m}'$; and $\mathbf{P}_\Phi(y^n|m,\hat{m}',z^n)$ is the second part of the decoder that reconstructs the source sequence.

We now concisely restate the behavior of the encoder and decoder, as these components of the induced distribution.

Codebook generation: We independently generate $2^{n(R+R')}$ sequences in $\mathcal{V}^n$ according to $\prod_{i=1}^n \overline{P}_V(v_i)$ and index by $(m,m') \in [1:2^{nR}] \times [1:2^{nR'}]$. We use $\mathcal{C}^{(n)}$ to denote the random codebook.

Encoder: The encoder $\mathbf{P}_{LE}(m,m'|x^n)$ is the likelihood encoder that chooses M and M' stochastically with probability proportional to the likelihood function given by

$$\mathcal{L}(m,m'|x^n) = \overline{P}_{X^n|V^n}(x^n|V^n(m,m')).$$

Decoder: The decoder has two steps. Let $\mathbf{P}_D(\hat{m}'|m,z^n)$ be a good channel decoder (e.g. the maximum likelihood decoder) with respect to the sub-codebook $\mathcal{C}^{(n)}(m) = \{v^n(m,a)\}_a$ and the memoryless channel $\overline{P}_{Z|V}$. For the second part of the decoder, let $\phi(\cdot,\cdot)$ be the function corresponding with the choice of $\overline{P}_{V|XZ}$ in (28); that is, $Y = \phi(V,Z)$ and $\mathbb{E}_{\overline{P}}[d(X,Y)] \leq D$. Define $\phi^n(v^n,z^n)$ as the concatenation $\{\phi(v_t,z_t)\}_{t=1}^n$ and set the decoder $\mathbf{P}_\Phi$ to be the deterministic function

$$\mathbf{P}_\Phi(y^n|m,\hat{m}',z^n) \triangleq \mathbb{1}\{y^n = \phi^n(V^n(m,\hat{m}'),z^n)\}.$$

Analysis: We will consider three distributions for the analysis, the induced distribution $\mathbf{P}$ and two approximating distributions $\mathbf{Q}^{(1)}$ and $\mathbf{Q}^{(2)}$. The idea is to show that 1) the system has nice

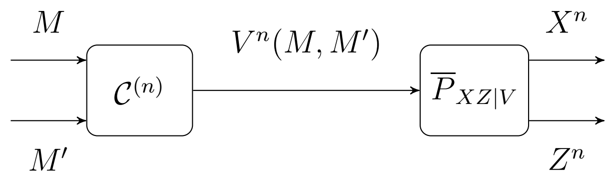

Fig. 4: Idealized distribution with test channel $\overline{P}_{XZ|V}$

behavior for distortion under $\mathbf{Q}^{(2)}$; and 2) $\mathbf{P}$ and $\mathbf{Q}^{(2)}$ are close in total variation (on average over the random codebook) through $\mathbf{Q}^{(1)}$.

The first approximating distribution, $\mathbf{Q}^{(1)}$, changes the distribution induced by the likelihood encoder to a distribution based on a reverse memoryless channel, as in the proof of point-to-point rate-distortion theory, and shown below in Fig. 4. This is shown to be a good approximation using the soft-covering lemma. The second approximating distribution, $\mathbf{Q}^{(2)}$, pretends that M', the index which is not transmitted, is used by the decoder to form the reconstruction. This is a good approximation because the decoder can accurately estimate M'.

Both approximating distributions $\mathbf{Q}^{(1)}$ and $\mathbf{Q}^{(2)}$ are builded upon the idealized distribution over the information sources and messages, according to the test channel, as shown in Fig. 4. Note that this idealized distribution $\mathbf{Q}$ is no different from the one we considered for the point-to-point case, except for the message indices. The joint distribution under $\mathbf{Q}$ in Fig. 4 can be written as

$$\begin{aligned}
&\mathbf{Q}_{X^nZ^nV^nMM'}(x^n, z^n, v^n, m, m') \\
=\ & Q_{MM'}(m,m')\mathbf{Q}_{V^n|MM'}(v^n|m,m')\mathbf{Q}_{X^nZ^n|MM'}(x^n,z^n|m,m') && (35)\\
=\ & \frac{1}{2^{n(R+R')}}\mathbb{1}\{v^n = V^n(m,m')\}\prod_{t=1}^{n}\overline{P}_{XZ|V}(x_t,z_t|V_t(m,m')) && (36)\\
=\ & \frac{1}{2^{n(R+R')}}\mathbb{1}\{v^n = V^n(m,m')\}\prod_{t=1}^{n}\overline{P}_{X|V}(x_t|v_t)\overline{P}_{Z|X}(z_t|x_t), && (37)
\end{aligned}$$

where (37) follows from the Markov chain under $\overline{P}$, $V - X - Z$. Note that by using the likelihood encoder, the idealized distribution $\mathbf{Q}$ satisfies

$$\mathbf{Q}_{MM'|X^nZ^n}(m,m'|x^n,z^n) = \mathbf{P}_{LE}(m,m'|x^n). \tag{38}$$

Furthermore, using the same technique as (12) and (17) given in the previous section, it can be verified that

$$\mathbb{E}_{\mathcal{C}^{(n)}}\left[\mathbf{Q}_{X^nZ^nV^n}(x^n,z^n,v^n)\right] = \overline{P}_{X^nZ^nV^n}(x^n,z^n,v^n), \tag{39}$$

where $\overline{P}_{X^nZ^nV^n}$ denotes the i.i.d. distribution $\prod_{t=1}^n \overline{P}_{XZV}$. Consequently,

$$\mathbb{E}_{\mathcal{C}^{(n)}}\left[\mathbb{E}_{\mathbf{Q}}\left[d\left(X^n,\phi^n(V^n,Z^n)\right)\right]\right] = \mathbb{E}_{\overline{P}}\left[d\left(X^n,\phi^n(V^n,Z^n)\right)\right].$$

Define the two distributions $\mathbf{Q}^{(1)}$ and $\mathbf{Q}^{(2)}$ based on $\mathbf{Q}$ as follows:

$$\begin{aligned} & \mathbf{Q}^{(1)}_{X^nZ^nMM'\hat{M}'Y^n}(x^n,z^n,m,m',\hat{m}',y^n) \\ \triangleq\; & \mathbf{Q}_{X^nZ^nMM'}(x^n,z^n,m,m')\mathbf{P}_D(\hat{m}'|m,z^n)\mathbf{P}_\Phi(y^n|m,\hat{m}',z^n) \qquad (40) \\ & \mathbf{Q}^{(2)}_{X^nZ^nMM'\hat{M}'Y^n}(x^n,z^n,m,m',\hat{m}',y^n) \\ \triangleq\; & \mathbf{Q}_{X^nZ^nMM'}(x^n,z^n,m,m')\mathbf{P}_D(\hat{m}'|m,z^n)\mathbf{P}_\Phi(y^n|m,m',z^n). \qquad (41) \end{aligned}$$

Notice that $\mathbf{Q}^{(2)}$ differs from $\mathbf{Q}^{(1)}$ by allowing the decoder to use m' rather than $\hat{m}'$ when forming its reconstruction through ϕ^n.

Therefore, on account of (39),

$$\mathbb{E}_{\mathcal{C}^{(n)}}\left[\mathbf{Q}^{(2)}_{X^nZ^nY^n}(x^n,z^n,y^n)\right] = \overline{P}_{X^nZ^nY^n}(x^n,z^n,y^n). \tag{42}$$

Now applying the soft-covering lemma, since $R+R' > I_{\overline{P}}(Z,X;V) = I_{\overline{P}}(X;V)$, we have

$$\mathbb{E}_{\mathcal{C}^{(n)}}\left[\|\overline{P}_{X^nZ^n} - \mathbf{Q}_{X^nZ^n}\|_{TV}\right] \le \epsilon_n \to_n 0.$$

And with (34), (38), (40) and Property $1(d)$, we obtain

$$\begin{aligned} & \mathbb{E}_{\mathcal{C}^{(n)}}\left[\|\mathbf{P}_{X^nZ^nMM'\hat{M}'Y^n} - \mathbf{Q}^{(1)}_{X^nZ^nMM'\hat{M}'Y^n}\|_{TV}\right] \\ =\; & \mathbb{E}_{\mathcal{C}^{(n)}}\left[\|\overline{P}_{X^nZ^n} - \mathbf{Q}_{X^nZ^n}\|_{TV}\right] \qquad (43) \\ \le\; & \epsilon_n. \qquad (44) \end{aligned}$$

Since by definition $\mathbf{Q}^{(1)}_{X^nZ^nMM'\hat{M}'} = \mathbf{Q}^{(2)}_{X^nZ^nMM'\hat{M}'}$,

$$\mathbb{P}_{\mathbf{Q}^{(1)}}[\hat{M}' \neq M'] = \mathbb{P}_{\mathbf{Q}^{(2)}}[\hat{M}' \neq M'].$$

Also, since $R' < I(V;Z)$, the codebook is randomly generated, and M' is uniformly distributed under Q, it is well known that the maximum likelihood decoder $\mathbf{P}_D$ (as well as a variety of other decoders) will drive the error probability to zero as n goes to infinity. This can be seen from Fig. 4, by identifying, for fixed M, that M' is the message to be transmitted over the memoryless channel $\overline{P}_{Z|V}$. Specifically,

$$\mathbb{E}_{\mathcal{C}^{(n)}}\left[\mathbb{P}_{\mathbf{Q}^{(1)}}[M' \neq \hat{M}']\right] \leq \delta_n \to_n 0.$$

Applying Lemma 2, we obtain

$$\mathbb{E}_{\mathcal{C}^{(n)}}\left[\|\mathbf{Q}^{(1)}_{X^n Z^n M \hat{M}'} - \mathbf{Q}^{(2)}_{X^n Z^n M M'}\|_{TV}\right] \leq \mathbb{E}_{\mathcal{C}^{(n)}}\left[\mathbb{P}_{\mathbf{Q}^{(1)}}[\hat{M}' \neq M']\right] \leq \delta_n. \tag{45}$$

Thus by Property $1(d)$ and definitions (40) and (41),

$$\mathbb{E}_{\mathcal{C}^{(n)}}\left[\|\mathbf{Q}^{(1)}_{X^n Z^n M \hat{M}' Y^n} - \mathbf{Q}^{(2)}_{X^n Z^n M M' Y^n}\|_{TV}\right] \leq \delta_n. \tag{46}$$

Combining (44) and (46) and using Property $1(c)$ and $1(e)$, we have

$$\mathbb{E}_{\mathcal{C}^{(n)}}\left[\|\mathbf{P}_{X^n Y^n} - \mathbf{Q}^{(2)}_{X^n Y^n}\|_{TV}\right] \leq \epsilon_n + \delta_n \tag{47}$$

where ϵ_n and δ_n are the error terms introduced from the soft-covering lemma and channel coding, respectively.

Repeating the same steps as (23) through (25) on $\mathbf{P}$, $\mathbf{Q}^{(2)}$, and $\overline{P}$, we obtain

$$\mathbb{E}_{\mathcal{C}^{(n)}}\left[\mathbb{E}_{\mathbf{P}}[d(X^n, Y^n)]\right] \leq \mathbb{E}_{\overline{P}}\left[d(X, Y)\right] + d_{max}(\epsilon_n + \delta_n) \leq_n D. \tag{48}$$

Therefore, there exists a codebook satisfying the requirement. ■

V. The Berger-Tung Inner Bound

The application of the likelihood encoder can go beyond single-user communications. In this section, we will demonstrate the use of the likelihood encoder via an alternative proof for achieving the Berger-Tung inner bound for the problem of multi-terminal source coding. Notice that no Markov lemma is needed in this proof. Similar to the single-user case, the key is to identify an auxiliary distribution that has nice properties and show that the system-induced distribution and the auxiliary distribution we choose are close in total variation.

A. Problem Setup and Result Review

We now consider a pair of correlated sources $({X_1}^n, {X_2}^n)$, distributed i.i.d. according to $(X_{1t}, X_{2t}) \sim \overline{P}_{X_1X_2}$, independent encoders, and a joint decoder, satisfying the following constraints:

- Encoder 1 $f_{1n} : {\mathcal{X}_1}^n \mapsto \mathcal{M}_1$ (possibly stochastic);
- Encoder 2 $f_{2n} : {\mathcal{X}_2}^n \mapsto \mathcal{M}_2$ (possibly stochastic);
- Decoder $g_n : \mathcal{M}_1 \times \mathcal{M}_2 \mapsto {\mathcal{Y}_1}^n \times {\mathcal{Y}_2}^n$ (possibly stochastic);
- Compression rates: R_1, R_2, i.e. $|\mathcal{M}_1| = 2^{nR_1}$, $|\mathcal{M}_2| = 2^{nR_2}$.

The system performance is measured according to the time-averaged distortion (as defined in the notation section):

- $d_1({X_1}^n, {Y_1}^n) = \frac{1}{n}\sum_{t=1}^n d_1(X_{1t}, Y_{1t})$,
- $d_2({X_2}^n, {Y_2}^n) = \frac{1}{n}\sum_{t=1}^n d_2(X_{2t}, Y_{2t})$,

  where $d_1(\cdot,\cdot)$ and $d_2(\cdot,\cdot)$ can be different distortion measures.

Definition 5. *(R_1, R_2) is achievable under distortion level (D_1, D_2) if there exists a sequence of rate (R_1, R_2) encoders and decoder (f_{1n}, f_{2n}, g_n) such that*

$$\mathbb{E}[d_1({X_1}^n, {Y_1}^n)] \leq_n D_1,$$

$$\mathbb{E}[d_2({X_2}^n, {Y_2}^n)] \leq_n D_2.$$

The achievable rate region is not yet known in general. But an inner bound, reproduced below, was given in [4] and [5] and is known as the Berger-Tung inner bound. The rates (R_1, R_2) are achievable if

$$R_1 > I_{\overline{P}}(X_1; U_1|U_2), \quad (49)$$

$$R_2 > I_{\overline{P}}(X_2; U_2|U_1), \quad (50)$$

$$R_1 + R_2 > I_{\overline{P}}(X_1, X_2; U_1, U_2) \quad (51)$$

for some $\overline{P}_{U_1X_1X_2U_2} = \overline{P}_{X_1X_2}\overline{P}_{U_1|X_1}\overline{P}_{U_2|X_2}$, and functions $\phi_k(\cdot,\cdot)$ such that $\mathbb{E}[d_k(X_k, Y_k)] \leq D_k$, where $Y_k \triangleq \phi_k(U_1, U_2), k = 1, 2.$ [1]

[1]This region, after optimizing over auxiliary variables, is in fact not convex, so it can be improved to the convex hull through time-sharing.

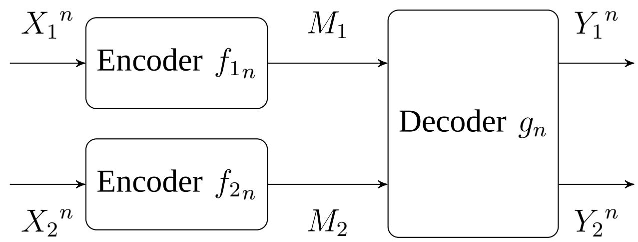

Fig. 5: Berger-Tung problem setup

B. Achievability Proof Using the Likelihood Encoder

For simplicity, we will focus on the corner points, $C_1 \triangleq (I_{\overline{P}}(X_1;U_1), I_{\overline{P}}(X_2;U_2|U_1))$ and $C_2 \triangleq (I_{\overline{P}}(X_1;U_1|U_2), I_{\overline{P}}(X_2;U_2))$, of the region given in (49) through (51) and use convexity to claim the complete region. Below we demonstrate how to achieve C_1. The point C_2 follows by symmetry.

Fix a $\overline{P}_{U_1U_2|X_1X_2} = \overline{P}_{U_1|X_1}\overline{P}_{U_2|X_2}$ and functions $\phi_k(\cdot,\cdot)$ such that $Y_k = \phi_k(U_1,U_2)$ and $\mathbb{E}_{\overline{P}}[d_k(X_k,Y_k)] < D_k$. Note that $U_1 - X_1 - X_2 - U_2$ forms a Markov chain under $\overline{P}$. We must show that any rate pair (R_1,R_2) satisfying $R_1 > I_{\overline{P}}(X_1;U_1)$ and $R_2 > I_{\overline{P}}(X_2;U_2|U_1)$ is achievable.

As expected, the decoder will use a lossy representation of one source as side information to assist reconstruction of the other source. We can choose an $R_2' < I_{\overline{P}}(U_1;U_2)$ such that $R_2 + R_2' > I_{\overline{P}}(X_2;U_2)$. Here R_2' corresponds to the rate of a virtual message M_2' which is produced by Encoder 2 but not physically transmitted to the receiver. This will play the role of the index of the codeword in the bin in a traditional covering and random-binning proof.

First we will use the likelihood encoder derived from $\overline{P}_{X_1U_1}$ and a random codebook $\{u_1{}^n(m_1)\}$ generated according to $\overline{P}_{U_1}$ for Encoder 1. Then we will use the likelihood encoder derived from $\overline{P}_{X_2U_2}$ and another random codebook $\{u_2{}^n(m_2,m_2')\}$ generated according to $\overline{P}_{U_2}$ for Encoder 2. The decoder will use the transmitted message M_1 to decode $U_1{}^n$, as in the point-to-point case, and use the transmitted message M_2 along with the decoded $U_1{}^n$ to decode M_2' as $\hat{M}_2'$, as in the Wyner-Ziv case, and reproduce $u_2^n(M_2,\hat{M}_2')$. Finally, the decoder outputs the reconstructions $Y_k{}^n$ according to the symbol-by-symbol functions $\phi_k(\cdot,\cdot)$ of $U_1{}^n$ and $U_2{}^n$.

The distribution induced by the encoders and decoder is

$$\mathbf{P}_{X_1{}^n X_2{}^n U_1{}^n M_1 M_2 M_2' \hat{M}_2' Y_1{}^n Y_2{}^n} = \overline{P}_{X_1{}^n X_2{}^n} \mathbf{P}_1 \mathbf{P}_2 \tag{52}$$

where

$$\begin{aligned} & \mathbf{P}_1(m_1, u_1{}^n | x_1{}^n) \\ \triangleq\ & \mathbf{P}_{M_1|X_1{}^n}(m_1|x_1{}^n) \mathbf{P}_{U_1{}^n|M_1}(u_1{}^n|m_1) && (53) \\ \triangleq\ & \mathbf{P}_{LE1}(m_1|x_1{}^n) \mathbf{P}_{D1}(u_1{}^n|m_1) && (54) \end{aligned}$$

and

$$\begin{aligned} & \mathbf{P}_2(m_2, m_2', \hat{m}_2', y_1{}^n, y_2{}^n | x_2{}^n, u_1{}^n) \\ \triangleq\ & \mathbf{P}_{M_2 M_2'|X_2{}^n}(m_2, m_2'|x_2{}^n) \mathbf{P}_{\hat{M}_2'|M_2 U_1{}^n}(\hat{m}_2'|m_2, u_1{}^n) \\ & \prod_{k=1,2} P_{Y_k{}^n|U_1{}^n M_2 \hat{M}_2'}(y_k{}^n|u_1{}^n, m_2, \hat{m}_2') && (55) \\ \triangleq\ & \mathbf{P}_{LE2}(m_2, m_2'|x_2{}^n) \mathbf{P}_{D2}(\hat{m}_2'|m_2, u_1{}^n) \\ & \prod_{k=1,2} \mathbf{P}_{\Phi,k}(y_k{}^n|u_1{}^n, m_2, \hat{m}_2'), && (56) \end{aligned}$$

where $\mathbf{P}_{LE1}$ and $\mathbf{P}_{LE2}$ are the likelihood encoders; $\mathbf{P}_{D1}$ is the first part of the decoder that does a codeword lookup on $\mathcal{C}_1^{(n)}$; $\mathbf{P}_{D2}$ is the second part of the decoder that decodes m_2' as $\hat{m}_2'$; and $\mathbf{P}_{\Phi,k}(y_k{}^n|u_1{}^n, m_2, \hat{m}_2')$ is the third part of the decoder that reconstructs the source sequences.

We now restate the behavior of the encoders and decoder, as components of the induced distribution.

Codebook generation: We independently generate 2^{nR_1} sequences in $\mathcal{U}_1{}^n$ according to $\prod_{t=1}^n \overline{P}_{U_1}(u_{1t})$ and index them by $m_1 \in [1 : 2^{nR_1}]$, and independently generate $2^{n(R_2+R_2')}$ sequences in $\mathcal{U}_2{}^n$ according to $\prod_{t=1}^n \overline{P}_{U_2}(u_{2t})$ and index them by $(m_2, m_2') \in [1 : 2^{nR_2}] \times [1 : 2^{nR_2'}]$. We use $\mathcal{C}_1^{(n)}$ and $\mathcal{C}_2^{(n)}$ to denote the two random codebooks, respectively.

Encoders: The first encoder $\mathbf{P}_{LE1}(m_1|x_1{}^n)$ is the likelihood encoder according to $\overline{P}_{X_1{}^n U_1{}^n}$ and $\mathcal{C}_1^{(n)}$. The second encoder $\mathbf{P}_{LE2}(m_2, m_2'|x_2{}^n)$ is the likelihood encoder according to $\overline{P}_{X_2{}^n U_2{}^n}$ and $\mathcal{C}_2^{(n)}$.

Decoder: First, let $\mathbf{P}_{D1}(u_1{}^n|m_1)$ be a $\mathcal{C}_1^{(n)}$ codeword lookup decoder. Then, let $\mathbf{P}_{D2}(\hat{m}_2'|m_2, u_1{}^n)$ be a good channel decoder with respect to the sub-codebook $\mathcal{C}_2^{(n)}(m_2) = \{u_2{}^n(m_2, a)\}_a$ and the memoryless channel $\overline{P}_{U_1|U_2}$. Last, define ${\phi_k}^n({u_1}^n, {u_2}^n)$ as the concatenation $\{\phi_k(u_{1t}, u_{2t})\}_{t=1}^n$ and set the decoders $\mathbf{P}_{\Phi,k}$ to be the deterministic functions

$$\mathbf{P}_{\Phi,k}({y_k}^n|{u_1}^n, m_2, \hat{m}_2') \triangleq \mathbb{1}\{{y_k}^n = {\phi_k}^n({u_1}^n, {U_2}^n(m_2, \hat{m}_2'))\}.$$

Analysis: We will need the following distributions: the induced distribution $\mathbf{P}$ and auxiliary distributions $\mathbf{Q}_1$ and $\mathbf{Q}_1^*$. The general idea of the proof is as follows: Encoder 1 makes $\mathbf{P}$ and $\mathbf{Q}_1$ close in total variation. Distribution $\mathbf{Q}_1^*$ (random only with respect to the second codebook $\mathcal{C}_2^{(n)}$) is the expectation of $\mathbf{Q}_1$ over the random codebook $\mathcal{C}_1^{(n)}$. This is really the key step in the proof. By considering the expectation of the distribution with respect to $\mathcal{C}_1^{(n)}$, we effectively remove Encoder 1 from the problem and turn the message from Encoder 1 into memoryless side information at the decoder. Hence, the two distortions (averaged over $\mathcal{C}_1^{(n)}$) under $\mathbf{P}$ are roughly the same as the distortions under $\mathbf{Q}_1^*$, which is a much simpler distribution. We then recognize $\mathbf{Q}_1^*$ as precisely $\mathbf{P}$ in (34) from the Wyner-Ziv proof of the previous section, with a source pair (X_1, X_2), a pair of reconstructions (Y_1, Y_2) and U_1 as the side information.

1) The auxiliary distribution $\mathbf{Q}_1$ takes the following form:

$$\mathbf{Q}_{1{X_1}^n{X_2}^n{U_1}^nM_1M_2M_2'\hat{M}_2'{Y_1}^n{Y_2}^n} = \mathbf{Q}_{1M_1{U_1}^n{X_1}^n{X_2}^n}\mathbf{P}_2$$

where

$$\begin{aligned} &\mathbf{Q}_{1M_1{U_1}^n{X_1}^n{X_2}^n}(m_1, {u_1}^n, {x_1}^n, {x_2}^n) \\ = \;& \frac{1}{2^{nR_1}}\mathbb{1}\{{u_1}^n = {U_1}^n(m_1)\}\overline{P}_{{X_1}^n|{U_1}^n}({x_1}^n|{u_1}^n)\overline{P}_{{X_2}^n|{X_1}^n}({x_2}^n|{x_1}^n). \end{aligned} \tag{57}$$

Note that $\mathbf{Q}_1$ is the idealized distribution with respect to the first message, as introduced in the point-to-point case. Hence, by the same arguments, since $R_1 > I_{\overline{P}}(X_1; U_1)$, using the soft-covering lemma,

$$\mathbb{E}_{\mathcal{C}_1^{(n)}}\left[\|\mathbf{Q}_1 - \mathbf{P}\|_{TV}\right] \leq \epsilon_{1n}, \tag{58}$$

where $\mathbf{Q}_1$ and $\mathbf{P}$ are distributions over random variables ${X_1}^n, {X_2}^n, {U_1}^n, M_1, M_2, M_2', \hat{M}_2', {Y_1}^n, {Y_2}^n$ and ϵ_{1n} is the error term introduced from soft-covering lemma.

2) Taking the expectation over codebook $\mathcal{C}_1^{(n)}$, we define

$$\mathbf{Q}^*_{1{X_1}^n{X_2}^n{U_1}^nM_2M_2'\hat{M}_2'{Y_1}^n{Y_2}^n} \triangleq \mathbb{E}_{\mathcal{C}_1^{(n)}}\left[\mathbf{Q}_{1{X_1}^n{X_2}^n{U_1}^nM_2M_2'\hat{M}_2'{Y_1}^n{Y_2}^n}\right]. \tag{59}$$

Note that under this definition of $\mathbf{Q}_1^*$, we have

$$\begin{aligned}
&\mathbf{Q}^*_{1X_1{}^nX_2{}^nU_1{}^nM_2M_2'\hat{M}_2'Y_1{}^nY_2{}^n}(x_1{}^n, x_2{}^n, u_1{}^n, m_2, m_2', \hat{m}_2', y_1{}^n, y_2{}^n) \\
=\ & \mathbb{E}_{\mathcal{C}_1^{(n)}}\left[\mathbf{Q}^*_{1X_1{}^nX_2{}^nU_1{}^n}(x_1{}^n, x_2{}^n, u_1{}^n)\right]\mathbf{P}_2(m_2, m_2', \hat{m}_2', y_1{}^n, y_2{}^n | x_2{}^n, u_1{}^n) \qquad (60)\\
=\ & \overline{P}_{X_1{}^nX_2{}^nU_1{}^n}(x_1{}^n, x_2{}^n, u_1{}^n)\mathbf{P}_2(m_2, m_2', \hat{m}_2', y_1{}^n, y_2{}^n | x_2{}^n, u_1{}^n), \qquad (61)
\end{aligned}$$

where the last step can be verified using the same technique as (12) given in Section III.

By Property $1(b)$,

$$\begin{aligned}
&\mathbb{E}_{\mathcal{C}_1^{(n)}}\left[\mathbb{E}_{\mathbf{P}}\left[d_k(X_k{}^n, Y_k{}^n)\right]\right] \\
\le\ & \mathbb{E}_{\mathcal{C}_1^{(n)}}\left[\mathbb{E}_{\mathbf{Q}_1}[d_k(X_k{}^n, Y_k{}^n)]\right] + d_{kmax}\epsilon_{1n} \qquad (62)\\
=\ & \mathbb{E}_{\mathcal{C}_1^{(n)}}\left[\sum_{x_k{}^n, y_k{}^n}\mathbf{Q}_1(x_k{}^n, y_k{}^n)d_k(x_k{}^n, y_k{}^n)\right] + d_{kmax}\epsilon_{1n} \qquad (63)\\
=\ & \sum_{x_k{}^n, y_k{}^n}\mathbb{E}_{\mathcal{C}_1^{(n)}}[\mathbf{Q}_1(x_k{}^n, y_k{}^n)]d_k(x_k{}^n, y_k{}^n) + d_{kmax}\epsilon_{1n} \qquad (64)\\
=\ & \sum_{x_k{}^n, y_k{}^n}\mathbf{Q}_1^*(x_k{}^n, y_k{}^n)d_k(x_k{}^n, y_k{}^n) + d_{kmax}\epsilon_{1n} \qquad (65)\\
=\ & \mathbb{E}_{\mathbf{Q}_1^*}\left[d_k(X_k{}^n, Y_k{}^n)\right] + d_{kmax}\epsilon_{1n}. \qquad (66)
\end{aligned}$$

Note that $\mathbf{Q}_1^*$ is exactly of the form of the induced distribution $\mathbf{P}$ in the Wyner-Ziv proof of the previous section, with the inconsequential modification that there are two reconstructions and two distortion functions. Thus, by (40) through (48), we obtain

$$\begin{aligned}
&\mathbb{E}_{\mathcal{C}_2^{(n)}}\left[\mathbb{E}_{\mathbf{Q}_1^*}\left[d_k(X_k{}^n, Y_k{}^n)\right]\right] \\
\le\ & \mathbb{E}_{\overline{P}}\left[d_k(X_k, Y_k)\right] + d_{kmax}(\epsilon_{2n} + \delta_n), \qquad (67)
\end{aligned}$$

where ϵ_{2n} and δ_n are error terms introduced from the soft-covering lemma and channel decoding, respectively.

Finally, taking expectation over $\mathcal{C}_1^{(n)}$ and using (66) and (67),

$$\begin{aligned}
&\mathbb{E}_{\mathcal{C}_2^{(n)}}\left[\mathbb{E}_{\mathcal{C}_1^{(n)}}\left[\mathbb{E}_{\mathbf{P}}\left[d_k(X_k{}^n, Y_k{}^n)\right]\right]\right] \\
\le\ & \mathbb{E}_{\mathcal{C}_2^{(n)}}\left[\mathbb{E}_{\mathbf{Q}_1^*}\left[d_k(X_k{}^n, Y_k{}^n)\right] + d_{kmax}\epsilon_{1n}\right] \qquad (68)\\
\le\ & \mathbb{E}_{\overline{P}}\left[d_k(X_k, Y_k)\right] + d_{kmax}\epsilon_{1n} + d_{kmax}(\epsilon_{2n} + \delta_n) \qquad (69)\\
\le_n\ & D_k, \qquad (70)
\end{aligned}$$

where (68) follows from (66); (69) follows from (66) and (67). ■

Remark 2. Note that the proof above uses the proof of Wyner-Ziv achievability from the previous section. To do the entire proof step by step, we would define a total of three auxiliary distributions, which would be the $\mathbf{Q}_1$ used in the proof, as well as $\mathbf{Q}_2^{(1)}$ and $\mathbf{Q}_2^{(2)}$ defined below for completeness. The steps outlined above show how to relate the induced distribution $\mathbf{P}$ to $\mathbf{Q}_1$ and its expectation $\mathbf{Q}_1^*$. This effectively converts the message from Encoder 1 into memoryless side information at the decoder. The omitted steps, as seen in the previous section, relate $\mathbf{Q}_1^*$ to $\mathbf{Q}_2^{(1)}$ through the soft-covering lemma and $\mathbf{Q}_2^{(1)}$ to $\mathbf{Q}_2^{(2)}$ through reliable channel decoding. The expected value of $\mathbf{Q}_2^{(2)}$ over codebooks is the desired distribution $\overline{P}$. For reference, the omitted auxiliary distributions are

$$\begin{aligned} & \mathbf{Q}_{2M_2M_2'U_2{}^nX_2{}^nX_1{}^nU_1{}^n} \\ = \; & \frac{1}{2^{n(R_2+R_2')}}\mathbb{1}\{u_2{}^n = U_2{}^n(m_2,m_2')\}\overline{P}_{X_2{}^n|U_2{}^n}(x_2{}^n|u_2{}^n) \\ & \overline{P}_{X_1{}^nU_1{}^n|X_2{}^n}(x_1{}^n,u_1{}^n|x_2{}^n), \end{aligned} \tag{71}$$

which is of the same structure as the idealized distribution described in Fig. 4, and

$$\begin{aligned} \mathbf{Q}_2^{(1)}{}_{X_1{}^nX_2{}^nU_1{}^nM_2M_2'\hat{M}_2'Y_1{}^nY_2{}^n} \triangleq \; & \mathbf{Q}_{2X_1{}^nX_2{}^nU_1{}^nM_2M_2'} \\ & \mathbf{P}_D(\hat{m}_2'|m_2,u_1{}^n)\prod_{k=1,2}\mathbf{P}_{\Phi,k}(y_k{}^n|u_1{}^n,m_2,\hat{m}_2') \end{aligned} \tag{72}$$

$$\begin{aligned} \mathbf{Q}_2^{(2)}{}_{X_1{}^nX_2{}^nU_1{}^nM_2M_2'\hat{M}_2'Y_1{}^nY_2{}^n} \triangleq \; & \mathbf{Q}_{2X_1{}^nX_2{}^nU_1{}^nM_2M_2'} \\ & \mathbf{P}_D(\hat{m}_2'|m_2,u_1{}^n)\prod_{k=1,2}\mathbf{P}_{\Phi,k}(y_k{}^n|u_1{}^n,m_2,m_2'). \end{aligned} \tag{73}$$

Remark 3. To see how this is a simpler proof than the traditional joint typicality encoder proof, recall from [8] that to bound the different error events, we would need the regular covering lemma, the conditional typicality lemma, the Markov lemma, and the mutual packing lemma, some of which are quite involving to verify. With the likelihood encoder, all we need is the soft-covering lemma and Lemma 2.

VI. Excess Distortion and Non-asymptotic Analysis

The proofs presented in the previous sections are for the average distortion criterion, i.e. $\mathbb{E}\left[\sum_{t=1}^n d(X_t,Y_t)\right] \leq_n D$. However, it is not hard to modify the proofs to show that they also hold for excess distortion. For brevity, we will demonstrate the analysis only for the point-to-point case.

With the same setup as in Section III, we change the average distortion requirement in the definition of achievability (Definition 1) to the requirement that

$$\mathbb{P}\left[d(X^n, Y^n) > D\right] \to_n 0.$$

The corresponding rate-distortion function is still given by $R(D)$ in (6).

A. Modified Proof for Excess Distortion

For the excess distortion, we will use the exact same encoding/decoding scheme, along with the same random codebook $\mathcal{C}^n$, from Section III. We make the following modifications.

We replace (13) to (17) with

$$\begin{aligned}
& \mathbb{E}_{\mathcal{C}^{(n)}}\left[\mathbb{P}_{\mathbf{Q}}\left[d(X^n, Y^n) > D\right]\right] \\
= \; & \mathbb{E}_{\mathcal{C}^{(n)}}\left[\sum_{x^n, y^n} \mathbf{Q}(x^n, y^n) \mathbb{1}\{d(X^n, Y^n) > D\}\right] && (74) \\
= \; & \sum_{x^n, y^n} \mathbb{E}_{\mathcal{C}^{(n)}}[\mathbf{Q}(x^n, y^n)] \mathbb{1}\{d(x^n, y^n) > D\} && (75) \\
= \; & \sum_{x^n, y^n} \overline{P}_{X^n, Y^n}(x^n, y^n) \mathbb{1}\{d(x^n, y^n) > D\} && (76) \\
= \; & \mathbb{P}_{\overline{P}}[d(X^n, Y^n) > D], && (77)
\end{aligned}$$

and replace (23) to (25) with

$$\begin{aligned}
& \mathbb{E}_{\mathcal{C}^{(n)}}\left[\mathbb{P}_{\mathbf{P}}[d(X^n, Y^n) > D]\right] \\
\leq \; & \mathbb{E}_{\mathcal{C}^{(n)}}\left[\mathbb{P}_{\mathbf{Q}}[d(X^n, Y^n) > D]\right] + \epsilon_n && (78) \\
= \; & \mathbb{P}_{\overline{P}}\left[d(X^n, Y^n) > D\right] + \epsilon_n && (79)
\end{aligned}$$

where the last step follows from (77). Therefore, there exists a codebook that satisfies the requirement. ■

B. Non-asymptotic Analysis

Let the achievable rate-distortion region $\mathcal{R}$ be

$$\mathcal{R} \triangleq \{(R, D) : R > R(D)\}.$$

For a fixed $(R, D) \in \mathcal{R}$, we aim to minimize the probability of excess distortion, using a random codebook and the likelihood encoder, over valid choices of $\overline{P}_{Y|X}$, and evaluate how fast the excess distortion decays with blocklength n under the optimal $\overline{P}_{Y|X}$. Mathematically, we want to obtain

$$\inf_{\overline{P}_{Y|X}} \mathbb{E}_{\mathcal{C}^n}\left[\mathbb{P}_{\mathbf{P}}\left[d(X^n,Y^n) > D\right]\right], \tag{80}$$

where the subscript $\mathbf{P}$ indicates probability taken with respect to the induced distribution.

To evaluate how fast the probability of excess distortion approaches zero, note in (79) that the first term is governed (approximately) by the gap $D - \mathbb{E}_{\overline{P}}[d(X,Y)]$ and the second term is governed (approximately) by the the gap $R - I_{\overline{P}}(X;Y)$. To see this, observe that for any $\beta > 0$,

$$\begin{aligned}
\epsilon_n' &\triangleq \mathbb{P}_{\overline{P}}[d(X^n,Y^n) > D] \\
&= \mathbb{P}_{\overline{P}}\left[\frac{1}{n}\sum_{t=1}^{n} d(X_t,Y_t) > D\right] && (81)\\
&\leq \inf_{\beta>0}\left[\frac{\mathbb{E}_{\overline{P}}[2^{\beta d(X,Y)}]}{2^{\beta D}}\right]^n && (82)\\
&= \exp\left(-n\log\left(\inf_{\beta>0}\mathbb{E}_{\overline{P}}\left[2^{\beta(d(X,Y)-D)}\right]\right)^{-1}\right) && (83)\\
&= \exp\left(-n\eta(\overline{P}_{Y|X})\right) && (84)
\end{aligned}$$

where (82) follows from the Chernoff bound and we have implicitly defined

$$\eta(\overline{P}_{Y|X}) \triangleq \log(\inf_{\beta>0}\mathbb{E}_{\overline{P}}\left[2^{\beta(d(X,Y)-D)}\right])^{-1}. \tag{85}$$

An upper bound on the second term in (79) is given in [15], reproduced below:

$$\epsilon_n \leq \frac{3}{2}\exp\left(-n\gamma(\overline{P}_{Y|X})\right), \tag{86}$$

where

$$\gamma(\overline{P}_{Y|X}) \triangleq \max_{\alpha\geq 1,\alpha'\leq 2}\frac{\alpha-1}{2\alpha-\alpha'}\left(R - \check{I}_{\overline{P},\alpha}(X;Y) + (\alpha'-1)(\check{I}_{\overline{P},\alpha}(X;Y) - \bar{I}_{\overline{P},\alpha'}(X;Y))\right) \tag{87}$$

$$\check{I}_{\overline{P},\alpha}(X;Y) \triangleq \frac{1}{\alpha-1}\log\left(\mathbb{E}_{\overline{P}}\left[\left(\frac{\overline{P}_{X,Y}(X,Y)}{\overline{P}_X(X)\overline{P}_Y(Y)}\right)^{\alpha-1}\right]\right) \tag{88}$$

$$\bar{I}_{\overline{P},\alpha'}(X,Y) \triangleq \frac{1}{\alpha'-1}\log\left(\left(\mathbb{E}_{\overline{P}_X}\left[\sqrt{\mathbb{E}_{\overline{P}_{Y|X}}\left[\left(\frac{\overline{P}_{XY}(X,Y)}{\overline{P}_X(X)\overline{P}_Y(Y)}\right)^{\alpha'-1}\right]}\right]\right)^2\right) \tag{89}$$

Both ϵ'_n and ϵ_n decay exponentially with n. To obtain an upper bound on the excess distortion given in (80), we now have a new optimization problem in the following form:

$$\inf_{\overline{P}_{Y|X}} \exp\left(-n\eta(\overline{P}_{Y|X})\right) + \frac{3}{2}\exp\left(-n\gamma(\overline{P}_{Y|X})\right), \tag{90}$$

where $\eta(\overline{P}_{Y|X})$ and $\gamma(\overline{P}_{Y|X})$ are defined in (85) and (87). Note that only choices of $\overline{P}_{Y|X}$ such that $\mathbb{E}_{\overline{P}}[d(X,Y)] < D$ and $I_{\overline{P}}(X;Y) < R$ should be considered for the optimization, as other choices render the bound degenerate.

We can relax (90) to obtain a simple upper bound on the excess distortion $\mathbb{P}_P[d(X^n,Y^n) > D]$ given in the following theorem.

Theorem 1. *The excess distortion* $\mathbb{P}_P[d(X^n,Y^n) > D]$ *using the likelihood encoder is upper bounded by*

$$\inf_{\overline{P}_{Y|X}} \frac{5}{2}\exp\left(-n\min\left\{\eta\left(\overline{P}_{Y|X}\right), \gamma\left(\overline{P}_{Y|X}\right)\right\}\right). \tag{91}$$

where $\eta(\overline{P}_{Y|X})$ *and* $\gamma(\overline{P}_{Y|X})$ *are given in* (85) *and* (87)*, respectively.*

Remark 4. Note that this bound does not achieve the exponent that we know to be optimal [22, Theorem 9.5] for rate-distortion theory. It may very well be that the likelihood encoder does not achieve the optimal exponent, though it may also be an artifact of our proof or the bound for the soft-covering lemma.

VII. Comparison with Random Binning Based Proof

The likelihood encoder proof technique is similar to the random binning based analysis approach presented in [23] in many ways. In this section, we will compare the two schemes along with their non-asymptotic behaviors.

We shall first provide a recap of the scheme for point-to-point lossy compression that uses the so-called "output statistics of random binning" in the proof. Below we modify the way it was originally presented in [23] to ease the comparison with the proof given in Section III-B.

A. The Proportional-Probability Encoder

We start by defining a source encoder that looks very similar in form to a likelihood encoder defined in Section II-C. Like any other source encoder, a *proportional-probability encoder* receives a sequence $x_1, ..., x_n$ and produces an index $m \in [1 : 2^{nR}]$.

A codebook is specified by a non-empty collection $\mathcal{C}$ of sequences $y^n \in \mathcal{Y}^n$ and indices $m(y^n)$ assigned to each $y^n \in \mathcal{Y}^n$. The codebook and a joint distribution P_{XY} specify the proportional-probability encoder.

Let $\mathcal{G}(m|x^n)$ be the probability, as a result of passing x^n through a memoryless channel given by $P_{Y|X}$, of finding Y^n in the collection $\mathcal{C}$ and retrieving the index m from the codebook:

$$\begin{aligned}\mathcal{G}(m|x^n) &\triangleq \mathbb{P}_{P_{Y^n|X^n}}\left[Y^n \in \mathcal{C}, m(Y^n) = m \mid X^n = x^n\right] \\ &= \sum_{y^n \in \mathcal{C}} P_{Y^n|X^n}(y^n|x^n)\mathbb{1}\{m(y^n) = m\}.\end{aligned}$$

A proportional-probability encoder is a stochastic encoder that determines the message index with probability proportional to $\mathcal{G}(m|x^n)$, i.e.

$$P_{M|X^n}(m|x^n) = \frac{\mathcal{G}(m|x^n)}{\sum_{m' \in [1:2^{nR}]} \mathcal{G}(m'|x^n)} \propto \mathcal{G}(m|x^n). \tag{92}$$

B. Scheme Using the Proportional-Probability Encoder

Before going into the achievability scheme, we first state a lemma that will be used in the analysis.

Lemma 3 (Independence of random binning - Theorem 1 of [23])**.** *Given a probability mass function P_{XY}, and each $y^n \in \mathcal{Y}^n$ is independently assigned to a bin index $b \in [1 : 2^{nR_b}]$ uniformly at random, where $B(y^n)$ denotes this random assignment. Define the joint distribution*

$$\mathbf{P}_{X^nY^nB}(x^n, y^n, b) \triangleq \prod_{i=1}^{n} P_{XY}(x_i, y_i)\mathbb{1}\{B(y^n) = b\}.$$

If $R_b < H(Y|X)$, then we have

$$\mathbb{E}_{\mathcal{B}}\left[\left\|\mathbf{P}_{X^nB} - P_{X^n}P_B^U\right\|_{TV}\right] \to_n 0,$$

where P_B^U is a uniform distribution on $[1 : 2^{nR_b}]$ and $\mathbb{E}_{\mathcal{B}}$ denotes expectation taken over the random binning.

We now outline the encoding-decoding scheme based on the proportional-probability encoder.

Fix a $\overline{P}_{Y|X}$ that satisfies $\mathbb{E}_{\overline{P}}[d(X, Y)] < D$ and choose the rates R and R' to satisfy $R' < H_{\overline{P}}(Y|X)$ and $R + R' > H_{\overline{P}}(Y)$.

Codebook generation: Each $y^n \in \mathcal{Y}^n$ is randomly and independently assigned to the codebook $\mathcal{C}$ with probability $2^{-nR'}$. Then, independent of the construction of $\mathcal{C}$, each $y^n \in \mathcal{Y}^n$ is independently assigned uniformly at random to one of 2^{nR} bins indexed by M.

Encoder: The encoder $\mathbf{P}_{PPE}(m|x^n)$ is the proportional-probability encoder with respect to $\overline{P}$. Specifically, the encoder chooses M stochastically according to (92), with $\mathcal{G}$ based on $\overline{P}$ as follows:

$$\mathcal{G}(m|x^n) = \sum_{y^n \in \mathcal{C}} \overline{P}_{Y^n|X^n}(y^n|x^n) \mathbb{1}\{m(y^n) = m\},$$

where $\overline{P}_{Y^n|X^n}(y^n|x^n) = \prod_{t=1}^n \overline{P}_{Y|X}(y_t|x_t)$.

Decoder: The decoder $\mathbf{P}_D(y^n|m)$ selects a y^n reconstruction that is in $\mathcal{C}$ and has index $m = M$. There will usually be more than one such y^n sequence, but rarely will there be more than one "good" choice, due to the rates used. The decoder can choose that most probable y^n sequence or the unique typical sequence, etc. The proof in [23] uses a "mismatch stochastic likelihood coder" (MSLC) [24] [20], and we will use their analysis for the performance bound in Section VII-C.

Remark 5. Intuitively, a decoder can successfully decode the sequence intended by the encoder since there are roughly $2^{nH_{\overline{P}}(Y)}$ typical y^n sequences, and the collection $\mathcal{C}$ together with the binning index M provides high enough rate $R' + R > H_{\overline{P}}(Y)$ to uniquely identify the sequence.

Analysis: The above scheme specifies a system induced distribution of the form:

$$\mathbf{P}_{X^nMY^n}(x^n, m, y^n) = \overline{P}_{X^n}\mathbf{P}_{PPE}(m|x^n)\mathbf{P}_D(y^n|m).$$

To analyze the above scheme, we start by replacing the codebook used for encoding and decoding with a set of codebooks. Recall that the codebook consists of a collection $\mathcal{C}$ and index assignments $m(y^n)$ that are both randomly constructed. Now consider a set of $2^{nR'}$ collections $\{\mathcal{C}_f\}_{f \in [1:2^{nR'}]}$, indexed by f, created by assigning each y^n sequence in $\mathcal{Y}^n$ randomly to exactly one collection equiprobably. From this we define a set of $2^{nR'}$ codebooks, one for each f, each one consisting of the collection $\mathcal{C}_f$ and the common message index function $m(y^n)$. We use $\mathcal{K}$ to denote this set of random codebooks.

By this construction, the original random collection $\mathcal{C}$ in the codebook used by the encoder and decoder is equivalent in probability to using the first codebook associated with $\mathcal{C}_1$. It is also equivalent to using a random codebook in the set, which is a point we will utilize shortly.

The purpose of defining multiple codebooks is to facilitate general proof tools associated with uniform random binning.

Here we summarize the proof given in [23]. In addition to the system induced random variables, we introduce a random variable F which is uniformly distributed on the set $\{1, .., 2^{nR'}\}$ and independent of X^n. The variable F selects the codebook to be used—everything else about the encoding and decoding remains the same. We have noted that the behavior and performance of this system with multiple codebooks is equivalent to that of the actual encoding and decoding. Nevertheless, we will formalize this connection in (108). For now, we refer to this new distribution that includes many codebooks as the pseudo induced distribution $\tilde{\mathbf{P}}$. According to $\tilde{\mathbf{P}}$, there is a set of randomly generated codebooks, and the one for use is selected by F.

The pseudo induced distribution can be expressed in the following form:

$$\begin{aligned} & \tilde{\mathbf{P}}_{FX^nMY^n}(f, x^n, m, y^n) \\ = \; & P_F(f)\overline{P}_{X^n}(x^n)\mathbf{P}_{PPE}(m|x^n, f)\mathbf{P}_D(y^n|m, f). \end{aligned} \tag{93}$$

We reiterate that

$$\mathbf{P}_{X^nMY^n} \stackrel{d}{=} \tilde{\mathbf{P}}_{X^nMY^n|F=f}, \quad \forall f \in [1:2^{nR'}]. \tag{94}$$

We now introduce one more random variable that never actually materialized during the implementation. Let $\tilde{Y}^n$ be the reconstruction sequence intended by the encoder. The encoding can be considered as a two step process. First, the encoder selects a $\tilde{Y}^n$ sequence from $\mathcal{C}_f$ with probability proportional to that induced by passing x^n through a memoryless channel given by $\overline{P}_{Y|X}$. Next, the encoder looks up the message index $m(\tilde{Y}^n)$ and transmits it as M.

Accordingly, we will replace the encoder in the pseudo induced distribution with the two parts discussed:

$$\mathbf{P}_{PPE}(m|x^n, f) = \sum_{\tilde{y}^n} \mathbf{P}_{E1}(\tilde{y}^n|x^n, f)\mathbf{P}_{E2}(m|\tilde{y}^n). \tag{95}$$

To analyze the expected distortion performance of the pseudo induced distribution $\tilde{\mathbf{P}}$, we introduce two approximating distributions $\mathbf{Q}^{(1)}$ and $\mathbf{Q}^{(2)}$.

Let us first define the distribution $\mathbf{Q}^{(1)}$:

$$\begin{aligned} & \mathbf{Q}^{(1)}_{FX^n\tilde{Y}^nMY^n}(f, x^n, \tilde{y}^n, m, y^n) \\ \triangleq \; & \overline{P}_{X^nY^n}(x^n, \tilde{y}^n)\mathbf{Q}_{F|\tilde{Y}^n}(f|\tilde{y}^n)\mathbf{P}_{E2}(m|\tilde{y}^n)\mathbf{P}_D(y^n|m, f) \end{aligned} \tag{96}$$

where $\mathbf{Q}_{F|\tilde{Y}^n}(f|\tilde{y}^n) = \mathbb{1}\{\tilde{y}^n \in \mathcal{C}_f\}$. In words, $\mathbf{Q}^{(1)}$ is constructed from an i.i.d. distribution according to $\overline{P}$ on $(X^n, \tilde{Y}^n)$, two random binnings F and M, as specified by the construction of the set of codebooks $\mathcal{K}$, and a decoding of Y^n from the random binnings.

Now we arrive at the reason for using the proportional-probability encoder. Part 1 of the encoder that selects the $\tilde{Y}^n$ sequences is precisely the conditional probability specified by $\mathbf{Q}^{(1)}$:

$$\mathbf{Q}^{(1)}_{\tilde{Y}^n|X^nF}(\tilde{y}^n|x^n, f) = \mathbf{P}_{E1}(\tilde{y}^n|x^n, f).$$

Therefore, the only difference between the pseudo induced distribution $\tilde{\mathbf{P}}$ and $\mathbf{Q}^{(1)}$ is the conditional distribution of F given X^n. This is where Lemma 3 plays a role.

Applying Lemma 3 by identifying F as the uniform binning, since $R' < H_{\overline{P}}(Y|X)$, we obtain

$$\mathbb{E}_{\mathcal{K}}\left[\left\|\mathbf{Q}^{(1)}_{X^nF} - \tilde{P}_{X^nF}\right\|_{TV}\right] \le \epsilon_n^{(rb)} \to_n 0. \tag{97}$$

Using Property 1 (d), we have

$$\mathbb{E}_{\mathcal{K}}\left[\left\|\tilde{\mathbf{P}}_{FX^nY^nM\hat{Y}^n} - \mathbf{Q}^{(1)}_{FX^nY^nM\hat{Y}^n}\right\|_{TV}\right] \le \epsilon_n^{(rb)}. \tag{98}$$

The next approximating distribution we define is $\mathbf{Q}^{(2)}$:

$$\mathbf{Q}^{(2)}_{FX^n\tilde{Y}^nMY^n}(f, x^n, \tilde{y}^n, m, y^n) \triangleq \mathbf{Q}^{(1)}_{FX^n\tilde{Y}^nM}(f, x^n, \tilde{y}^n, m)\mathbb{1}\{y^n = \tilde{y}^n\}. \tag{99}$$

Recall from Remark 5, decoding $\tilde{Y}^n$ will succeed with high probability if the total rate of the binnings is above the entropy rate of the sequence that was binned. This is well known from the Slepian-Wolf coding result [25] [26]. Therefore, since the total binning rate $R + R' > H_{\overline{P}}(Y)$, according to the definition of total variation, we obtain

$$\mathbb{E}_{\mathcal{K}}\left[\left\|\mathbf{Q}^{(1)}_{\tilde{Y}^nY^n} - \mathbf{Q}^{(2)}_{\tilde{Y}^nY^n}\right\|_{TV}\right] \le \epsilon_n^{(sw)} \to_n 0, \tag{100}$$

where $\epsilon_n^{(sw)}$ is the decoding error.

Again by Property 1 (d), we have

$$\mathbb{E}_{\mathcal{K}}\left[\left\|\mathbf{Q}^{(1)}_{FX^n\tilde{Y}^nMY^n} - \mathbf{Q}^{(2)}_{FX^n\tilde{Y}^nMY^n}\right\|_{TV}\right] \le \epsilon_n^{(sw)}. \tag{101}$$

Combining (98) and (101) using the triangle inequality, we obtain

$$\mathbb{E}_{\mathcal{K}}\left[\left\|\tilde{\mathbf{P}}_{FX^n\tilde{Y}^nMY^n} - \mathbf{Q}^{(2)}_{FX^n\tilde{Y}^nMY^n}\right\|_{TV}\right] \le \epsilon_n^{(rb)} + \epsilon_n^{(sw)}. \tag{102}$$

Note that the distortion under any realization of $\mathbf{Q}^{(2)}$, regardless of the codebook, is

$$\mathbb{E}_{Q^{(2)}}[d(X^n, Y^n)] = \mathbb{E}_{Q^{(2)}}[d(X^n, Y^n)] \tag{103}$$

$$= \mathbb{E}_{\overline{P}}[d(X, Y)]. \tag{104}$$

Applying Property $1(b)$, we can obtain

$$\mathbb{E}_{\mathcal{K}}\left[\mathbb{E}_{\tilde{\mathbf{P}}}[d(X^n, Y^n)]\right] \leq \mathbb{E}_{\overline{P}}[d(X, Y)] + d_{max}(\epsilon_n^{(rb)} + \epsilon_n^{(sw)}). \tag{105}$$

Furthermore, by symmetry and the law of total expectation, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{K}}\left[\mathbb{E}_{\tilde{\mathbf{P}}}[d(X^n, Y^n)]\right] \\
&= \mathbb{E}_F\left[\mathbb{E}_{\mathcal{K}}\left[\mathbb{E}_{\tilde{\mathbf{P}}}[d(X^n, Y^n)] \mid F\right]\right] && (106) \\
&= \mathbb{E}_{\mathcal{K}}\left[\mathbb{E}_{\tilde{\mathbf{P}}}[d(X^n, Y^n)] \mid F = 1\right] && (107) \\
&= \mathbb{E}_{\mathcal{K}}\left[\mathbb{E}_{\mathbf{P}}[d(X^n, Y^n)]\right], && (108)
\end{aligned}$$

where the last equality comes from the observation in (94).

Finally, applying the random coding argument, there exists a code that gives

$$\mathbb{E}_P[d(X^n, Y^n)] \leq \mathbb{E}_{\overline{P}}[d(X, Y)] + d_{max}\left(\epsilon_n^{(rb)} + \epsilon_n^{(sw)}\right),$$

which is less than D for n large enough.

C. Comparing the Likelihood Encoder with Proportional-Probability Encoder

Let us now compare the achievability proofs using the likelihood encoder approach and the *proportional-probability encoder* (random binning based) approach for the point-to-point rate distortion function.

We shall notice that the error term in the likelihood encoder approach only arises from the soft-covering lemma, while the error terms in the proportional-probability approach come from two places, random binning and MSLC decoding.

Next, we will provide a non-asymptotic comparison between the two approaches with respect to excess distortion.

Some asymptotic analysis was given in [24] on channel coding with random binning. We can extend this to give non-asymptotic bounds for source coding problems also. Using Theorems 1 and 2 from [24], we can obtain the following theorem.

Theorem 2. *The excess distortion* $\mathbb{P}_P[d(X^n, Y^n) > D]$ *using the proportional-probability encoder is upper bounded by*

$$\inf_{\overline{P}_{Y|X}} \left\{ \exp\left(-n\eta(\overline{P}_{Y|X})\right) + \sigma_n(\overline{P}_{Y|X}) \right\} \tag{109}$$

where

$$\sigma_n(\overline{P}_{Y|X}) = \inf_{R' \in (H(Y)-R, H(Y|X))} \{A_n + B_n\} \tag{110}$$

and

$$A_n = \inf_{\delta \in (0, H(Y|X)-R')} \left\{ \mathbb{P}_{\overline{P}}\left[-\log \overline{P}_{Y^n|X^n}(Y^n|X^n) \le n(R'+\delta)\right] + \frac{1}{\sqrt{2}} 2^{-\frac{n\delta}{2}} \right\} \tag{111}$$

$$B_n = \inf_{\tau > 0} \left\{ \mathbb{P}_{\overline{P}}\left[n(R+R') - h(Y^n) \le n\tau\right] + 3 \times 2^{-n\tau} \right\}. \tag{112}$$

We can further bound the quantities in A_n and B_n in Theorem 2 by the Chernoff inequality following the steps (81) through (84) and obtain the following exponential forms:

$$\begin{aligned} & \mathbb{P}_{\overline{P}}\left[-\log \overline{P}_{Y^n|X^n}(Y^n|X^n) \le n(R'+\delta)\right] \\ \le \ & \inf_{\beta_1 > 0} \left\{ \exp\left(-n \log\left(\mathbb{E}_{\overline{P}}\left[2^{\beta_1\left(R'+\delta-\log\frac{1}{\overline{P}_{Y|X}(Y|X)}\right)}\right]\right)^{-1}\right) \right\}, \end{aligned} \tag{113}$$

$$\begin{aligned} & \mathbb{P}_{\overline{P}}\left[n(R+R') - h(Y^n) \le n\tau\right] \\ \le \ & \inf_{\beta_2 > 0} \left\{ \exp\left(-n \log\left(\mathbb{E}_{\overline{P}}\left[2^{\beta_2\left(\log\frac{1}{\overline{P}_Y(Y)}-R-R'+\tau\right)}\right]\right)^{-1}\right) \right\}. \end{aligned} \tag{114}$$

D. Numerical Example

Next, we would like to compare the bounds given by the likelihood encoder in Theorem 1 and given by the proportional-probability encoder in Theorem 2.

Here we give a numerical comparison between the likelihood encoder and the proportional-probability encoder for a Bernoulli $\frac{1}{2}$ source and Hamming distortion. For simplicity, we consider only symmetric test channels of the form $\overline{P}_{Y|X}(0|0) = \overline{P}_{Y|X}(1|1) = a_0$.

Assume $D < \frac{1}{2}$ and fix a_0. Observe that $\eta(a_0) \triangleq \eta(\overline{P}_{Y|X})$ is a term shared by both the likelihood encoder and the proportional-probability encoder methods and it takes the following form:

$$\eta(a_0) = -\log_2\left(a_0 2^{-\beta^* D} + (1-a_0) 2^{\beta^*(1-D)}\right), \tag{115}$$

where

$$\beta^* = \log_2 \frac{Da_0}{(1-D)(1-a_0)}. \tag{116}$$

For a Bernoulli $\frac{1}{2}$ source, the quantities from the likelihood encoder satisfies

$$\check{I}_\alpha(a_0) \triangleq \check{I}_{\overline{P},\alpha} = \bar{I}_{\overline{P},\alpha} = 1 + \frac{1}{\alpha-1}\log_2\left(a_0^\alpha + (1-a_0)^\alpha\right) \tag{117}$$

$$\gamma(a_0) = \max_{\alpha\geq 1, \alpha'\leq 2} \frac{\alpha-1}{2\alpha-\alpha'}\left(R - 1 + \frac{\alpha'-2}{\alpha-1}\log_2(a_0^\alpha + (1-a_0)^\alpha) - \log_2(a_0^{\alpha'} + (1-a_0)^{\alpha'})\right). \tag{118}$$

Observe that the first term in B_n given in (112) is deterministic; therefore, we can choose

$$\tau^* = R + R' - 1. \tag{119}$$

The optimum β_1 in (113) is given by

$$\beta_1^* = \left[\log_{\frac{a_0}{1-a_0}}\left(-\frac{R'+\delta+\log_2(1-a_0)}{R'+\delta+\log_2(a_0)}\right) - 1\right]^+. \tag{120}$$

Consequently, the exponent of the first term of A_n is given by

$$A_{1n}(R',\delta,a_0) \triangleq -\log_2\left(a_0 2^{\beta_1^*(R'+\delta+\log_2(a_0))} + (1-a_0)2^{\beta_1^*(R'+\delta+\log_2(1-a_0))}\right). \tag{121}$$

Let us define

$$\lambda(a_0) \triangleq \max_{R',\delta}\left(R+R'-1, \frac{\delta}{2}, A_1(R',\delta,a_0)\right),$$

where the domains of R' and δ are omitted.

To summarize, for the likelihood encoder, we still need to optimize over α and α', and for the proportional-probability encoder, we need to optimize over R' and δ. Finally, for both, we optimize over a_0. The derived error exponent bounds for the likelihood encoder and the proportional-probability encoder are given by the following, respectively:

$$\text{Error exponent for the likelihood encoder} = \max_{a_0}\min(\eta(a_0),\gamma(a_0)) \tag{122}$$

$$\text{Error exponent for the proportional-probability encoder} = \max_{a_0}\min(\eta(a_0),\lambda(a_0)). \tag{123}$$

Comparisons of the error exponents given in (122) and (123) are shown in Fig. 6, plotted as functions of D and R. The numerical comparisons show that the likelihood encoder has a better error exponent than the proportional-probability encoder, at least according to these derived upper bounds on the error.

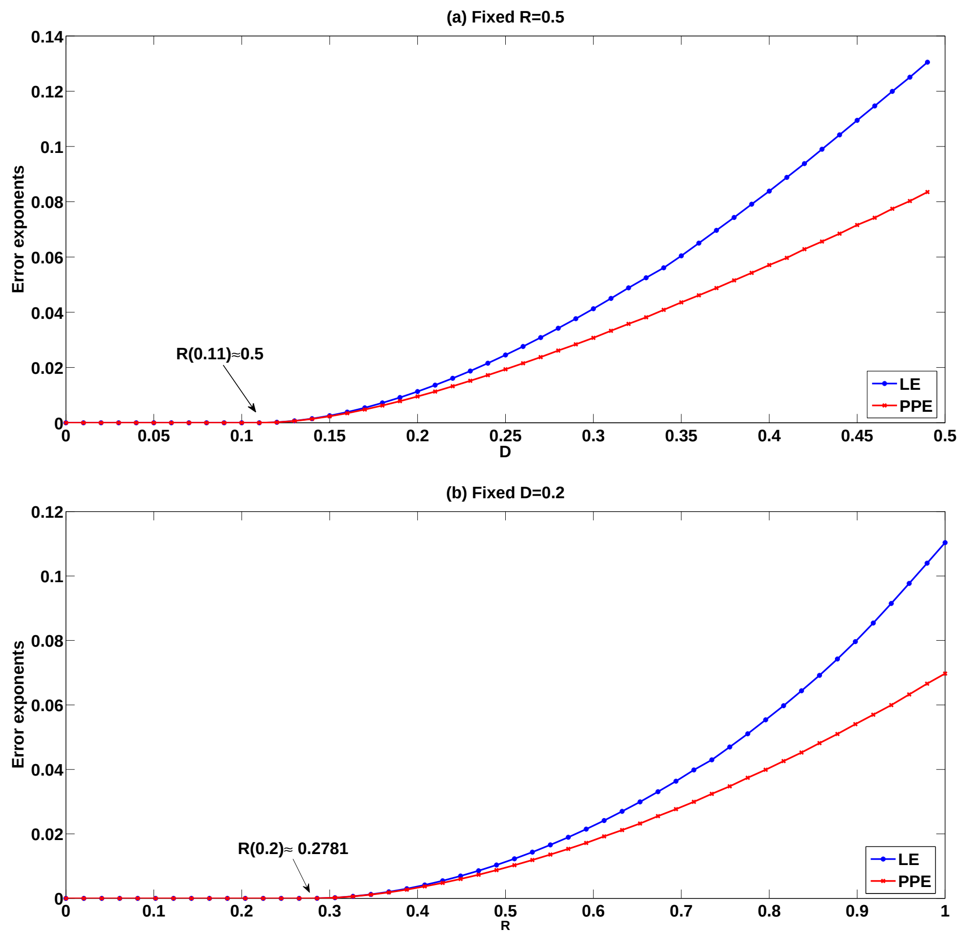

Fig. 6: Error exponents by the likelihood encoder and the proportional-probability encoder (random binning based analysis) for a Bernoulli $\frac{1}{2}$ source and Hamming distortion, in (a) as a function of D for fixed $R = \frac{1}{2}$, and in (b) as a function of R for fixed $D = 0.2$. Notice that for this particular example, the optimal excess error actually decays super-exponentially, but this is not achieved with either of the proof techniques discussed.

VIII. Conclusion

In this paper, we have demonstrated how the likelihood encoder can be used to obtain achievability results for various lossy source coding problems. The analysis of the likelihood encoder relies on the soft-covering lemma. Although the proof method is unusual, we hope to have demonstrated that this method of proof is simple, both conceptually and mechanically. The simplicity is accentuated when used for distributed source coding because it bypasses the need for a Markov lemma of any form and it avoids the technical complications that can arise in analyzing the decoder whenever random binning is involved in lossy compression. This proof method applies directly to continuous sources as well with no need for additional arguments, because the soft-covering lemma is not restricted to discrete sources. The likelihood encoder also simplifies analysis in secrecy settings, though this was not demonstrated within this paper.

A parallel comparison of the non-asymptotic performance of the likelihood encoder and the "proportional-probability encoder" has been provided along with a numerical example. In this example, the likelihood encoder achieves better error exponents than does the proportional-probability encoder.

References

[1] C. E. Shannon, “A mathematical theory of communication,” *Bell Sys. Tech. Journal*, vol. 27, pp. 379–423, 623–656, 1948.

[2] C. E. Shannon, “Coding theorems for a discrete source with a fidelity criterion,” *IRE National Convention Record, Part 4*, pp. 142–163, 1959.

[3] A. 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, 1976.

[4] S.-Y. Tung, *Multiterminal Source Coding*. PhD thesis, Cornell University, Ithaca, NY, May, 1978.

[5] T. Berger, “Multiterminal source coding,” *The Information Theory Approach to Communications*, vol. 229, pp. 171–231, 1977.

[6] T. Berger and R. W. Yeung, “Multiterminal source encoding with one distortion criterion,” *IEEE Transactions on Information Theory*, vol. 35, no. 2, pp. 228–236, 1989.

[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, 2012.

[8] A. El Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge University Press, 2011.

[9] P. Minero, S. H. Lim, and Y.-H. Kim, “Hybrid coding: An interface for joint source-channel coding and network communication,” *arXiv preprint arXiv:1306.0530*, 2013.

[10] A. Lapidoth and S. Tinguely, “Sending a bivariate gaussian over a gaussian mac,” *IEEE Transactions on Information Theory*, vol. 56, pp. 2714–2752, June 2010.

[11] P. Cuff and E. C. Song, “The likelihood encoder for source coding,” in *Proc. 2013 IEEE Information Theory Workshop (ITW)*, pp. 1–2, Sept 2013.

[12] A. Wyner, “The common information of two dependent random variables,” *IEEE Transactions on Information Theory*, vol. 21, no. 2, pp. 163–179, 1975.

[13] T. Han and S. Verdú, “Approximation theory of output statistics,” *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 752–772, 1993.

[14] P. Cuff, “Communication requirements for generating correlated random variables,” in *Proc. 2008 IEEE International Symposium on Information Theory (ISIT)*, pp. 1393–1397, 2008.

[15] P. Cuff, “Distributed channel synthesis,” *IEEE Transactions on Information Theory*, vol. 59, pp. 7071–7096, Nov 2013.

[16] C. Schieler and P. Cuff, “Rate-distortion theory for secrecy systems,” *CoRR*, vol. abs/1305.3905, 2013.

[17] P. Cuff, H. Permuter, and T. Cover, “Coordination capacity,” *IEEE Transactions on Information Theory*, vol. 56, pp. 4181–4206, Sept 2010.

[18] J. Jeon, “A generalized typicality for abstract alphabets,” *arXiv preprint arXiv:1401.6728*, 2014.

[19] P. Mitran, “Typical sequences for polish alphabets,” *arXiv preprint arXiv:1005.2321*, 2010.

[20] M. Yassaee, M. Aref, and A. Gohari, “A technique for deriving one-shot achievability results in network information theory,” in *Proc. 2013 IEEE International Symposium on Information Theory (ISIT)*, pp. 1287–1291, July 2013.

[21] C. Schieler and P. Cuff, “A connection between good rate-distortion codes and backward dmcs,” in *Proc. 2013 IEEE Information Theory Workshop (ITW)*, pp. 1–5, Sept 2013.

[22] I. Csiszar and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2011.

[23] M. Yassaee, M.-R. Aref, and A. Gohari, “Achievability proof via output statistics of random binning,” in *Proc. 2012 IEEE International Symposium on Information Theory (ISIT)*, pp. 1044–1048, July 2012.

[24] M. H. Yassaee, M. R. Aref, and A. Gohari, “Non-asymptotic output statistics of random binning and its applications,” *arXiv preprint arXiv:1303.0695*, 2013.

[25] D. Slepian and J. Wolf, “Noiseless coding of correlated information sources,” *IEEE Trans. Inf. Theory*, vol. 19, pp. 471–480, Jul 1973.

[26] T. Cover, “A proof of the data compression theorem of slepian and wolf for ergodic sources (corresp.),” *IEEE Trans. Inf. Theory*, vol. 21, pp. 226–228, Mar 1975.