

# Non-Asymptotic and Second-Order Achievability Bounds for Coding With Side-Information

Shun Watanabe *Member, IEEE*, Shigeaki Kuzuoka *Member, IEEE*, and Vincent Y. F. Tan *Member, IEEE*

**Abstract**—We present novel non-asymptotic or finite blocklength achievability bounds for three side-information problems in network information theory. These include (i) the Wyner-Ahlsvede-Körner (WAK) problem of almost-lossless source coding with rate-limited side-information, (ii) the Wyner-Ziv (WZ) problem of lossy source coding with side-information at the decoder and (iii) the Gel'fand-Pinsker (GP) problem of channel coding with noncausal state information available at the encoder. The bounds are proved using ideas from channel simulation and channel resolvability. Our bounds for all three problems improve on all previous non-asymptotic bounds on the error probability of the WAK, WZ and GP problems—in particular those derived by Verdú. Using our novel non-asymptotic bounds, we recover the general formulas for the optimal rates of these side-information problems. Finally, we also present achievable second-order coding rates by applying the multidimensional Berry-Esséen theorem to our new non-asymptotic bounds. Numerical results show that the second-order coding rates obtained using our non-asymptotic achievability bounds are superior to those obtained using existing finite blocklength bounds.

**Index Terms**—Source coding, channel coding, side-information, Wyner-Ahlsvede-Körner, Wyner-Ziv, Gel'fand-Pinsker, finite blocklength, non-asymptotic, second-order coding rates

## I. INTRODUCTION

The study of *network information theory* [1] involves characterizing the optimal rate regions or capacity regions for problems involving compression and transmission from multiple sources to multiple destinations. Apart from a few special channels or source models, optimal rate regions and capacity regions for many network information theory problems are still not known. In this paper, we revisit three coding problems whose asymptotic rate characterizations are well known. These include

- The *Wyner-Ahlsvede-Körner* (WAK) problem of almost-lossless source coding with rate-limited (aka coded) side-information [2], [3],
- The *Wyner-Ziv* (WZ) problem of lossy source coding with side-information at the decoder [4], and

This paper was presented in part at the 2013 IEEE International Symposium on Information Theory.

The first author is with the Department of Information Science and Intelligent Systems, University of Tokushima, 2-1, Minami-josanjima, Tokushima, 770-8506, Japan, and with the Institute for Systems Research, University of Maryland, College Park, MD 20742, USA, e-mail:shun-wata@is.tokushima-u.ac.jp.

The second author is with the Department of Computer and Communication Sciences, Wakayama University, Wakayama, 640-8510, Japan, e-mail:kuzuoka@ieee.org.

The third author is with the Department of Electrical and Computer Engineering and Department of Mathematics, National University of Singapore (NUS), e-mail:vtan@nus.edu.sg

Manuscript received ; revised

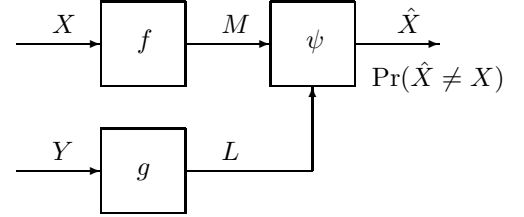


Fig. 1. Illustration of the WAK problem

- The *Gel'fand-Pinsker* (GP) problem of channel coding with noncausal state information at the encoder [5].

These problems fall under the class of coding problems with *side-information*. That is, a subset of terminals has access to either a correlated source or the state of the channel. In most cases, this knowledge helps to strictly improve the rates of compression or transmission over the case where there is no side-information.

While the study of asymptotic characterizations of network information theory problems has been of key interest and importance for the past 50 years, it is important to analyze non-asymptotic (or finite blocklength) limits of various network information theory problems. This is because there may be hard constraints on decoding complexity or delay in modern, heavily-networked systems. The paper derives new non-asymptotic bounds on the error probability for the WAK and GP problems as well as the probability of excess distortion for the WZ problem. Our bounds improve on all existing finite blocklength bounds for these problems such as those in [6]. In addition, we use these bounds to recover known general formulas [7]–[10] and we also derive achievable second-order coding rates [11], [12] for these side-information problems.

Traditionally, achievability proofs of the direct parts of these coding problems are common and involve a covering step, a packing step and the use of the Markov lemma [2] (also known as conditional typicality lemma in El Gamal and Kim [1]). As such to prove tighter bounds, it is necessary to develop new proof techniques in place of these lemmas [1] and their non-asymptotic versions [6], [7]. These new techniques are based on the notion of *channel resolvability* [7], [13], [14] and *channel simulation* [15]–[17]. We use the former in the helper's code construction.

To illustrate our idea at a high level, let us use the WAK problem as a canonical example of all three problems of interest. Recall that in the classical WAK problem, there is an independent and identically distributed (i.i.d.) joint source  $P_{XY}^n(x^n, y^n) = \prod_{i=1}^n P_{XY}(x_i, y_i)$ . The main source  $X^n \sim P_X^n$  is to be reconstructed almost losslessly from rate-

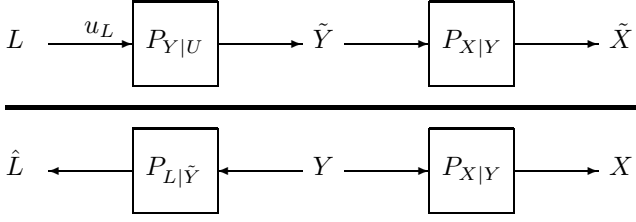


Fig. 2. High level description of helper's coding scheme for WAK. The upper row is a virtual scheme in which the uniform random number  $L$  is sent over channel  $P_{Y|U}$ . The lower row is the corresponding actual scheme in which message  $\hat{L}$  is stochastically generated via  $P_{L|\tilde{Y}}$ .

limited versions of both  $X^n$  and  $Y^n$ , where  $Y^n$  is a correlated random variable regarded as side-information. See Fig. 1. The compression rates of  $X^n$  and  $Y^n$  are denoted as  $R_1$  and  $R_2$  respectively. The optimal rate region is the set of rate pairs  $(R_1, R_2)$  for which there exists a *reliable* code, that is one whose error probability can be made arbitrarily small with increasing blocklengths. WAK [2], [3] showed that the optimal rate region is

$$R_1 \geq H(X|U), \quad R_2 \geq I(U; Y) \quad (1)$$

for some  $P_{U|Y}$ . For the direct part, the helper encoder compresses the side-information and transmits a description represented by  $U^n$ . By the covering lemma [1], this results in the rate constraint  $R_2 \geq I(U; Y)$ . The main encoder then uses binning [18] as in the achievability proof of the Slepian-Wolf theorem [19] to help the decoder recover  $X$  given the description  $U$ . This results in the rate constraint  $R_1 \geq H(X|U)$ .

The main idea in our proof of the new non-asymptotic upper bound on the error probability of the WAK problem is as follows: In the channel resolvability problem, for given channel  $P_{Y|U}$  and input distribution  $P_U$ , the goal is to approximate the output distribution  $P_Y$  (induced by  $(P_{Y|U}, P_U)$ ) by the output distribution  $P_{\tilde{Y}}$  of codewords for a codebook<sup>1</sup>  $\mathcal{C} = \{u_1, \dots, u_{|\mathcal{C}|}\}$  and the uniform random number  $L \in \mathcal{L}$ . Asymptotically, the approximation can be done successfully if the rate  $R_2$  of the random number  $L$  satisfies  $R_2 \geq I(U; Y)$ . In our helper's coding scheme (see Fig. 2), we use channel resolvability as a virtual scheme that is applied to the reverse test channel  $P_{Y|U}$  of a given test channel and the marginal  $P_U$  of the auxiliary random variable as the input distribution. Then, we flip the roles of the input and the output, i.e., we construct the conditional distribution  $P_{L|\tilde{Y}}$  from the joint distribution  $P_{L\tilde{Y}}$ . In the actual coding scheme, the message  $\hat{L}$  on  $\mathcal{L}$  is stochastically generated from helper's source  $Y$  via  $P_{L|\tilde{Y}}$ , which is known as the *likelihood encoder* [17]. Since the successful approximation in the channel resolvability guarantees  $P_{\tilde{Y}} \simeq P_Y$ , the joint distributions in the virtual scheme and the actual scheme are also close, i.e.,

$$P_{\tilde{L}XY} = P_Y P_{L|\tilde{Y}} P_{X|Y} \simeq P_{\tilde{Y}} P_{L|\tilde{Y}} P_{X|Y} = P_{L\tilde{X}\tilde{Y}}. \quad (2)$$

The decoder reproduces  $X$  via a Slepian-Wolf decoder by using  $u_{\hat{L}}$  as the side-information. Because of (2), the analysis

<sup>1</sup>Usually, the codebook is randomly generated according to the input distribution  $P_U$ .

of error probability can be done as if the decoder's observation is  $u_L$  and the underlying distribution is the virtual one  $P_{L\tilde{X}\tilde{Y}}$ . Moreover, by taking the average over the randomly generated codebook  $\mathcal{C}$ , since the codeword  $u_L$  is distributed according to  $P_U$ ,  $(\tilde{X}, u_L)$  behaves like  $(X, U)$ . Thus, the analysis of error probability can be done in the same manner as the Slepian-Wolf coding with full side-information  $U$ . The above argument enables us to circumvent the need to use the so-called piggy-back coding lemma (PBL) and the Markov lemma [2] which result in much poorer estimates on the error probability.

### A. Main Contributions

We now describe the three main contributions in this paper.

Our first main contribution in this paper is to show improved bounds on the probabilities of error for WAK, WZ and GP coding. We briefly describe the form of the bound for WAK coding here. The primary part of the new upper bound on the error probability  $P_e(\Phi)$  for WAK coding depends on two positive constants  $\gamma_b$  and  $\gamma_c$  and is essentially given by

$$P_e(\Phi) \lesssim \Pr(\mathcal{E}_c \cup \mathcal{E}_b) \quad (3)$$

where the *covering error* is

$$\mathcal{E}_c := \left\{ \log \frac{P_{Y|U}(U|Y)}{P_Y(Y)} \geq \gamma_c \right\} \quad (4)$$

and the *binning error* is

$$\mathcal{E}_b := \left\{ \log \frac{1}{P_{X|U}(X|U)} \geq \gamma_b \right\}. \quad (5)$$

The notation  $\lesssim$  is not meant to be precise and, in fact, we are dropping several residual terms that do not contribute to the second-order coding rates in the  $n$ -fold i.i.d. setting if  $\gamma_b$  and  $\gamma_c$  are chosen appropriately. This result is stated precisely in Theorem 5. From (3), we deduce that in the  $n$ -fold i.i.d. setting, if we choose  $\gamma_c$  and  $\gamma_b$  to be fixed numbers that are strictly larger than the mutual information  $I(U; Y)$  and the conditional entropy  $H(X|U)$  respectively, we are guaranteed that the error probability  $P_e(\Phi)$  decays to zero. This follows from Khintchine's law of large numbers [7, Ch. 1]. Thus, we recover the direct part of WAK's result. In fact, we can take this one step further (Theorem 12) to obtain an achievable *general formula* (in the sense of Verdú-Han [7], [20]) for the WAK problem with general source [7, Ch. 1]. This was previously done by Miyake-Kanaya [8] but their derivation is based on a different non-asymptotic formula more akin to Wyner's PBL. Also, since we have the freedom to design  $\gamma_c$  and  $\gamma_b$  as sequences instead of fixed positive numbers, if we let them be  $O(\frac{1}{\sqrt{n}})$ -larger than  $I(U; Y)$  and  $H(X|U)$ , then the error probability is smaller than a prescribed constant depending on the implied constants in the  $O(\cdot)$ -notations. This follows from the multivariate Berry-Esséen theorem [21]. This bound is useful because it is a *union* of two events and  $\mathcal{E}_c$  and  $\mathcal{E}_b$  are both information spectrum [7] events which are easy to analyze.

Secondly, the preceding discussion shows that the bound in (3) also yields an achievable second-order coding rate [11], [12]. However, unlike in the point-to-point setting [11], [12],

[22], the achievable second-order coding rate is expressed in terms of a so-called *dispersion matrix* [23]. We can easily show that if  $\mathcal{R}_{\text{WAK}}(n, \varepsilon)$  is the set of all rate pairs  $(R_1, R_2)$  for which there exists a length- $n$  WAK code with error probability not exceeding  $\varepsilon > 0$  (i.e., the  $(n, \varepsilon)$ -optimal rate region), then for any  $P_{U|Y}$ , the set

$$\left[ \frac{I(U; Y)}{H(X|U)} \right] + \frac{\mathcal{S}(\mathbf{V}, \varepsilon)}{\sqrt{n}} + O\left(\frac{\log n}{n}\right) \mathbf{1}_2 \quad (6)$$

is an inner bound to  $\mathcal{R}_{\text{WAK}}(n, \varepsilon)$ . In (6),  $\mathcal{S}(\mathbf{V}, \varepsilon) \subset \mathbb{R}^2$  denotes the analogue of the  $Q^{-1}$  function [23] and it depends on the covariance matrix of the so-called information-entropy density vector

$$\left[ \log \frac{P_{Y|U}(U|Y)}{P_Y(Y)} \quad \log \frac{1}{P_{X|U}(X|U)} \right]^T. \quad (7)$$

The precise statement for the second-order coding rate for the WAK problem is given in Theorem 15. We see from (6) that for a fixed test channel  $P_{U|Y}$ , the redundancy at blocklength  $n$  in order to achieve an error probability  $\varepsilon > 0$  is governed by the term  $\frac{\mathcal{S}(\mathbf{V}, \varepsilon)}{\sqrt{n}}$ . The pre-factor of this term  $\mathcal{S}(\mathbf{V}, \varepsilon)$ , is likened to the *dispersion* [22], [24]–[26], and depends not only the variances of the information and entropy densities but also their correlations.

Thirdly, we note that the same flavour of non-asymptotic bounds and second-order coding rates hold verbatim for the WZ and GP problems. In addition, since the canonical rate-distortion problem [27] is a special case of the WZ problem, we show that our non-asymptotic achievability bound for the WZ problem, when suitably specialized, yields the correct dispersion for lossy source coding [25], [26]. We do so using two methods: (i) the method of types [28] and (ii) results involving the  $D$ -tilted information [26]. Finally, we not only improve on the existing bounds for the GP problem [6], [10], but we also consider an almost sure cost constraint on the channel input.

## B. Related Work

Wyner [2] and Ahlswede-Körner [3] were the first to consider and solve (in the first-order sense) the problem of almost-lossless source coding with coded side information. Weak converses were proved in [2], [3] and a strong converse was proved in [29] using the “blowing-up lemma”. An information spectrum characterization was provided by Miyake and Kanaya [8] and Kuzuoka [30] leveraged on the non-asymptotic bound which can be extracted from [8] to derive the redundancy for the WAK problem. Verdú [6] strengthened the non-asymptotic bound and showed that the error probability for the WAK problem is essentially bounded as

$$P_e(\Phi) \lesssim \Pr(\mathcal{E}_c) + \Pr(\mathcal{E}_b), \quad (8)$$

which is the result upon using the union bound on our bound in (3). The notation  $\lesssim$  means that the residual terms do not affect the second-order coding rates.

Wyner and Ziv [4] derived the rate-distortion function for lossy source coding with decoder side-information. However, they do not consider the probability of excess distortion. Rather, the quantity of interest is the expected distortion.

The generalization of the WZ problem for general correlated sources was considered by Iwata and Muramatsu [9] who showed that the general WZ function can be written as a difference of a limit superior in probability and a limit inferior in probability, reflecting the covering and packing components in the classical achievability proof.

The problem of channel coding with noncausal random state information was solved by Gel’fand and Pinsker [5]. A general formula for the GP problem (with general channel and general state) was provided by Tan [10]. Tyagi and Narayan [31] proved the strong converse for this problem and used it to derive a sphere-packing bound. For both the WZ and GP problems, Verdú [6] used generalizations of the packing and covering lemmas in [1] to derive non-asymptotic bounds on the probability of excess distortion (for WZ) and the average error probability (for GP). However, they yield worse second-order rates because the main part of the bound is a sum of two or three probabilities as in (8), rather than the probability of the union as in (3).

In our work, we derive tight non-asymptotic bounds by using ideas from channel resolvability [13] [7, Ch. 6] and channel simulation [15]<sup>2</sup> to replace the covering part and Markov lemma. It was shown by Han and Verdú [13] that this problem is closely connected to channel coding and channel identification. Hayashi also studied the channel resolvability problem [14] and derived a non-asymptotic formula. We leverage on a key lemma in Hayashi [14] (and also Cuff [17]) to derive our bounds.

In [15], Bennett *et al.* proposed a problem to simulate a channel by the aid of common randomness. An application of the channel simulation to simulate the test channel in the rate-distortion problem was first investigated by Winter [16], and then extensively studied mainly in the field of the quantum information. Cuff investigated the trade-off between the rates of the message and common randomness for the channel simulation [17] (see also [33]). For a thorough list of literatures related to the channel simulation, see [17], [33]. In these works, channel resolvability is used as a building block for channel simulation. In particular, a code construction and analysis techniques that do not rely on the typicality argument were developed in [17]. The idea to use channel simulation instead of the Markov lemma is motivated by aforementioned papers, and our code construction and analysis are based on the ones in [17]. However, we stress that the derivations of our non-asymptotic bounds are not straightforward applications of channel simulation and channel resolvability. Indeed, our code construction is tailored to derive the bound as in (3), and we also introduce bounding techniques that have not appeared previously to the best of our knowledge.

Recently, Yassaee-Aref-Gohari (YAG) [34] proposed an alternative approach for channel simulation, in which they exploited the (multi-terminal version of) intrinsic randomness [7, Ch. 2] instead of channel resolvability. This approach is coined *output statistics of random binning* (OSRB). Although their approach is also used to replace the Markov lemma [2], it

<sup>2</sup>Steinberg and Verdú also studied the channel simulation problem [32]. However, their problem formulation is slightly different from the one in [15].

was not *a priori* yet clear when [34] was published whether our bounds can be also derived from the OSRB approach [34]. One of difficulties to apply the OSRB approach for non-asymptotic analysis is that the amount of common randomness that can be used in the channel simulation is limited by the randomness of sources involved in a coding problem, which is not the case with the approach using the channel resolvability. It was shown more recently by YAG [35] that a modification of the OSRB framework can, in fact, be used to obtain achievable dispersions of Marton's region for the broadcast channel [36] and the wiretap channel [37]. In fact, in another concurrent work by YAG [38], the authors derived very similar second-order results to the ones presented here. They derive bounds on the probability of error for Gel'fand-Pinsker, Heegard-Berger and multiple description coding [1] among others. The main idea in their proofs is to use the *stochastic likelihood coder* (SLC) and exploit the convexity of  $(x_1, x_2) \mapsto 1/(x_1 x_2)$  (for  $x_1, x_2 > 0$ ) to lower bound the probability of correct detection. Although the results in this paper and those in [38] partly overlap, the approaches to derive the results are different. To the best of our knowledge, this paper is the first to demonstrate usefulness of the channel simulation in non-asymptotic analysis of network information theory problems, which we believe to be interesting in its own right.

Our main motivation in this work is to derive tight non-asymptotic bounds on the error probabilities. We are also interested in second-order coding rates. The study of the asymptotic expansion of the logarithm of the maximum number codewords that are achievable for  $n$  uses a channel with maximum error probability no larger than  $\varepsilon$  was first done by Strassen [39]. This was re-popularized in recent times by Kontoyiannis [40], Baron-Khojastepour-Baraniuk [41], Hayashi [11], [12], and Polyanskiy-Poor-Verdú [22] among others. Second-order analysis for network information theory problems were considered in Tan and Kosut [23] as well as other authors [42]–[45]. However, this is the first work that considers second-order rates for problems with side-information.

### C. Paper Organization

In Section II, we state our notation and formally define the three coding problems with side-information. We then review existing first-order asymptotic results in Section III. In Section IV, we state our new non-asymptotic bounds for the three problems. We then use these bounds to re-derive (direct parts of) known general formulas [8]–[10] in Section V. Following that, we present achievable second-order coding rates for these coding problems. We will see that just as in the Slepian-Wolf setting [23], [44], the dispersion is in fact a matrix. In Section VII, we show via numerical examples that our non-asymptotic bounds lead to larger  $(n, \varepsilon)$ -rate regions compared with [6]. Concluding remarks and directions for future work are provided Section VIII. This paper only contains achievability bounds. In the conclusion, we also discuss the difficulties associated with obtaining non-asymptotic converse bounds. To ensure that the main ideas are seamlessly communicated in the main text, we relegate all proofs to the appendices.

## II. PRELIMINARIES

In this section, we introduce our notation and recall the WAK, WZ and GP problems.

### A. Notations

Random variables (e.g.,  $X$ ) and their realizations (e.g.,  $x$ ) are in capital and lower case respectively. All random variables take values in some alphabets which are denoted in calligraphic font (e.g.,  $\mathcal{X}$ ). The cardinality of  $\mathcal{X}$ , if finite, is denoted as  $|\mathcal{X}|$ . Let the random vector  $X^n := (X_1, \dots, X_n)$  and similarly for a realization  $x^n = (x_1, \dots, x_n)$ . The set of all distributions supported on alphabet  $\mathcal{X}$  is denoted as  $\mathcal{P}(\mathcal{X})$ . The set of all channels with the input alphabet  $\mathcal{X}$  and the output alphabet  $\mathcal{Y}$  is denoted by  $\mathcal{P}(\mathcal{Y}|\mathcal{X})$ . We will at times use the method of types [28]. The joint distribution induced by a marginal distribution  $P \in \mathcal{P}(\mathcal{X})$  and a channel  $V \in \mathcal{P}(\mathcal{Y}|\mathcal{X})$  is denoted interchangeably as  $P \times V$  or  $PV$ . This should be clear from the context.

For a sequence  $x^n = (x_1, \dots, x_n) \in \mathcal{X}^n$  in which  $|\mathcal{X}|$  is finite, its *type* or *empirical distribution* is the probability mass function  $P(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{x = x_i\}$  where the indicator function  $\mathbf{1}\{x \in \mathcal{A}\} = 1$  if  $x \in \mathcal{A}$  and 0 otherwise. The set of types with denominator  $n$  supported on alphabet  $\mathcal{X}$  is denoted as  $\mathcal{P}_n(\mathcal{X})$ . The *type class* of  $P$  is denoted as  $\mathcal{T}_P := \{x^n \in \mathcal{X}^n : x^n \text{ has type } P\}$ . For a sequence  $x^n \in \mathcal{T}_P$ , the set of sequences  $y^n \in \mathcal{Y}^n$  such that  $(x^n, y^n)$  has joint type  $PV = P(x)V(y|x)$  is the *V-shell*  $\mathcal{T}_V(x^n)$ . Let  $\mathcal{V}_n(\mathcal{Y}; P)$  be the family of stochastic matrices  $V : \mathcal{X} \rightarrow \mathcal{Y}$  for which the *V-shell* of a sequence of type  $P \in \mathcal{P}_n(\mathcal{X})$  is not empty. Information-theoretic quantities are denoted in the usual way. For example,  $I(X; Y)$  and  $I(P, V)$  denote the mutual information where the latter expression makes clear that the joint distribution of  $(X, Y)$  is  $PV$ . All logarithms are with respect to base 2 so information quantities are measured in bits.

The multivariate normal distribution with mean  $\mu$  and covariance matrix  $\Sigma$  is denoted as  $\mathcal{N}(\mu, \Sigma)$ . The complementary Gaussian cumulative distribution function  $Q(t) := \int_t^\infty \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du$  and its inverse is denoted as  $Q^{-1}(\varepsilon) := \min\{t \in \mathbb{R} : Q(t) \leq \varepsilon\}$ . Finally,  $|z|^+ := \max\{z, 0\}$ .

### B. The Wyner-Ahlsvede-Körner (WAK) Problem

In this section, we recall the WAK problem of lossless source coding with coded side-information [2], [3]. Let us consider a correlated source  $(X, Y)$  taking values in  $\mathcal{X} \times \mathcal{Y}$  and having joint distribution  $P_{XY}$ . Throughout,  $X$ , a discrete random variable, is the main source while  $Y$  is the helper or side-information. The WAK problem involves reconstructing  $X$  losslessly given rate-limited (or coded) versions of both  $X$  and  $Y$ . See Fig. 1.

**Definition 1.** A (possibly stochastic) source coding with side-information code or Wyner-Ahlsvede-Körner (WAK) code  $\Phi = (f, g, \psi)$  is a triple of mappings that includes two encoders  $f : \mathcal{X} \rightarrow \mathcal{M}$  and  $g : \mathcal{Y} \rightarrow \mathcal{L}$  and a decoder

$\psi : \mathcal{M} \times \mathcal{L} \rightarrow \mathcal{X}$ . The error probability of the WAK code  $\Phi$  is defined as

$$P_e(\Phi) := \Pr\{X \neq \psi(f(X), g(Y))\}. \quad (9)$$

In the following, we may call  $f$  as the main encoder and  $g$  the helper.

In Section VI, we consider  $n$ -fold i.i.d. extensions of  $X$  and  $Y$ , denoted as  $X^n$  and  $Y^n$ . In this case, we use the subscript  $n$  to specify the blocklength, i.e., the code is  $\Phi_n = (f_n, g_n, \psi_n)$  and the compression index sets are  $\mathcal{M}_n = f_n(\mathcal{X}^n)$  and  $\mathcal{L}_n = g_n(\mathcal{Y}^n)$ . In this case, we can define the pair of rates of the code  $\Phi_n$  as

$$R_1(\Phi_n) := \frac{1}{n} \log |\mathcal{M}_n|, \quad (10)$$

$$R_2(\Phi_n) := \frac{1}{n} \log |\mathcal{L}_n|. \quad (11)$$

**Definition 2.** The  $(n, \varepsilon)$ -optimal rate region for the WAK problem  $\mathcal{R}_{\text{WAK}}(n, \varepsilon)$  is defined as the set of all pairs of rates  $(R_1, R_2)$  for which there exists a blocklength- $n$  WAK code  $\Phi_n$  with rates at most  $(R_1, R_2)$  and with error probability not exceeding  $\varepsilon$ . In other words,

$$\mathcal{R}_{\text{WAK}}(n, \varepsilon) := \left\{ (R_1, R_2) \in \mathbb{R}_+^2 : \exists \Phi_n \text{ s.t.} \right. \\ \left. \begin{aligned} \frac{1}{n} \log |\mathcal{M}_n| &\leq R_1, \\ \frac{1}{n} \log |\mathcal{L}_n| &\leq R_2, \\ P_e(\Phi_n) &\leq \varepsilon \end{aligned} \right\} \quad (12)$$

We also define the asymptotic rate regions

$$\mathcal{R}_{\text{WAK}}(\varepsilon) := \text{cl} \left[ \bigcup_{n \geq 1} \mathcal{R}_{\text{WAK}}(n, \varepsilon) \right], \quad (13)$$

$$\mathcal{R}_{\text{WAK}} := \bigcap_{0 < \varepsilon < 1} \mathcal{R}_{\text{WAK}}(\varepsilon). \quad (14)$$

where cl denotes set closure in  $\mathbb{R}^2$ .

In the following, we will provide an inner bound to  $\mathcal{R}_{\text{WAK}}(n, \varepsilon)$  that improves on inner bounds that can be derived from previously obtained non-asymptotic bounds on  $P_e(\Phi_n)$  [6], [30].

### C. The Wyner-Ziv (WZ) Problem

In this section, we recall the WZ problem of lossy source coding with full side-information at the decoder [4]. Here, as in the WAK problem, we have a correlated source  $(X, Y)$  taking values in  $\mathcal{X} \times \mathcal{Y}$  and having joint distribution  $P_{XY}$ . Again,  $X$  is the main source and  $Y$  is the helper or side-information. Neither  $X$  nor  $Y$  has to be a discrete random variable. Unlike the WAK problem, it is not required to reconstruct  $X$  exactly, rather a distortion  $D$  between  $X$  and its reproduction  $\hat{X}$  is allowed. Let  $\hat{\mathcal{X}}$  be the reproduction alphabet and let  $d : \mathcal{X} \times \hat{\mathcal{X}} \rightarrow [0, \infty)$  be a bounded distortion measure such that for every  $x \in \mathcal{X}$  there exists a  $\hat{x} \in \hat{\mathcal{X}}$  such that  $d(x, \hat{x}) = 0$  and  $\max_{x, \hat{x}} d(x, \hat{x}) = D_{\max} < \infty$ . See Fig. 3.

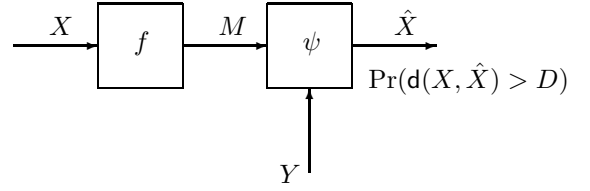


Fig. 3. Illustration of the WZ problem with probability of excess distortion criterion

**Definition 3.** A (possibly stochastic) lossy source coding with side-information or Wyner-Ziv (WZ) code  $\Phi = (f, \psi)$  is a pair of mappings that includes an encoder  $f : \mathcal{X} \rightarrow \mathcal{M}$  and a decoder  $\psi : \mathcal{M} \times \mathcal{Y} \rightarrow \hat{\mathcal{X}}$ . The probability of excess distortion for the WZ code  $\Phi$  at distortion level  $D$  is defined as

$$P_e(\Phi; D) := \Pr\{d(X, \psi(f(X), Y)) > D\}. \quad (15)$$

We will again consider  $n$ -fold extensions of  $X$  and  $Y$ , denoted as  $X^n$  and  $Y^n$  in Section VI. The code is indexed by the blocklength as  $\Phi_n = (f_n, \psi_n)$ . Furthermore, the compression index set is denoted as  $\mathcal{M}_n = f_n(\mathcal{X}^n)$ . The rate of the code  $\Phi_n$  is defined as

$$R(\Phi_n) := \frac{1}{n} \log |\mathcal{M}_n|. \quad (16)$$

The distortion between two length- $n$  sequences  $x^n \in \mathcal{X}^n$  and  $\hat{x}^n \in \hat{\mathcal{X}}^n$  is defined as

$$d_n(x^n, \hat{x}^n) := \frac{1}{n} \sum_{i=1}^n d(x_i, \hat{x}_i). \quad (17)$$

**Definition 4.** The  $(n, \varepsilon)$ -Wyner-Ziv rate-distortion region  $\mathcal{R}_{\text{WZ}}(n, \varepsilon) \subset \mathbb{R}_+^2$  is the set of all rate-distortion pairs  $(R, D)$  for which there exists a blocklength- $n$  WZ code  $\Phi_n$  at distortion level  $D$  with rate at most  $R$  and probability of excess distortion not exceeding  $\varepsilon$ . In other words,

$$\mathcal{R}_{\text{WZ}}(n, \varepsilon) := \left\{ (R, D) \in \mathbb{R}_+^2 : \exists \Phi_n \text{ s.t.} \right. \\ \left. \begin{aligned} \frac{1}{n} \log |\mathcal{M}_n| &\leq R, \\ P_e(\Phi_n; D) &\leq \varepsilon \end{aligned} \right\} \quad (18)$$

We also define the asymptotic rate-distortion regions

$$\mathcal{R}_{\text{WZ}}(\varepsilon) := \text{cl} \left[ \bigcup_{n \geq 1} \mathcal{R}_{\text{WZ}}(n, \varepsilon) \right], \quad (19)$$

$$\mathcal{R}_{\text{WZ}} := \bigcap_{0 < \varepsilon < 1} \mathcal{R}_{\text{WZ}}(\varepsilon). \quad (20)$$

The  $(n, \varepsilon)$ -Wyner-Ziv rate-distortion function  $R_{\text{WZ}}(n, \varepsilon, D)$  is defined as

$$R_{\text{WZ}}(n, \varepsilon, D) := \inf\{R : (R, D) \in \mathcal{R}_{\text{WZ}}(n, \varepsilon)\} \quad (21)$$

We also define the asymptotic rate-distortion functions

$$R_{\text{WZ}}(\varepsilon, D) = \inf\{R : (R, D) \in \mathcal{R}_{\text{WZ}}(\varepsilon)\} \quad (22)$$

$$R_{\text{WZ}}(D) = \lim_{\varepsilon \rightarrow 0} R_{\text{WZ}}(\varepsilon, D) \quad (23)$$

Note that the use of the limit (as opposed to the limit superior or limit inferior) in (23) is justified because  $R_{\text{WZ}}(\varepsilon, D)$

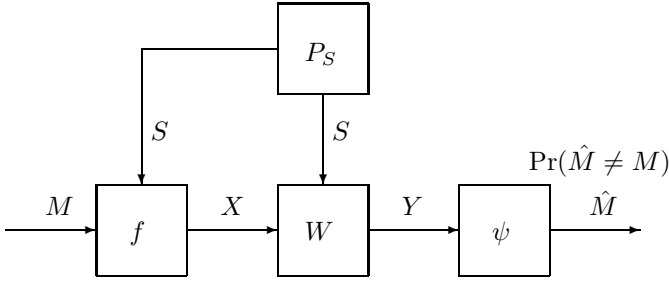


Fig. 4. Illustration of the GP problem. The channel input  $X$  must satisfy (25).

is, from its definition, monotonically non-increasing in  $\varepsilon$ . In the sequel, we will provide an inner bound to  $\mathcal{R}_{\text{WZ}}(n, \varepsilon)$  and thus an upper bound on  $R_{\text{WZ}}(n, \varepsilon, D)$  by appealing to a new non-asymptotic upper bound on the probability of excess distortion  $P_e(\Phi_n; D)$ . In addition, note that if  $Y = \emptyset$ , i.e., side-information is not available, this reduces to the point-to-point rate-distortion (lossy source coding) problem.

Conventionally [1], [4], the WZ problem is stated not with the probability of excess distortion criterion but with the *average fidelity criterion*. That is, the requirement that  $P_e(\Phi_n; D) \rightarrow 0$  (implicit in (23)) is replaced by

$$\limsup_{n \rightarrow \infty} \mathbb{E}[d_n(X^n, \psi_n(f_n(X^n), Y^n))] \leq D. \quad (24)$$

#### D. The Gel'fand-Pinsker (GP) Problem

In the previous two subsections, we dealt exclusively with source coding problems, either lossless (WAK) or lossy (WZ). In this section, we review the setup of the GP problem [5] which involves channel coding with noncausal state information at the encoder. It is the dual to the WZ problem [46]. In this problem, there is a state-dependent channel  $W : \mathcal{X} \times \mathcal{S} \rightarrow \mathcal{Y}$  and a random variable representing the state  $S$  with distribution  $P_S$  taking values in some set  $\mathcal{S}$ . A message  $M$  chosen uniformly at random from  $\mathcal{M}$  is to be sent and the encoder has information about which message is to be sent as well as the channel state information  $S$ , which is known *noncausally*. (Noncausality only applies when the blocklength is larger than 1.) It is assumed that the message and the state are independent. Let  $g : \mathcal{X} \rightarrow [0, \infty)$  be some cost function. The encoder  $f$  encodes the message and state into a codeword (channel input)  $X = f(M, S)$  that satisfies the cost constraint

$$g(X) \leq \Gamma, \quad (25)$$

for some  $\Gamma \geq 0$  with high probability. See precise definition/requirement in (26) as well as Proposition 1. The decoder receives the channel output  $Y | \{X = x, S = s\} \sim W(\cdot | x, s)$  and decides which message was sent via a decoder  $\psi : \mathcal{Y} \rightarrow \mathcal{M}$ . See Fig. 4. More formally, we have the following definition.

**Definition 5.** A (possibly stochastic) code for the channel coding problem with noncausal state information or Gel'fand-Pinsker (GP) code  $\Phi = (f, \psi)$  is a pair of mappings that includes an encoder  $f : \mathcal{M} \times \mathcal{S} \rightarrow \mathcal{X}$  and a decoder

$\psi : \mathcal{Y} \rightarrow \mathcal{M}$ . The average probability of error for the GP code is defined as

$$P_e(\Phi; \Gamma) := \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \sum_{s \in \mathcal{S}} P_S(s) \sum_{y \in \mathcal{Y}} W(y | f(m, s), s) \mathbf{1} \{g(f(m, s)) > \Gamma \cup y \in \mathcal{Y} \setminus \psi^{-1}(m)\}. \quad (26)$$

More simply,  $P_e(\Phi; \Gamma) = \Pr(\{g(f(M, S)) > \Gamma\} \cup \{\hat{M} \neq M\})$  where  $M$  is uniform on  $\mathcal{M}$  and independent of  $S \sim P_S$ ,  $\hat{M} := \psi(Y)$  and  $Y$  is the random variable whose conditional distribution given  $M = m$  and  $S = s$  is  $W(\cdot | f(m, s), s)$ .

The following proposition, which will be proved in Appendix A, guarantees that we can always convert a code in the sense of Definition 5 into a code in the sense of an almost sure cost constraint.

**Proposition 1** (Expurgated Code). *Let the set of admissible inputs in  $\mathcal{X}$  be*

$$\mathcal{T}_g^{\text{GP}}(\Gamma) := \{x \in \mathcal{X} : g(x) \leq \Gamma\}. \quad (27)$$

For any (stochastic) encoder  $P_{X|MS}$  (this plays the role of  $f$  in Definition 5) and decoder  $P_{\hat{M}|Y}$  (this plays the role of  $\psi$  in Definition 5), there exists an encoder  $\tilde{P}_{X|MS}$  such that

$$\tilde{P}_X(\mathcal{T}_g^{\text{GP}}(\Gamma)) = 1 \quad (28)$$

and

$$\tilde{P}_{MSXY\hat{M}}[m \neq \hat{m}] \leq P_{MSXY\hat{M}}[g(x) > \Gamma \cup m \neq \hat{m}], \quad (29)$$

where

$$P_{MSXY\hat{M}} := P_M P_S P_{X|MS} W P_{\hat{M}|Y}, \quad (30)$$

$$\tilde{P}_{MSXY\hat{M}} := P_M P_S \tilde{P}_{X|MS} W P_{\hat{M}|Y}. \quad (31)$$

From Proposition 1, noting that  $P_e((P_{X|MS}, P_{\hat{M}|Y}); \Gamma) = P_{MSXY\hat{M}}[g(x) > \Gamma \cup m \neq \hat{m}]$ , we see that the constraint in (25) is equivalent to  $g(X) \leq \Gamma$  *almost surely* (implied by (28)). For the purposes of deriving channel simulation-based bounds in Section IV-C, it is easier to work with the error criterion in (26) so we adopt Definition 5.

In order to obtain achievable second-order coding rates for the GP problem, we consider  $n$ -fold i.i.d. extensions of the channel and state. Hence, for every  $(s^n, x^n, y^n)$ , we have  $W^n(y^n | x^n, s^n) = \prod_{i=1}^n W(y_i | x_i, s_i)$  and the state  $S^n$  evolves in a stationary, memoryless fashion according to  $P_S$ . For blocklength  $n$ , the code and message set are denoted as  $\Phi_n = (f_n, \psi_n)$  and  $\mathcal{M}_n$  respectively. The cost function is denoted as  $g_n : \mathcal{X}^n \rightarrow [0, \infty)$  and is defined as the average of the per-letter costs, i.e.,

$$g_n(x^n) := \frac{1}{n} \sum_{i=1}^n g(x_i) \quad (32)$$

For example, in the Gaussian GP problem (which is also known as *dirty paper coding* [47]),  $g(x) = x^2$ . This corresponds to a power constraint and  $\Gamma$  is the upper bound on the permissible power. The rate of the code is the normalized logarithm of the number of messages, i.e.,

$$R(\Phi_n) := \frac{1}{n} \log |\mathcal{M}_n|. \quad (33)$$

**Definition 6.** The  $(n, \varepsilon)$ -GP capacity-cost region  $\mathcal{C}_{\text{GP}}(n, \varepsilon) \subset \mathbb{R}_+^2$  is the set of all rate-cost pairs  $(R, \Gamma)$  for which there exists a blocklength- $n$  GP code  $\Phi_n$  with cost not exceeding  $\Gamma$ , with rate at least  $R$  and probability of error not exceeding  $\varepsilon$ . In other words,

$$\mathcal{C}_{\text{GP}}(n, \varepsilon) := \left\{ (R, \Gamma) \in \mathbb{R}_+^2 : \exists \Phi_n \text{ s.t.} \right. \\ \left. \frac{1}{n} \log |\mathcal{M}_n| \geq R, \right. \\ \left. P_e(\Phi_n; \Gamma) \leq \varepsilon \right\}. \quad (34)$$

We also define the asymptotic capacity-cost regions

$$\mathcal{C}_{\text{GP}}(\varepsilon) := \text{cl} \left[ \bigcup_{n \geq 1} \mathcal{C}_{\text{GP}}(n, \varepsilon) \right], \quad (35)$$

$$\mathcal{C}_{\text{GP}} := \bigcap_{0 < \varepsilon < 1} \mathcal{C}_{\text{GP}}(\varepsilon). \quad (36)$$

The  $(n, \varepsilon)$ -capacity-cost function  $C_{\text{GP}}(n, \varepsilon, \Gamma)$  is defined as

$$C_{\text{GP}}(n, \varepsilon, \Gamma) := \sup \{ R : (R, \Gamma) \in \mathcal{C}_{\text{GP}}(n, \varepsilon) \} \quad (37)$$

We also define the asymptotic capacity-cost functions

$$C_{\text{GP}}(\varepsilon, \Gamma) := \sup \{ R : (R, \Gamma) \in \mathcal{C}_{\text{GP}}(\varepsilon) \} \quad (38)$$

$$C_{\text{GP}}(\Gamma) := \lim_{\varepsilon \rightarrow 0} C_{\text{GP}}(\varepsilon, \Gamma) \quad (39)$$

If the cost constraint (25) is absent (i.e., every codeword in  $\mathcal{X}^n$  is admissible), we will write  $C_{\text{GP}}(n, \varepsilon)$  instead of  $C_{\text{GP}}(n, \varepsilon, \infty)$ ,  $P_e(\Phi_n)$  instead of  $P_e(\Phi_n; \infty)$  and so on.

Once again, the limit in (39) exists because the function  $C_{\text{GP}}(\varepsilon, \Gamma)$  is monotonically non-decreasing in  $\varepsilon$ . In the sequel, we will provide a lower bound on  $C_{\text{GP}}(n, \varepsilon, \Gamma)$  by appealing to a new non-asymptotic upper bound on the average probability of error  $P_e(\Phi_n; \Gamma)$ .

### III. REVIEW OF EXISTING FIRST-ORDER RESULTS

#### A. First-Order Result for the WAK Problem

Let  $\mathcal{P}(P_{XY})$  be the set of all joint distributions  $P_{UXY} \in \mathcal{P}(\mathcal{U} \times \mathcal{X} \times \mathcal{Y})$  such that the  $\mathcal{X} \times \mathcal{Y}$ -marginal of  $P_{UXY}$  is the source distribution  $P_{XY}$ ,  $U - Y - X$  forms a Markov chain in that order and<sup>3</sup>  $|\mathcal{U}| \leq |\mathcal{Y}| + 1$ . Define

$$\mathcal{R}_{\text{WAK}}^* := \bigcup_{P_{UXY} \in \mathcal{P}(P_{XY})} \{ (R_1, R_2) \in \mathbb{R}_+^2 : R_1 \geq H(X|U), \\ R_2 \geq I(U; Y) \}. \quad (40)$$

Wyner [2] and Ahlswede-Körner [3] proved the following:

**Theorem 2** (Wyner [2], Ahlswede-Körner [3]). *For every  $0 < \varepsilon < 1$ , we have*

$$\mathcal{R}_{\text{WAK}}(\varepsilon) = \mathcal{R}_{\text{WAK}} = \mathcal{R}_{\text{WAK}}^*, \quad (41)$$

<sup>3</sup>The cardinality bound on  $\mathcal{U}$  in the definition of  $\mathcal{P}(P_{XY})$  is applied when we consider the single letter characterization  $\mathcal{R}_{\text{WAK}}^*$  and the inner bound to the  $(n, \varepsilon)$ -optimal rate region  $\mathcal{R}_{\text{WAK}}(n, \varepsilon)$ . It is not applied when we consider non-asymptotic analysis. Similar remarks are also applied for the WZ and GP problems.

where  $\mathcal{R}_{\text{WAK}}(\varepsilon)$  and  $\mathcal{R}_{\text{WAK}}$  are defined in (13) and (14) respectively.

To prove the direct part, Wyner used the PBL and the Markov lemma [2] while Ahlswede-Körner [3] used a maximal code construction. Only weak converses were provided in [2] and [3]. Ahlswede-Gács-Körner [29] proved the strong converse using entropy and image-size characterizations [28, Ch. 15], which are based on the so-called blowing-up lemma [28, Ch. 5]. See [28, Thm. 16.4].

#### B. First-Order Result for the WZ Problem

Let  $\mathcal{P}_D(P_{XY})$  be the set of all pairs  $(P_{UXY}, g)$  where  $P_{UXY} \in \mathcal{P}(\mathcal{U} \times \mathcal{X} \times \mathcal{Y})$  is a joint distribution and  $g : \mathcal{U} \times \mathcal{Y} \rightarrow \hat{\mathcal{X}}$  is a (reproduction) function such that the  $\mathcal{X} \times \mathcal{Y}$ -marginal of  $P_{UXY}$  is the source distribution  $P_{XY}$ ,  $U - X - Y$  forms a Markov chain in that order,  $|\mathcal{U}| \leq |\mathcal{X}| + 1$  and the distortion constraint is satisfied, i.e.,

$$\mathbb{E}[d(X, g(U, Y))] = \sum_{u, x, y} P_{UXY}(u, x, y) d(x, g(u, y)) \leq D. \quad (42)$$

In Section VI-B, we allow  $g$  to be stochastic (i.e., represented by a conditional probability mass function  $P_{\hat{X}|UY}$ ) but we still retain the use of the notation  $\mathcal{P}_D(P_{XY})$ . Define the function

$$R_{\text{WZ}}^*(D) := \min_{(P_{UXY}, g) \in \mathcal{P}_D(P_{XY})} I(U; X) - I(U; Y). \quad (43)$$

Note from Markovity that  $I(U; X) - I(U; Y) = I(U; X|Y)$ . Then, we have the following asymptotic characterization of the WZ rate-distortion function.

**Theorem 3** (Wyner-Ziv [4]). *We have*

$$R_{\text{WZ}}(D) = R_{\text{WZ}}^*(D), \quad (44)$$

where  $R_{\text{WZ}}(D)$  is defined in (23).

The direct part of the proof of the theorem in the original Wyner-Ziv paper [4] is based on the average fidelity criterion in (24). It relies on the *compress-bin* idea. That is, binning is used to reduce the rate of the description of the main source to the receiver. The encoder transmits the bin index and the decoder searches within that bin for the transmitted codeword. The reproduction function  $g$  is then used to reproduce the source to within a distortion  $D$ . To prove Theorem 3 for the probability of excess distortion criterion, we may use the new non-asymptotic bound in Section IV-B or the weaker non-asymptotic bounds in [9] or [6].

#### C. First-Order Result for the GP Problem

We conclude this section by stating the capacity of the GP problem [5]. Recall that in the GP problem, we have a channel  $W : \mathcal{X} \times \mathcal{S} \rightarrow \mathcal{Y}$  and a state distribution  $P_S \in \mathcal{P}(\mathcal{S})$ . Assume for simplicity that all alphabets are finite sets. Let  $\mathcal{P}_\Gamma(W, P_S)$  be the collection of all joint distributions  $P_{UXSY} \in \mathcal{P}(\mathcal{U} \times \mathcal{X} \times \mathcal{S} \times \mathcal{Y})$  such that the  $\mathcal{S}$ -marginal is  $P_S$ , the conditional distribution  $P_{Y|XS} = W$ ,  $U - (X, S) - Y$  forms a Markov chain in that order,

$$\mathbb{E}[g(X)] \leq \Gamma \quad (45)$$

and<sup>4</sup>  $|\mathcal{U}| \leq \min\{|\mathcal{X}||\mathcal{S}|, |\mathcal{S}| + |\mathcal{Y}|\}$ . Define the quantity

$$C_{\text{GP}}^*(\Gamma) := \max_{P_{UXSY} \in \mathcal{P}_{\Gamma}(W, P_S)} I(U; Y) - I(U; S), \quad (46)$$

where  $I(U; Y)$  and  $I(U; S)$  are computed with respect to the joint distribution  $P_{UXSY}$ . If there is no cost constraint (45), we simply write  $C_{\text{GP}}^*$  instead of  $C_{\text{GP}}^*(\infty)$ . Then, we have the following asymptotic characterization.

**Theorem 4** (Gel'fand-Pinsker [5]). *If the alphabets  $\mathcal{S}, \mathcal{X}$  and  $\mathcal{Y}$  are discrete, for every  $0 < \varepsilon < 1$ , we have*

$$C_{\text{GP}}(\varepsilon) = C_{\text{GP}} = C_{\text{GP}}^* \quad (47)$$

where  $C_{\text{GP}}(\varepsilon)$  and  $C_{\text{GP}}$  are defined in (38) and (39) respectively.

The direct part was proved using a covering-packing argument as well as the conditional typicality lemma (using the notion of strong typicality). Essentially, each message  $m \in \mathcal{M}$  is uniquely associated to a subcodebook of size  $L$ . To send message  $m$ , the encoder looks in the  $m$ -th subcodebook for a codeword that is jointly typical with the noncausal state. The decoder then searches for the unique subcodebook which contains at least one codeword that is jointly typical with the channel output. The weak converse in the original Gel'fand-Pinsker paper was proved using the Csiszár-sum-identity. See [1, Thm. 7.3]. In fact the weak converse shows that encoding function  $P_{X|US}$  can be restricted to the set of deterministic functions. Tyagi and Narayan proved a strong converse [31] using entropy and image-size characterizations via judicious choices of auxiliary channels. Their proof only applies to discrete memoryless channels with discrete state distribution without cost constraints.

#### IV. MAIN RESULTS: NOVEL NON-ASYMPTOTIC ACHIEVABILITY BOUNDS

In this section, we describe our results concerning novel non-asymptotic achievability bounds for the WAK, WZ and GP problems. We show using ideas from channel resolvability [7, Ch. 6] [13] [14] and channel simulation [15]–[17] that the bounds obtained by Verdú in [6] can be refined so as to obtain better second-order coding rates. The definition of and techniques involving channel resolvability and channel simulation are reviewed in Appendices B and C respectively. These are concepts that form crucial components of the proofs of the Channel-Simulation-type (CS-type) bounds in the sequel.

The following quantity, introduced in [17], will be used extensively in this section so we provide its definition here. For a joint distribution  $P_{UY} \in \mathcal{P}(\mathcal{U} \times \mathcal{Y})$  and a positive constant  $\gamma_c$ , define

$$\begin{aligned} \Delta(\gamma_c, P_{UY}) &:= \sum_{y \in \mathcal{Y}} P_Y(y) \\ &\times \sqrt{\sum_{u \in \mathcal{U}} P_{U|Y}(u|y) \frac{P_{Y|U}(y|u)}{P_Y(y)} \mathbf{1} \left\{ \log \frac{P_{Y|U}(y|u)}{P_Y(y)} \leq \gamma_c \right\}} \end{aligned} \quad (48)$$

<sup>4</sup>Because of cost constraint, the second entry of the cardinality bound is increased by one compared to the case without cost constraint [1, Thm. 7.3].

By applying the Jensen inequality, we find that  $\Delta(\gamma_c, P_{UY})$  has the property that

$$\Delta(\gamma_c, P_{UY}) \leq \sqrt{2\gamma_c}. \quad (49)$$

#### A. Novel Non-Asymptotic Achievability Bound for the WAK Problem

Fix an auxiliary alphabet  $\mathcal{U}$  and a joint distribution  $P_{UXY} \in \mathcal{P}(P_{XY})$ . See definition of  $\mathcal{P}(P_{XY})$  prior to (40). For arbitrary non-negative constants  $\gamma_b$  and  $\gamma_c$ , define two sets

$$\mathcal{T}_b^{\text{WAK}}(\gamma_b) := \left\{ (u, x) \in \mathcal{U} \times \mathcal{X} : \log \frac{1}{P_{X|U}(x|u)} \leq \gamma_b \right\}, \quad (50)$$

$$\mathcal{T}_c^{\text{WAK}}(\gamma_c) := \left\{ (u, y) \in \mathcal{U} \times \mathcal{Y} : \log \frac{P_{Y|U}(y|u)}{P_Y(y)} \leq \gamma_c \right\}. \quad (51)$$

These sets are similar to the *typical* sets used extensively in network information theory [1] but note that these sets only involve the entropy and information densities. Consequently, the probabilities of these sets (events) are entropy and information spectrum quantities [7]. The subscripts b and c refer respectively to *binning* and *covering*. Similar subscripts and will be used in the sequel for the other side-information problems to demonstrate the similarities between the proof techniques all of which leverage on ideas from channel resolvability [7, Ch. 6] [14] and channel simulation [15]–[17].

**Theorem 5** (CS-type bound for WAK coding). *For arbitrary  $\gamma_b, \gamma_c \geq 0$ , there exists a WAK code  $\Phi$  with error probability satisfying*

$$\begin{aligned} P_e(\Phi) &\leq P_{UXY} [(u, x) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)^c \cup (u, y) \in \mathcal{T}_c^{\text{WAK}}(\gamma_c)^c] \\ &\quad + \frac{1}{|\mathcal{M}|} \sum_{(u, \tilde{x}) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)} P_U(u) + \frac{\Delta(\gamma_c, P_{UY})}{2\sqrt{|\mathcal{L}|}}. \end{aligned} \quad (52)$$

See Appendix D for the proof of Theorem 5. Observe that the primary novelty of the bound in (52) lies in the fact that both error events  $\{(u, x) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)^c\}$  and  $\{(u, y) \in \mathcal{T}_c^{\text{WAK}}(\gamma_c)^c\}$  lie under the *same* probability and so can be bounded together (as a vector) in second-order coding analysis. The sum of the information spectrum terms (first two terms) in Verdú's bound in [6, Thm. 1] is the result upon invoking the union bound on the first term in (52). We illustrate the differences in the resulting second-order coding rates numerically in Section VII. The bound in (52) is rather unwieldy. We can simplify it without losing too much. Indeed, using the definition of  $\mathcal{T}_b^{\text{WAK}}(\gamma_b)$ , we observe that the second



term in (52) can be bounded as

$$\frac{1}{|\mathcal{M}|} \sum_{(u, \tilde{x}) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)} P_U(u) \quad (53)$$

$$= \frac{1}{|\mathcal{M}|} \sum_{(u, \tilde{x}) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)} P_U(u) \frac{P_{X|U}(\tilde{x}|u)}{P_{X|U}(\tilde{x}|u)} \quad (54)$$

$$\leq \frac{1}{|\mathcal{M}|} \sum_{(u, \tilde{x}) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)} P_U(u) P_{X|U}(\tilde{x}|u) 2^{\gamma_b} \quad (55)$$

$$\leq \frac{2^{\gamma_b}}{|\mathcal{M}|}. \quad (56)$$

Together with (49), we have the following simplified CS-type bound, which resembles a Feinstein-type [48] achievability bound (but average instead of maximum error probability).

**Corollary 6** (Simplified CS-type bound for WAK coding). *For arbitrary  $\gamma_b, \gamma_c \geq 0$ , there exists a WAK code  $\Phi$  with error probability satisfying*

$$P_e(\Phi) \leq P_{UXY} [(u, x) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)^c \cup (u, y) \in \mathcal{T}_c^{\text{WAK}}(\gamma_c)^c] + \frac{2^{\gamma_b}}{|\mathcal{M}|} + \frac{1}{2} \sqrt{\frac{2^{\gamma_c}}{|\mathcal{L}|}}. \quad (57)$$

If  $(X^n, Y^n)$  is drawn from the product distribution  $P_{XY}^n$ , then by designing  $\gamma_b$  and  $\gamma_c$  appropriately, we see that the dominating term in (57) is the first one. The other terms vanish with  $n$ .

By modifying the helper in the proof of Theorem 5, we can show the following theorem.

**Theorem 7** (Modified CS-type bound for WAK coding). *For arbitrary  $\gamma_b, \gamma_c \geq 0$ , and positive integer  $J$ , there exists a WAK code  $\Phi$  with error probability satisfying*

$$P_e(\Phi) \leq P_{UXY} [(u, x) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)^c \cup (u, y) \in \mathcal{T}_c^{\text{WAK}}(\gamma_c)^c] + \frac{1}{|\mathcal{M}|} \sum_{(u, \tilde{x}) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)} P_U(u) + \frac{J}{|\mathcal{M}||\mathcal{L}|} \sum_{(u, \tilde{x}) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)} P_U(u) + \frac{\Delta(\gamma_c, P_{UY})}{2\sqrt{J}}. \quad (58)$$

See Appendix E for the proof of Theorem 7. By letting  $J = |\mathcal{L}|$  in (58), we recover (52) up to an additional residual term, which is unimportant in second-order analysis. A close inspection of the proof reveals that the additional term is due to additional random bin coding at the helper, which is not needed if  $J = |\mathcal{L}|$ .

**Remark 1.** *For the special case such that test channel  $P_{U|Y}$  is noiseless, we can show that there exists a WAK code satisfying*

$$P_e(\Phi) \leq P_{XY} [(x, y) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)^c \cup \mathcal{T}_s^{\text{WAK}}(\gamma_s)^c] + \frac{2^{\gamma_b}}{|\mathcal{M}|} + \frac{2^{\gamma_s}}{|\mathcal{M}||\mathcal{L}|} \quad (59)$$

for any  $\gamma_b, \gamma_s \geq 0$ , where

$$\mathcal{T}_s^{\text{WAK}}(\gamma_s) := \left\{ (x, y) \in \mathcal{X} \times \mathcal{Y} : \log \frac{1}{P_{XY}(x, y)} \leq \gamma_s \right\}. \quad (60)$$

We can prove the bound (59) by using the standard Slepian-Wolf type bin coding for both the main encoder and the helper [23], [44]. As it will turn out later in Section VI-A, this simple bound gives tighter second-order achievability in some cases.

### B. Novel Non-Asymptotic Achievability Bound for the WZ Problem

We now turn our attention to the WZ problem where we derive a similar bound as in Theorem 5. This improves on Verdú's bound in Theorem [6, Thm. 2]. It again uses the same CS idea for the covering part.

Define the three sets for fixed  $(P_{UXY}, g) \in \mathcal{P}_D(P_{XY})$  and non-negative constants  $\gamma_p$  and  $\gamma_c$ :

$$\mathcal{T}_p^{\text{WZ}}(\gamma_p) := \left\{ (u, y) \in \mathcal{U} \times \mathcal{Y} : \log \frac{P_{Y|U}(y|u)}{P_Y(y)} \geq \gamma_p \right\} \quad (61)$$

$$\mathcal{T}_c^{\text{WZ}}(\gamma_c) := \left\{ (u, x) \in \mathcal{U} \times \mathcal{X} : \log \frac{P_{X|U}(x|u)}{P_X(x)} \leq \gamma_c \right\} \quad (62)$$

$$\mathcal{T}_d^{\text{WZ}}(D) := \{(u, x, y) \in \mathcal{U} \times \mathcal{X} \times \mathcal{Y} : d(x, g(u, y)) \leq D\}. \quad (63)$$

These sets have intuitive explanations:  $\mathcal{T}_c^{\text{WZ}}(\gamma_c)^c$  represents the *covering* error that  $U$  is unable to describe  $X$  to the desired level indicated by  $\gamma_c$ ;  $\mathcal{T}_p^{\text{WZ}}(\gamma_p)^c$  represents the *packing* error in which the decoder is unable to decode the correct codeword  $U$  given  $Y$  using a threshold test based on the information density statistic and  $\gamma_p$ ;  $\mathcal{T}_d^{\text{WZ}}(D)^c$  represents the *distortion* error in which the reproduction  $\hat{X}$  not within a distortion of  $D$  of the source  $X$ .

In the following, we allow the reproduction function  $g: \mathcal{U} \times \mathcal{Y} \rightarrow \hat{\mathcal{X}}$  to be stochastic; i.e., we consider a reproduction channel  $P_{\hat{X}|UY}: \mathcal{U} \times \mathcal{Y} \rightarrow \hat{\mathcal{X}}$ . When we consider a stochastic function instead of a deterministic one, we will use the set

$$\mathcal{T}_{d, \text{st}}^{\text{WZ}}(D) := \{(x, \hat{x}) \in \mathcal{X} \times \hat{\mathcal{X}} : d(x, \hat{x}) \leq D\} \quad (64)$$

instead of  $\mathcal{T}_d^{\text{WZ}}(D)$ ; see (66) and Remark 2 below.

In this subsection, a pair  $(P_{U|X}, P_{\hat{X}|UY})$  of a test channel  $P_{U|X}: \mathcal{X} \rightarrow \mathcal{U}$  and a reproduction channel  $P_{\hat{X}|UY}: \mathcal{U} \times \mathcal{Y} \rightarrow \hat{\mathcal{X}}$  is fixed. Note that the joint distribution  $P_{UXY\hat{X}}$  of  $U, X, Y, \hat{X}$  is also fixed as

$$P_{UXY\hat{X}}(u, x, y, \hat{x}) = P_{XY}(x, y) P_{U|X}(u|x) P_{\hat{X}|UY}(\hat{x}|u, y). \quad (65)$$

**Theorem 8** (CS-type bound for WZ coding). *For arbitrary constants  $\gamma_p, \gamma_c \geq 0$  and positive integer  $L$ , there exists a WZ code  $\Phi$  with probability of excess distortion satisfying*

$$P_e(\Phi; D) \leq P_{UXY\hat{X}} [(u, x) \in \mathcal{T}_c^{\text{WZ}}(\gamma_c)^c \cup (x, \hat{x}) \in \mathcal{T}_{d, \text{st}}^{\text{WZ}}(D)^c \cup (u, y) \in \mathcal{T}_p^{\text{WZ}}(\gamma_p)^c] + \frac{L}{|\mathcal{M}|} \sum_{(u, y) \in \mathcal{T}_p^{\text{WZ}}(\gamma_p)} P_U(u) P_Y(y) + \frac{\Delta(\gamma_c, P_{UX})}{2\sqrt{L}}. \quad (66)$$

where  $\Delta(\gamma_c, P_{UX})$  is defined in (48).

**Remark 2.** If  $P_{\hat{X}|UY}$  is deterministic and represented by  $g: \mathcal{U} \times \mathcal{Y} \rightarrow \hat{\mathcal{X}}$  then the event  $\{(x, \hat{x}) \in \mathcal{T}_{d, \text{st}}^{\text{WZ}}(D)^c\}$  can be replaced by  $\{(u, x, y) \in \mathcal{T}_d^{\text{WZ}}(D)^c\}$ . In fact, by an application of the functional representation lemma [1, Appendix A], the assumption that the reproduction channel  $P_{\hat{X}|UY}$  is deterministic can be made without any loss of generality.

The proof of Theorem 10 is provided in Appendix F. As with Theorem 5, the main novelty of our bound lies in the fact that the three error events lie under the same probability, making it amendable to treat all three error events *jointly*. The residual terms in (66) (namely, the second, third and fourth terms) are relatively small with a proper choice of constants  $\gamma_p, \gamma_c$  and  $L \in \mathbb{N}$  as we shall see in the sequel. We can again relax the somewhat cumbersome second and third terms in (66) by noting the definition of  $\mathcal{T}_p^{\text{WZ}}(\gamma_p)$  and by going through the same steps to upper bound  $\Delta$ ; cf. (49). We thus obtain:

**Corollary 9** (Simplified CS-type bound for WZ coding). *For arbitrary constants  $\gamma_p, \gamma_c \geq 0$  and positive integer  $L$ , there exists a WZ code  $\Phi$  with probability of excess distortion satisfying*

$$P_e(\Phi; D) \leq P_{USXY}[(u, y) \in \mathcal{T}_p^{\text{WZ}}(\gamma_p)^c \cup (u, x) \in \mathcal{T}_c^{\text{WZ}}(\gamma_c)^c \cup (x, \hat{x}) \in \mathcal{T}_{d, \text{st}}^{\text{WZ}}(D)^c] + \frac{L}{2^{\gamma_p} |\mathcal{M}|} + \frac{1}{2} \sqrt{\frac{2^{\gamma_c}}{L}}. \quad (67)$$

To obtain achievable second-order coding rates for the WZ problem, we evaluate the bound in (67) for appropriate choices of  $\gamma_p, \gamma_c \geq 0$  and  $L \in \mathbb{N}$  in Section VI-B. Since the lossy source coding problem is a special case of WZ coding, we use a specialization of the bound in (67) to derive an achievable dispersion (or second-order coding rate) of lossy source coding [25], [26], which turns out to be tight.

### C. Novel Non-Asymptotic Achievability Bound for the GP Problem

This section presents with a novel non-asymptotic achievability bound for the GP problem, which is the dual of the WZ problem [46]. Our bound improves on Verdú's non-asymptotic bound for GP coding [6, Thm. 3] and uses the same Channel-Simulation idea for the covering part.

To state the bound, we define the sets

$$\mathcal{T}_p^{\text{GP}}(\gamma_p) := \left\{ (u, y) \in \mathcal{U} \times \mathcal{Y} : \log \frac{P_{Y|U}(y|u)}{P_Y(y)} \geq \gamma_p \right\} \quad (68)$$

$$\mathcal{T}_c^{\text{GP}}(\gamma_c) := \left\{ (u, s) \in \mathcal{U} \times \mathcal{S} : \log \frac{P_{S|U}(s|u)}{P_S(s)} \leq \gamma_c \right\} \quad (69)$$

These are analogous to the typical sets used extensively in network information theory [1] but they only involve the information densities. The first set in (68) represents *packing* event while the second in (69) represents *covering* event. Also recall the definition of the set  $\mathcal{T}_g^{\text{GP}}(\Gamma)$  in (27) which represents satisfaction of the cost constraints.

In the following, the distribution  $P_{USXY} \in \mathcal{P}(\mathcal{U} \times \mathcal{S} \times \mathcal{X} \times \mathcal{Y})$  satisfying (i) the  $\mathcal{S}$ -marginal of  $P_{USXY}$  is  $P_S$ , (ii)  $P_{Y|XS} = W$  and (iii)  $U - (X, S) - Y$  forms a Markov chain

is fixed. Note the encoding function  $P_{X|US}$  is allowed to be stochastic but just as in Remark 2, there is no loss in assuming  $P_{X|US}$  is deterministic by the functional representation lemma. We prefer to use  $P_{X|US}$  for convenience.

**Theorem 10** (CS-type bound for GP coding). *For arbitrary constants  $\gamma_p, \gamma_c \geq 0$  and positive integer  $L$ , there exists a GP code  $\Phi$  with average error probability satisfying*

$$P_e(\Phi; \Gamma) \leq P_{USXY}[(u, y) \in \mathcal{T}_p^{\text{GP}}(\gamma_p)^c \cup (u, s) \in \mathcal{T}_c^{\text{GP}}(\gamma_c)^c \cup x \in \mathcal{T}_g^{\text{GP}}(\Gamma)^c] + L |\mathcal{M}| \sum_{(u, y) \in \mathcal{T}_p^{\text{GP}}(\gamma_p)} P_U(u) P_Y(y) + \frac{\Delta(\gamma_c, P_{US})}{2\sqrt{L}} \quad (70)$$

where  $\Delta(\gamma_c, P_{US})$  is defined in (48).

Because the technique to prove Theorem 10 is similar to that for Theorems 5 and 8, we only sketch the code construction in Appendix G. In the second-order asymptotics sense, Theorem 10 improves on [6, Thm. 3] because the error events are under the *same* error probability. Notice that unlike the existing asymptotic and non-asymptotic results for GP coding [6], [10], [49], the channel input  $x$  satisfies the cost constraint (25) or its almost sure equivalent (cf. Proposition 1). Direct application of (49) to bound  $\Delta(\gamma_c, P_{US})$  and the definition of  $\mathcal{T}_p^{\text{GP}}(\gamma_p)$  in (68) yields the following:

**Corollary 11** (Simplified CS-type bound for GP coding). *For arbitrary constants  $\gamma_p, \gamma_c \geq 0$  and positive integer  $L$ , there exists a GP code  $\Phi$  with average error probability satisfying*

$$P_e(\Phi; \Gamma) \leq P_{USXY}[(u, y) \in \mathcal{T}_p^{\text{GP}}(\gamma_p)^c \cup (u, s) \in \mathcal{T}_c^{\text{GP}}(\gamma_c)^c \cup x \in \mathcal{T}_g^{\text{GP}}(\Gamma)^c] + \frac{L |\mathcal{M}|}{2^{\gamma_p}} + \frac{1}{2} \sqrt{\frac{2^{\gamma_c}}{L}}. \quad (71)$$

To obtain achievable second-order coding rates for the GP problem, we evaluate the bound in (71) for appropriate choices of  $\gamma_p, \gamma_c$  and  $L \in \mathbb{N}$  in Section VI-C.

## V. GENERAL FORMULAS

In this section, we use the simplified CS-type bounds in Corollaries 6, 9 and 11 to derive achievable general formulas for the optimal rate region of the WAK problem, the rate-distortion function of the WZ problem and the capacity of the GP problem. This allows us to recover known results in [8]–[10]. By *general formula*, we mean that we consider sequences of these problems and do not place any underlying structure such as stationarity, memorylessness and ergodicity on the source and channel [7], [20]. To state our results, let us first recall the following probabilistic limit operations. Their properties are similar to the limit superior and limit inferior for numerical sequences in mathematical analysis and are summarized in [7].

**Definition 7.** Let  $\mathbf{U} := \{U_n\}_{n=1}^\infty$  be a sequence of real-valued random variables. The limit superior in probability of  $\mathbf{U}$  is

defined as

$$\mathbf{p}\text{-lim sup}_{n \rightarrow \infty} U_n := \inf \left\{ \alpha \in \mathbb{R} : \lim_{n \rightarrow \infty} \Pr(U_n > \alpha) = 0 \right\}. \quad (72)$$

The limit inferior in probability of  $\mathbf{U}$  is defined as

$$\mathbf{p}\text{-lim inf}_{n \rightarrow \infty} U_n := -\mathbf{p}\text{-lim sup}_{n \rightarrow \infty} (-U_n) \quad (73)$$

We also recall the following definitions from Han [7]. These definitions play a prominent role in the rest of this section.

**Definition 8.** Given a pair of stochastic processes  $(\mathbf{X}, \mathbf{Y}) = \{X^n, Y^n\}_{n=1}^\infty$  with joint distributions  $\{P_{X^n, Y^n}\}_{n=1}^\infty$ , the spectral sup-mutual information rate is defined as

$$\bar{I}(\mathbf{X}; \mathbf{Y}) := \mathbf{p}\text{-lim sup}_{n \rightarrow \infty} \frac{1}{n} \log \frac{P_{Y^n|X^n}(Y^n|X^n)}{P_{Y^n}(Y^n)}. \quad (74)$$

The spectral inf-mutual information rate  $\underline{I}(\mathbf{X}; \mathbf{Y})$  is defined as in (72) with  $\mathbf{p}\text{-lim inf}$  in place of  $\mathbf{p}\text{-lim sup}$ . The spectral sup- and inf-conditional mutual information rates are defined similarly.

The spectral sup-conditional entropy rates is defined as

$$\bar{H}(\mathbf{Y}|\mathbf{X}) := \mathbf{p}\text{-lim sup}_{n \rightarrow \infty} \frac{1}{n} \log \frac{1}{P_{Y^n|X^n}(Y^n|X^n)}. \quad (75)$$

The spectral inf-conditional entropy rates is defined as in (75) with  $\mathbf{p}\text{-lim inf}$  in place of  $\mathbf{p}\text{-lim sup}$ .

#### A. General Formula for the WAK problem

In this section, we consider sequences of the WAK problem indexed by the blocklength  $n$  where the sequence of source distributions  $\{P_{X^n Y^n}\}_{n=1}^\infty$  is *general*, i.e., we do not place any assumptions on the structure of the source such as stationarity, memorylessness and ergodicity. We aim to characterize an inner bound to the optimal rate region defined in (14). We show that our inner bound coincides with that derived by Miyake and Kanaya [8] but is derived based on the upper bound on the error probability provided in our CS-type bound in Corollary 6. The choice of the parameters  $\gamma_b, \gamma_c$  and  $\delta$  plays a crucial role and guides our choice of these parameters for second-order coding analysis in the following section.

Let  $\mathcal{P}(\{P_{X^n Y^n}\}_{n=1}^\infty)$  be the set of all sequences of distributions  $\{P_{U^n X^n Y^n}\}_{n=1}^\infty$  such that for every  $n \geq 1$ ,  $U^n - Y^n - X^n$  forms a Markov chain and the  $(\mathcal{X}^n \times \mathcal{Y}^n)$ -marginal of  $P_{U^n X^n Y^n}$  is  $P_{X^n Y^n}$ . Define the set

$$\begin{aligned} \hat{\mathcal{R}}_{\text{WAK}}^* := & \bigcup_{\{P_{U^n X^n Y^n}\}_{n=1}^\infty \in \mathcal{P}(\{P_{X^n Y^n}\}_{n=1}^\infty)} \\ & \{(R_1, R_2) \in \mathbb{R}_+^2 : R_1 \geq \bar{H}(\mathbf{X}|\mathbf{U}), R_2 \geq \bar{I}(\mathbf{U}; \mathbf{Y})\} \end{aligned} \quad (76)$$

**Theorem 12** (Inner Bound to the Optimal Rate Region for WAK [8]). We have

$$\hat{\mathcal{R}}_{\text{WAK}}^* \subset \mathcal{R}_{\text{WAK}}. \quad (77)$$

We remark that by using techniques from [32], Miyake and Kanaya [8] showed that (77) is in fact an equality, i.e.,  $\hat{\mathcal{R}}_{\text{WAK}}^*$  is also an outer bound to  $\mathcal{R}_{\text{WAK}}$ . In addition, when the source distributions  $\{P_{X^n Y^n}\}_{n=1}^\infty$  are stationary

and memoryless (and the alphabets  $\mathcal{X}$  and  $\mathcal{Y}$  are discrete and finite),  $\hat{\mathcal{R}}_{\text{WAK}}^*$  reduces to the single-letter region  $\mathcal{R}_{\text{WAK}}^*$  defined in (40). This follows easily from the law of large numbers. The proof of Theorem 12 follows directly from the finite blocklength bound in Corollary 6. In fact, the weaker bounds in [30] and [6] suffice for this purpose.

*Proof:* Consider (57) and let us fix a process  $\{P_{U^n X^n Y^n}\}_{n=1}^\infty \in \mathcal{P}(\{P_{X^n Y^n}\}_{n=1}^\infty)$  and a constant  $\eta > 0$ . Set

$$\frac{1}{n} \log |\mathcal{M}| := \bar{H}(\mathbf{X}|\mathbf{U}) + 2\eta \quad (78)$$

$$\frac{1}{n} \log |\mathcal{L}| := \bar{I}(\mathbf{U}; \mathbf{Y}) + 2\eta \quad (79)$$

$$\gamma_b := n(\bar{H}(\mathbf{X}|\mathbf{U}) + \eta) \quad (80)$$

$$\gamma_c := n(\bar{I}(\mathbf{U}; \mathbf{Y}) + \eta) \quad (81)$$

Then for blocklength  $n$ , the probability on the RHS of (57) can be written as

$$\begin{aligned} & P_{U^n X^n Y^n} \left[ \left\{ \frac{1}{n} \log \frac{1}{P_{X^n|U^n}(X^n|U^n)} \geq \bar{H}(\mathbf{X}|\mathbf{U}) + \eta \right\} \right. \\ & \left. \bigcup \left\{ \frac{1}{n} \log \frac{P_{Y^n|U^n}(Y^n|U^n)}{P_{Y^n}(Y^n)} \geq \bar{I}(\mathbf{U}; \mathbf{Y}) + \eta \right\} \right] \end{aligned} \quad (82)$$

By the definition of the spectral sup-entropy rate and the spectral sup-mutual information rate, the probabilities of both events in (82) tend to zero. Further,

$$\frac{2^{\gamma_b}}{|\mathcal{M}|} = 2^{-n\eta} \rightarrow 0, \quad \text{and} \quad \frac{1}{2} \sqrt{\frac{2^{\gamma_c}}{|\mathcal{L}|}} = \frac{1}{2} \cdot 2^{-n\eta/2} \rightarrow 0. \quad (83)$$

Hence,  $P_e(\Phi_n) \rightarrow 0$ . Since  $\eta > 0$  is arbitrary, from (78) and (79) we deduce that any pair of rates  $(R_1, R_2)$  satisfying  $R_1 > \bar{H}(\mathbf{X}|\mathbf{U})$  and  $R_2 > \bar{I}(\mathbf{U}; \mathbf{Y})$  is achievable. ■

#### B. General Formula for the WZ problem

In a similar way, we can recover the general formula for WZ coding derived by Iwata and Muramatsu [9]. Note however, that we directly work with the probability of excess distortion, which is related to but different from the maximum-distortion criterion employed in [9]. Once again, we assume that the source is  $\{P_{X^n Y^n}\}_{n=1}^\infty$  is *general* in the sense explained in Section V-A.

Let  $\mathcal{P}_D(\{P_{X^n Y^n}\}_{n=1}^\infty)$  be the set of all sequences of distributions  $\{P_{U^n X^n Y^n}\}_{n=1}^\infty$  and reproduction functions  $\{g_n : U^n \times Y^n \rightarrow \hat{\mathcal{X}}^n\}$  such that for every  $n \geq 1$ ,  $U^n - X^n - Y^n$  forms a Markov chain, the  $(\mathcal{X}^n \times \mathcal{Y}^n)$ -marginal of  $P_{U^n X^n Y^n}$  is  $P_{X^n Y^n}$  and

$$\mathbf{p}\text{-lim sup}_{n \rightarrow \infty} d_n(X^n, g_n(U^n, Y^n)) \leq D \quad (84)$$

Define the rate-distortion function

$$\hat{R}_{\text{WZ}}^*(D) := \inf \{ \bar{I}(\mathbf{U}; \mathbf{X}) - \underline{I}(\mathbf{U}; \mathbf{Y}) \} \quad (85)$$

where the infimum is over all  $\{P_{U^n X^n Y^n}, g_n\}_{n=1}^\infty \in \mathcal{P}_D(\{P_{X^n Y^n}\}_{n=1}^\infty)$ .

**Theorem 13** (Upper Bound to the Rate-Distortion Function for WZ [9]). *We have*

$$R_{WZ}(D) \leq \hat{R}_{WZ}^*(D). \quad (86)$$

Iwata and Muramatsu [9] showed in fact that (86) is an equality by proving a converse along the lines of [32]. It can be shown that the general rate-distortion function defined in (85) reduces to the one derived by Wyner and Ziv [4] in the case where the alphabets are finite and the source is stationary and memoryless. Also Iwata and Muramatsu [9] showed that deterministic reproduction functions  $g_n : \mathcal{U}^n \times \mathcal{Y}^n \rightarrow \mathcal{X}^n$  suffice and we do not need the more general stochastic reproduction functions  $P_{\hat{X}^n|U^n Y^n}$ .

*Proof:* Let  $\eta > 0$ . We start from the bound on the probability of excess distortion in (67), where we first consider  $D + \eta$  instead of  $D$ . Let us fix the sequence of distribution and the sequence of functions  $\{(P_{U^n X^n Y^n}, g_n)\}_{n=1}^\infty \in \mathcal{P}_D(\{P_{X^n Y^n}\}_{n=1}^\infty)$ . Set

$$\frac{1}{n} \log |\mathcal{M}| := \bar{I}(\mathbf{U}; \mathbf{X}) - \underline{I}(\mathbf{U}; \mathbf{Y}) + 4\eta \quad (87)$$

$$\frac{1}{n} \log L := \bar{I}(\mathbf{U}; \mathbf{X}) + 2\eta \quad (88)$$

$$\gamma_p := n(\underline{I}(\mathbf{U}; \mathbf{Y}) - \eta) \quad (89)$$

$$\gamma_c := n(\bar{I}(\mathbf{U}; \mathbf{X}) + \eta). \quad (90)$$

Then, the probability in (67) for blocklength  $n$  can be written as

$$P_{U^n X^n Y^n} \left[ \left\{ \frac{1}{n} \log \frac{P_{Y^n|U^n}(Y^n|U^n)}{P_{Y^n}(Y^n)} \leq \underline{I}(\mathbf{U}; \mathbf{Y}) - \eta \right\} \right. \\ \left. \bigcup \left\{ \frac{1}{n} \log \frac{P_{X^n|U^n}(X^n|U^n)}{P_{X^n}(X^n)} \geq \bar{I}(\mathbf{U}; \mathbf{X}) + \eta \right\} \right. \\ \left. \bigcup \left\{ d_n(X^n, g_n(U^n, Y^n)) \geq D + \eta \right\} \right] \quad (91)$$

By the definition of the spectral sup- and inf-mutual information rates and the distortion condition in (84), we observe that the probability in (91) tends to zero as  $n$  grows. By a similar calculation as in (83), the other terms in (67) also tend to zero. Hence, the probability of excess distortion  $P_e(\Phi_n; D + \eta) \rightarrow 0$  as  $n$  grows. This holds for every  $\eta > 0$ . By (87), the any rate below  $\bar{I}(\mathbf{U}; \mathbf{X}) - \underline{I}(\mathbf{U}; \mathbf{Y}) + 4\eta$  is achievable. In order to complete the proof, we choose a positive sequence satisfying  $\eta_1 > \eta_2 > \dots > 0$  and  $\eta_k \rightarrow 0$  as  $k \rightarrow \infty$ . Then, by using the diagonal line argument [7, Thm. 1.8.2], we complete the proof of (86). ■

### C. General Formula for the GP problem

We conclude this section by showing that the non-asymptotic bound on the average probability of error derived in Corollary 11 can be adapted to recover the general formula for the GP problem derived in Tan [10]. Here, both the state distribution  $\{P_{S^n} \in \mathcal{P}(S^n)\}_{n=1}^\infty$  and the channel  $\{W^n : \mathcal{X}^n \times \mathcal{S}^n \rightarrow \mathcal{Y}^n\}_{n=1}^\infty$  are general. In particular, the only requirement on the stochastic mapping  $W^n$  is that for every  $(x^n, s^n) \in \mathcal{X}^n \times \mathcal{S}^n$ ,

$$\sum_{y^n \in \mathcal{Y}^n} W^n(y^n|x^n, s^n) = 1. \quad (92)$$

Let  $\mathcal{P}_\Gamma(\{W^n, P_{S^n}\}_{n=1}^\infty)$  be the family of joint distributions  $P_{U^n S^n X^n Y^n}$  such that for every  $n \geq 1$ ,  $U^n - (X^n, S^n) - Y^n$  forms a Markov chain, the  $S^n$ -marginal of  $P_{U^n S^n X^n Y^n}$  is  $P_{S^n}$ , the channel law  $P_{Y^n|X^n, S^n} = W^n$  and

$$\mathbf{p}\text{-}\limsup_{n \rightarrow \infty} g_n(X^n) \leq \Gamma \quad (93)$$

Define the quantity

$$\hat{C}_{GP}^*(\Gamma) := \sup \{ \underline{I}(\mathbf{U}; \mathbf{Y}) - \bar{I}(\mathbf{U}; \mathbf{S}) \} \quad (94)$$

where the supremum is over all joint distributions  $\{P_{U^n S^n X^n Y^n}\}_{n=1}^\infty \in \mathcal{P}_\Gamma(\{W^n, P_{S^n}\}_{n=1}^\infty)$ .

**Theorem 14** (Lower Bound to the GP capacity [10]). *We have*

$$C_{GP}(\Gamma) \geq \hat{C}_{GP}^*(\Gamma). \quad (95)$$

Tan [10] also showed that the inequality in (95) is, in fact, tight. However, unlike in the general WZ scenario, the encoding function  $P_{X^n|U^n S^n}$  cannot be assumed to be deterministic in general. When the channel and state are discrete, stationary and memoryless, Tan [10] showed that the general formula in (94) reduces to the conventional one derived by Gel'fand-Pinsker [5] in (46). The proof of Theorem 14 parallels that for Theorem 13 and thus, we omit it.

## VI. ACHIEVABLE SECOND-ORDER CODING RATES

In this section, we demonstrate achievable second-order coding rates [11], [12], [22], [39], [40] for the three side-information problems of interest. Essentially, we are interested in characterizing the  $(n, \varepsilon)$ -optimal rate region for the WAK problem, the  $(n, \varepsilon)$ -Wyner-Ziv rate-distortion function and the  $(n, \varepsilon)$ -capacity of GP problem up to the second-order term. We do this by applying the multidimensional Berry-Esséen theorem [21], [50] to the finite blocklength CS-type bounds in Corollaries 6, 9 and 11. Throughout, we will not concern ourselves with optimizing the third-order terms.

The following important definition will be used throughout this section.

**Definition 9.** Let  $k$  be a positive integer. Let  $\mathbf{V} \in \mathbb{R}^{k \times k}$  be a positive-semidefinite matrix that is not the all-zeros matrix but is allowed to be rank-deficient. Let the Gaussian random vector  $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{V})$ . Define the set

$$\mathcal{S}(\mathbf{V}, \varepsilon) := \{\mathbf{z} \in \mathbb{R}^k : \Pr(\mathbf{Z} \leq \mathbf{z}) \geq 1 - \varepsilon\}. \quad (96)$$

This set was introduced in [23] and is, roughly speaking, the multidimensional analogue of the  $Q^{-1}$  function. Indeed, for  $k = 1$  and any standard deviation  $\sigma > 0$ ,

$$\mathcal{S}(\sigma^2, \varepsilon) = [\sigma Q^{-1}(\varepsilon), \infty). \quad (97)$$

Also,  $\mathbf{1}_k$  and  $\mathbf{0}_{k \times k}$  denote the length- $k$  all-ones column vector and the  $k \times k$  all-zeros matrix respectively.

### A. Achievable Second-Order Coding Rates for the WAK problem

In this section, we derive an inner bound to  $\mathcal{R}_{\text{WAK}}(n, \varepsilon)$  in (12) by the use of Gaussian approximations. Instead of simply applying the Berry-Esséen theorem to the information spectrum term within the simplified CS-type bound in (57), we enlarge our inner bound by using a “time-sharing” variable  $T$ , which is independent of  $(X, Y)$ . This technique was also used for the multiple access channel (MAC) by Huang and Moulin [42]. Note that in the finite blocklength setting, the region  $\mathcal{R}_{\text{WAK}}(n, \varepsilon)$  does not have to be convex unlike in the asymptotic case; cf. (40). For fixed finite sets  $\mathcal{U}$  and  $\mathcal{T}$ , let  $\tilde{\mathcal{P}}(P_{XY})$  be the set of all  $P_{UTXY} \in \mathcal{P}(\mathcal{U} \times \mathcal{T} \times \mathcal{X} \times \mathcal{Y})$  such that the  $\mathcal{X} \times \mathcal{Y}$ -marginal of  $P_{UTXY}$  is  $P_{XY}$ ,  $U - (Y, T) - X$  forms a Markov chain and  $T$  is independent of  $(X, Y)$ .

**Definition 10.** The entropy-information density vector for the WAK problem for  $P_{UTXY} \in \tilde{\mathcal{P}}(P_{XY})$  is defined as

$$\mathbf{j}(U, X, Y|T) := \begin{bmatrix} \log \frac{P_{X|UT}(X|U, T)}{P_Y(Y)} \\ \log \frac{P_{Y|UT}(Y|U, T)}{P_Y(Y)} \end{bmatrix}. \quad (98)$$

Note that the mean of the entropy-information density vector in (98) is the vector of the entropy and mutual information, i.e.,

$$\mathbf{J}(P_{UTXY}) := \mathbb{E}[\mathbf{j}(U, X, Y|T)] = \begin{bmatrix} H(X|U, T) \\ I(U; Y|T) \end{bmatrix}. \quad (99)$$

The mutual information  $I(U; Y|T) = I(U, T; Y)$  because  $T$  and  $Y$  are independent.

**Definition 11.** The entropy-information dispersion matrix for the WAK problem for a fixed  $P_{UTXY} \in \tilde{\mathcal{P}}(P_{XY})$  is defined as

$$\mathbf{V}(P_{UTXY}) := \mathbb{E}_T[\text{Cov}(\mathbf{j}(U, X, Y|T))] \quad (100)$$

$$= \sum_{t \in \mathcal{T}} P_T(t) \text{Cov}(\mathbf{j}(U, X, Y|t)). \quad (101)$$

We abbreviate the deterministic quantities  $\mathbf{J}(P_{UTXY}) \in \mathbb{R}_+^2$  and  $\mathbf{V}(P_{UTXY}) \succeq 0$  as  $\mathbf{J}$  and  $\mathbf{V}$  respectively when the distribution  $P_{UTXY} \in \tilde{\mathcal{P}}(P_{XY})$  is obvious from the context.

**Definition 12.** If  $\mathbf{V}(P_{UTXY}) \neq \mathbf{0}_{2 \times 2}$ , define  $\mathcal{R}_{\text{in}}(n, \varepsilon; P_{UTXY})$  to be the set of rate pairs  $(R_1, R_2)$  such that  $\mathbf{R} := [R_1, R_2]^T$  satisfies

$$\mathbf{R} \in \mathbf{J} + \frac{\mathcal{S}(\mathbf{V}, \varepsilon)}{\sqrt{n}} + \frac{2 \log n}{n} \mathbf{1}_2. \quad (102)$$

If  $\mathbf{V}(P_{UTXY}) = \mathbf{0}_{2 \times 2}$ , define  $\mathcal{R}_{\text{in}}(n, \varepsilon; P_{UTXY})$  to be the set of rate pairs  $(R_1, R_2)$  such that

$$\mathbf{R} \in \mathbf{J} + \frac{2 \log n}{n} \mathbf{1}_2. \quad (103)$$

From the simplified CS-type bound for the WAK problem in Corollary 6, we can derive the following:

**Theorem 15** (Inner Bound to  $(n, \varepsilon)$ -Optimal Rate Region). *For every  $0 < \varepsilon < 1$  and all  $n$  sufficiently large, the  $(n, \varepsilon)$ -optimal rate region  $\mathcal{R}_{\text{WAK}}(n, \varepsilon)$  satisfies*

$$\bigcup_{P_{UTXY} \in \tilde{\mathcal{P}}(P_{XY})} \mathcal{R}_{\text{in}}(n, \varepsilon; P_{UTXY}) \subset \mathcal{R}_{\text{WAK}}(n, \varepsilon). \quad (104)$$

Furthermore, the union over  $P_{UTXY}$  can be restricted to those distributions for which the supports  $\mathcal{U}$  and  $\mathcal{T}$  of auxiliary random variables  $U$  and  $T$  satisfy that  $|\mathcal{U}| \leq |\mathcal{Y}| + 4$  and  $|\mathcal{T}| \leq 5$  respectively.

From the modified CS-type bound for the WAK problem in Theorem 7, we can derive the following:

**Theorem 16** (Modified Inner Bound to  $(n, \varepsilon)$ -Optimal Rate Region). *For every  $0 < \varepsilon < 1$  and all  $n$  sufficiently large, the  $(n, \varepsilon)$ -optimal rate region  $\mathcal{R}_{\text{WAK}}(n, \varepsilon)$  satisfies*

$$\bigcup_{P_{UTXY} \in \tilde{\mathcal{P}}(P_{XY})} \mathcal{R}'_{\text{in}}(n, \varepsilon; P_{UTXY}) \subset \mathcal{R}_{\text{WAK}}(n, \varepsilon), \quad (105)$$

where  $\mathcal{R}'_{\text{in}}(n, \varepsilon; P_{UTXY})$  is the set defined by replacing (102) with

$$\mathbf{R} \in \bigcup_{\rho \geq 0} \left\{ \mathbf{J} + \frac{\mathcal{S}(\mathbf{V}, \varepsilon) + [\rho, -\rho]^T}{\sqrt{n}} + \frac{2 \log n}{n} \mathbf{1}_2 \right\}. \quad (106)$$

**Remark 3.** We can also restrict the cardinalities  $|\mathcal{U}|$  and  $|\mathcal{T}|$  of auxiliary random variables in Theorem 16 in the same way as in Theorem 15. The bound in Theorem 16 is at least as tight as that in Theorem 15, and the former is strictly tighter than the latter for a fixed test channel. However, it is not clear whether the improvement is strict or not when we take the union over the test channels.

By setting  $T = Y = U = \emptyset$  and  $R_2 = 0$  in Theorem 16,<sup>5</sup> we obtain a result first discovered by Strassen [39].

**Corollary 17** (Achievable Second-Order Coding Rate for Lossless Source Coding). *Define the second-order coding rate for lossless source coding to be*

$$\sigma(P_X, \varepsilon) := \limsup_{n \rightarrow \infty} \sqrt{n}(R_X(n, \varepsilon) - H(X)) \quad (107)$$

where  $R_X(n, \varepsilon)$  is the minimal rate of almost-lossless compression of source  $P_X$  at blocklength  $n$  with error probability not exceeding  $\varepsilon$ . Then,

$$\sigma(P_X, \varepsilon) \leq \sqrt{\text{Var}(\log P_X(X))} Q^{-1}(\varepsilon). \quad (108)$$

It is well-known that the result in Corollary 17 is tight, i.e.,  $\sqrt{\text{Var}(\log P_X(X))} Q^{-1}(\varepsilon)$  is indeed the second-order coding rate for lossless source coding [11], [39], [40].

We refer to the reader to Appendix I for the proof of Theorem 15 (Appendix J for the proof of Theorem 16). The proof is based on the CS-type bound in (57) and the non-i.i.d. version of the multidimensional Berry-Esséen theorem by Götze [21]. The proof of the cardinality bounds is provided in Appendix M. The interpretation of this result is clear: From (102) which is the non-degenerate case, we see that the second-order coding rate region for a fixed  $P_{UTXY}$  is represented by the set  $\mathcal{S}(\mathbf{V}(P_{UTXY}), \varepsilon)/\sqrt{n}$ . Thus, the  $(n, \varepsilon)$ -optimal rate region converges to the asymptotic WAK region at a rate of  $O(1/\sqrt{n})$  which can be predicted by the central limit theorem. More importantly, because our finite blocklength bound in (57) treats both the covering and binning error events jointly,

<sup>5</sup>In fact, to be precise, we cannot derive Corollary 17 from Theorem 15 because there is the residual term  $\frac{2 \log n}{n}$  and we cannot set  $R_2 = 0$ . However, we can use Corollary 6 with  $U = \emptyset$  to obtain Corollary 17 easily.

this results in the coupling of the second-order rates through the set  $\mathcal{S}(\mathbf{V}(P_{UTXY}), \varepsilon)$  and hence, the dispersion matrix  $\mathbf{V}(P_{UTXY})$ . This shows that the correlation between the entropy and information densities matters in the determination of the second-order coding rate.

More specifically, Theorems 15 and 16 are proved by taking  $P_{U^n|Y^n}(u^n|y^n)$  to be equal to  $P_{U|TY}^n(u^n|t^n, y^n)$  for some fixed (time-sharing) sequence  $t^n \in \mathcal{T}^n$  and some joint distribution  $P_{UTXY} \in \mathcal{P}(P_{XY})$ . If  $\mathcal{T} = \emptyset$ , this is essentially using i.i.d. codes. Theorems 15 and 16 also show that  $|\mathcal{T}|$  can be upper bounded by 5. An alternative to this proof strategy is to use conditionally constant composition codes as was done in Kelly-Wagner [51] to prove their error exponent result. The advantage of this strategy is that it may yield better dispersion matrices because the unconditional dispersion matrix always dominates the conditional dispersion matrix [22, Lemma 62] (in the partial order induced by semi-definiteness). For using conditionally constant composition codes, we fix a conditional type  $V_{Q_Y} \in \mathcal{V}_n(\mathcal{U}; Q_Y)$  for every marginal type  $Q_Y \in \mathcal{P}_n(\mathcal{Y})$ . Then, codewords are generated uniformly at random from  $\mathcal{T}_{V_{Q_Y}}(y^n)$  if  $y^n \in \mathcal{T}_{Q_Y}$ . However, it does not appear that this strategy yields improved second-order coding rates compared to using i.i.d. codes as given in Theorems 15 and 16.

We emphasize here that the restriction of the sizes of the alphabets  $\mathcal{U}$  and  $\mathcal{T}$  only allows us to only preserve the *second-order region* defined by the vector  $\mathbf{J}(P_{UTXY})$  and the matrix  $\mathbf{V}(P_{UTXY})$  over all  $P_{UTXY} \in \mathcal{P}(P_{XY})$ . An optimized third-order term in (102) might be dependent on higher-order statistics of the entropy-information density vector  $\mathbf{j}(U, X, Y|T)$  and the quantities that define this third-order term are *not preserved* by the bounds  $|\mathcal{U}| \leq |\mathcal{Y}| + 4$  and  $|\mathcal{T}| \leq 5$ . This remark is also applicable to the second-order rate regions for WZ and GP in Subsections VI-B and VI-C. However, we note that for lossless source coding [39] or channel coding [22], [52], under some regularity conditions, the third-order term is neither dependent on higher-order statistics nor on the alphabet sizes.

To compare our Theorems 15 and 16 to that of Verdú [6], for a fixed  $P_{UXY} \in \mathcal{P}(P_{XY})$ , define  $\mathcal{R}_{\text{in}}^V(n, \varepsilon; P_{UXY})$  to be the set of rate pairs that satisfy

$$R_1 \geq H(X|U) + \sqrt{\frac{V_H(X|U)}{n}} Q^{-1}(\lambda\varepsilon) + \frac{2 \log n}{n} \quad (109)$$

$$R_2 \geq I(U; Y) + \sqrt{\frac{V_I(U; Y)}{n}} Q^{-1}((1 - \lambda)\varepsilon) + \frac{2 \log n}{n} \quad (110)$$

for some  $\lambda \in [0, 1]$  where the marginal entropy and information dispersions are defined as

$$V_H(X|U) := \text{Var} \left( \log \frac{1}{P_{X|U}(X|U)} \right) \quad (111)$$

$$V_I(U; Y) := \text{Var} \left( \log \frac{P_{Y|U}(Y|U)}{P_Y(Y)} \right) \quad (112)$$

respectively. Note that if  $T = \emptyset$ , then  $V_H(X|U)$  and  $V_I(U; Y)$  are the diagonal elements of the matrix  $\mathbf{V}(P_{UTXY})$  in (100). It can easily be seen that Verdú's bound on the error probability

of the WAK problem (8) yields the following inner bound on  $\mathcal{R}_{\text{WAK}}(n, \varepsilon)$ .

$$\bigcup_{P_{UXY} \in \mathcal{P}(P_{XY})} \mathcal{R}_{\text{in}}^V(n, \varepsilon; P_{UXY}) \subset \mathcal{R}_{\text{WAK}}(n, \varepsilon). \quad (113)$$

This “splitting” technique of  $\varepsilon$  into  $\lambda\varepsilon$  and  $(1 - \lambda)\varepsilon$  in (109) and (110) was used by MolavianJazi and Laneman [43] in their work on finite blocklength analysis for the MAC. In Section VII, we numerically compare the inner bounds for the WAK problem provided in (104), (105) and (113).

**Remark 4.** From the non-asymptotic bound in Remark 1, we can also show that

$$\hat{\mathcal{R}}_{\text{in}}(n, \varepsilon) \subset \mathcal{R}_{\text{WAK}}(n, \varepsilon), \quad (114)$$

where  $\hat{\mathcal{R}}_{\text{in}}(n, \varepsilon)$  is the set of rate pairs  $(R_1, R_2)$  such that

$$\begin{bmatrix} R_1 \\ R_1 + R_2 \end{bmatrix} \in \begin{bmatrix} H(X|Y) \\ H(X, Y) \end{bmatrix} + \frac{\mathcal{S}(\hat{\mathbf{V}}, \varepsilon)}{\sqrt{n}} + \frac{2 \log n}{n} \mathbf{1}_2 \quad (115)$$

for the covariance matrix

$$\hat{\mathbf{V}} = \text{Cov} \left( \begin{bmatrix} -\log P_{X|Y}(X|Y) \\ -\log P_{XY}(X, Y) \end{bmatrix} \right). \quad (116)$$

#### B. Achievable Second-Order Coding Rates for the WZ problem

In this section, we leverage on the simplified CS-type bound in Corollary 9 to derive an achievable second-order coding rate for the WZ problem. We do so by first finding an inner bound to the  $(n, \varepsilon)$ -Wyner-Ziv rate-distortion region  $\mathcal{R}_{\text{WZ}}(n, \varepsilon)$  defined in (18). Subsequently we find an upper bound to the  $(n, \varepsilon)$ -Wyner-Ziv rate-distortion function  $R_{\text{WZ}}(n, \varepsilon)$  defined in (21). We also show that the (direct part of the) dispersion of lossy source coding found by Ingber-Kochman [25] and Kostina-Verdú [26] can be recovered from the CS-type bound in Corollary 9. This is not unexpected because the lossy source coding (rate-distortion) problem is a special case of the Wyner-Ziv problem where the side-information is absent.

We will again employ the “time-sharing” strategy used in Section VI-A and show that the cardinality of the time-sharing alphabet  $\mathcal{T}$  can be bounded. Note again that in the finite-blocklength setting  $\mathcal{R}_{\text{WZ}}(n, \varepsilon)$  does not have to be convex, unlike in the asymptotic setting. For fixed finite sets  $\mathcal{U}$  and  $\mathcal{T}$ , let  $\tilde{\mathcal{P}}(P_{XY})$  be the collection of all joint distributions  $P_{UTXY} \in \mathcal{P}(\mathcal{U} \times \mathcal{T} \times \mathcal{X} \times \mathcal{Y})$  such that the  $\mathcal{X} \times \mathcal{Y}$ -marginal of  $P_{UTXY}$  is  $P_{XY}$ ,  $U - (X, T) - Y$  forms a Markov chain and  $T$  is independent of  $(X, Y)$ . A pair  $(P_{UTXY}, P_{\hat{X}|UYT})$  of a joint distribution  $P_{UTXY} \in \tilde{\mathcal{P}}(P_{XY})$  and a reproduction channel  $P_{\hat{X}|UYT}: \mathcal{U} \times \mathcal{Y} \times \mathcal{T} \rightarrow \hat{\mathcal{X}}$  defines a joint distribution  $P_{UTXY\hat{X}}$  such that

$$\begin{aligned} P_{UTXY\hat{X}}(u, t, x, y, \hat{x}) \\ = P_{XY}(x, y) P_T(t) P_{U|YT}(u|y, t) P_{\hat{X}|UYT}(\hat{x}|u, y, t). \end{aligned} \quad (117)$$

Further, a pair of  $P_{UTXY} \in \tilde{\mathcal{P}}(P_{XY})$  and  $P_{\hat{X}|UYT}$  induces a random variable

$$d(X, \hat{X}|T) := d(X_T, \hat{X}_T) \quad (118)$$

where  $(X_t, \hat{X}_t)$  for any  $t \in \mathcal{T}$  has distribution  $P_{X\hat{X}|T=t}$ . In other words, for fixed  $t \in \mathcal{T}$ ,

$$\begin{aligned} & \Pr\{d(X, \hat{X}|T=t) = d\} \\ &= \sum_{\substack{x, \hat{x}: \\ d(x, \hat{x})=d}} \sum_{u, y} P_{XY}(x, y) P_{U|YT}(u|y, t) P_{\hat{X}|UYT}(\hat{x}|u, y, t). \end{aligned} \quad (119)$$

**Definition 13.** For a pair  $(P_{UTXY}, P_{\hat{X}|UYT})$  of  $P_{UTXY} \in \tilde{\mathcal{P}}(P_{XY})$  and  $P_{\hat{X}|UYT}$ , the information-density-distortion vector for the WZ problem is defined as

$$\mathbf{j}(U, X, Y, \hat{X}|T) := \begin{bmatrix} -\log \frac{P_{Y|UT}(Y|U, T)}{P_Y(Y)} \\ \log \frac{P_{X|UT}(X|U, T)}{P_X(X)} \\ d(X, \hat{X}|T) \end{bmatrix}. \quad (120)$$

Since  $\mathbb{E}[d(X, \hat{X})] = \sum_t P_T(t) \mathbb{E}_{P_{X\hat{X}|T}}[d(X_T, \hat{X}_T)|T = t]$ , the expectation of information-density-distortion vector is given by

$$\mathbf{J}(P_{UTXY}, P_{\hat{X}|UYT}) := \mathbb{E}[\mathbf{j}(U, X, Y, \hat{X}|T)] \quad (121)$$

$$= \begin{bmatrix} -I(U; Y|T) \\ I(U; X|T) \\ \mathbb{E}[d(X, \hat{X})] \end{bmatrix}. \quad (122)$$

Observe that the sum of the first two components of (122) resembles the Wyner-Ziv rate-distortion function defined in (43). As such when stating an achievable  $(n, \varepsilon)$ -Wyner-Ziv rate-distortion region, we project the first two terms onto an affine subspace representing their sum. See (125) and (126) below.

**Definition 14.** The information-distortion dispersion matrix for the WZ problem for a pair of  $P_{UTXY} \in \tilde{\mathcal{P}}(P_{XY})$  and  $P_{\hat{X}|UYT}$  is defined as

$$\mathbf{V}(P_{UTXY}, P_{\hat{X}|UYT}) := \mathbb{E}_T [\text{Cov}(\mathbf{j}(U, X, Y, \hat{X}|T))]. \quad (123)$$

**Definition 15.** Let  $\mathbf{M} \in \mathbb{R}^{2 \times 3}$  be the matrix

$$\mathbf{M} := \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \quad (124)$$

If  $\mathbf{V}(P_{UTXY}, P_{\hat{X}|UYT}) \neq \mathbf{0}_{3 \times 3}$ , define  $\mathcal{R}_{\text{in}}(n, \varepsilon; P_{UTXY}, P_{\hat{X}|UYT})$  to be the set of all rate-distortion pairs  $(R, D)$  satisfying

$$\begin{bmatrix} R \\ D \end{bmatrix} \in \mathbf{M} \left( \mathbf{J} + \frac{\mathcal{S}(\mathbf{V}, \varepsilon)}{\sqrt{n}} + \frac{2 \log n}{n} \mathbf{1}_3 \right). \quad (125)$$

where  $\mathbf{J} := \mathbf{J}(P_{UTXY}, P_{\hat{X}|UYT})$  and  $\mathbf{V} := \mathbf{V}(P_{UTXY}, P_{\hat{X}|UYT})$ . Else if  $\mathbf{V}(P_{UTXY}, P_{\hat{X}|UYT}) = \mathbf{0}_{3 \times 3}$ , define  $\mathcal{R}_{\text{in}}(n, \varepsilon; P_{UTXY}, P_{\hat{X}|UYT})$  to be the set of all rate-distortion pairs  $(R, D)$  satisfying

$$\begin{bmatrix} R \\ D \end{bmatrix} \in \mathbf{M} \left( \mathbf{J} + \frac{2 \log n}{n} \mathbf{1}_3 \right). \quad (126)$$

In (125), the matrix  $\mathbf{M}$  serves project the three-dimensional set  $\mathbf{J} + \mathcal{S}(\mathbf{V}, \varepsilon)/\sqrt{n} \subset \mathbb{R}^3$  onto two dimensions by linearly combining the first two mutual information terms to give

$I(U; X|T) - I(U; Y|T) = I(U; X|Y, T)$  (by the Markov chain  $U - (X, T) - Y$ ). From the simplified CS-type bound for the WZ problem in Corollary 9 and the multidimensional Berry-Esséen theorem [21], we can derive the following:

**Theorem 18** (Inner Bound to the  $(n, \varepsilon)$ -Wyner-Ziv Rate-Distortion Region). *For every  $0 < \varepsilon < 1$  and all  $n$  sufficiently large, the  $(n, \varepsilon)$ -Wyner-Ziv rate-distortion region  $\mathcal{R}_{\text{WZ}}(n, \varepsilon)$  satisfies*

$$\begin{aligned} & \bigcup_{P_{UTXY} \in \tilde{\mathcal{P}}(P_{XY}), P_{\hat{X}|UYT}} \mathcal{R}_{\text{in}}(n, \varepsilon; P_{UTXY}, P_{\hat{X}|UYT}) \\ & \subset \mathcal{R}_{\text{WZ}}(n, \varepsilon). \end{aligned} \quad (127)$$

Furthermore, the union over a pair of  $P_{UTXY}$  and  $P_{\hat{X}|UYT}$  can be restricted to those distributions for which the supports  $\mathcal{U}$  and  $\mathcal{T}$  of auxiliary random variables  $U$  and  $T$  satisfy that  $|\mathcal{U}| \leq |\mathcal{X}| + 8$  and  $|\mathcal{T}| \leq 9$  respectively.

**Remark 5.** The assumption that the reproduction channel  $P_{\hat{X}|UTX}$  is stochastic is used to establish bounds on the cardinalities of the auxiliary random variables  $U$  and  $T$  (see Remark 10). This is because even though the functional representation lemma [1, Appendix A] ensures that the first two entries of  $\mathbf{j}(u, x, y, \hat{x}|t)$  in (120) are preserved using a deterministic reproduction channel and appropriate bounds on  $|\mathcal{U}|$  and  $|\mathcal{T}|$ , the last entry concerning the distortion  $d(x, \hat{x}|t)$  may not be preserved using the same techniques.

The proof of this result is provided in Appendix K. Further projecting onto the first dimension (the rate) for a fixed distortion level  $D$  yields the following:

**Theorem 19** (Upper Bound to the  $(n, \varepsilon)$ -Wyner-Ziv Rate-Distortion Function). *For every  $0 < \varepsilon < 1$  and all  $n$  sufficiently large, the  $(n, \varepsilon)$ -Wyner-Ziv rate-distortion function  $R_{\text{WZ}}(n, \varepsilon, D)$  satisfies*

$$\begin{aligned} R_{\text{WZ}}(n, \varepsilon, D) & \leq \inf \left\{ R : (R, D) \in \bigcup_{P_{UTXY} \in \tilde{\mathcal{P}}(P_{XY}), P_{\hat{X}|UYT}} \mathcal{R}_{\text{in}}(n, \varepsilon; P_{UTXY}, P_{\hat{X}|UYT}) \right\}. \end{aligned} \quad (128)$$

Theorems 18 and 19 are very similar in spirit to the result on the achievable second-order coding rate for the WAK problem. The marginal contributions from the distortion error event, the packing error event, the covering error event as well as their correlations are all involved in the dispersion matrix  $\mathbf{V}(P_{UTXY}, P_{\hat{X}|UYT})$ .

It is worth mentioning why for the inner bound to the second-order region in Theorem 18, we should, in general, employ stochastic reproduction functions  $P_{\hat{X}|UYT}$  instead of a deterministic ones  $g : \mathcal{U} \times \mathcal{Y} \rightarrow \hat{\mathcal{X}}$ . The reasons are twofold: First, this is to facilitate the bounding of the cardinalities of the auxiliary alphabets  $\mathcal{U}$  and  $\mathcal{T}$  in Theorem 18. This is done using variants of the support lemma [1, Appendix A]. See Lemma 37 and 38 in Appendix M. The preservation of the expected distortion  $\mathbb{E}d(X, \hat{X})$  requires that  $P_{\hat{X}|UYT}$  is stochastic. See Theorem 35 in Appendix M. Second, and more importantly, it is not *a priori* clear without a converse (outer)

bound on  $\mathcal{R}_{\text{WZ}}(n, \varepsilon)$  that the second-order inner bound we have in (127) cannot be enlarged via the use of a stochastic reproduction function  $P_{\hat{X}|UYT}$ . The same observation holds verbatim for the GP problem where we use  $P_{X|US}$  instead of a deterministic encoding function from  $\mathcal{U} \times \mathcal{S}$  to  $\mathcal{X}$ .

At this juncture, it is natural to wonder whether we are able to recover the dispersion for lossy source coding [25], [26] as a special case of Theorem 19 (like Corollary 17 is a special case of Theorem 16). This does not seem straightforward because of the distortion error event in (67). However, we can start from the CS-type bound in (67), set  $Y = \emptyset$ ,  $U = \hat{X}$  and use the method of types [28] or the notion of the  $D$ -tilted information [26] to obtain the specialization for the direct part. Before stating the result, we define a few quantities. Let the rate-distortion function of the source  $X \sim Q \in \mathcal{P}(\mathcal{X})$  be denoted as

$$R(Q, D) := \min_{P_{\hat{X}, X}: P_X = Q, \mathbb{E}d(X, \hat{X}) \leq D} I(X; \hat{X}), \quad (129)$$

where  $\mathbb{E}d(X, \hat{X}) := \sum_{x, \hat{x}} P_{\hat{X}, X}(\hat{x}, x) d(x, \hat{x})$ . Also, define the  $D$ -tilted information to be

$$j(x, D) := -\log \mathbb{E} \left[ \exp \left( \lambda^* D - \lambda^* d(x, \hat{X}^*) \right) \right] \quad (130)$$

where the expectation is with respect to the unconditional distribution of  $\hat{X}^*$ , the output distribution that optimizes the rate-distortion function in (129) and

$$\lambda^* := -\frac{\partial}{\partial D} R(P_X, D). \quad (131)$$

**Theorem 20** (Achievable Second-Order Coding Rate for Lossy Source Coding). *Define the second-order coding rate for lossy source coding to be*

$$\sigma(P_X, D, \varepsilon) := \limsup_{n \rightarrow \infty} \sqrt{n} (R_X(n, \varepsilon; D) - R(P_X, D)) \quad (132)$$

where  $R_X(n, \varepsilon; D)$  is the minimal rate of compression of source  $X \sim P_X$  up to distortion  $D$  at blocklength  $n$  and probability of excess distortion not exceeding  $\varepsilon$ . We have

$$\sigma(P_X, D, \varepsilon) \leq \sqrt{\text{Var}(j(X, D))} Q^{-1}(\varepsilon) \quad (133)$$

Two proofs of Theorem 20 are provided in Appendix L, one based on the method of types and the other based on the  $D$ -tilted information in (130). For the former proof based on the method of types, we need to assume that  $Q \mapsto R(Q, D)$  is differentiable in a small neighborhood of  $P_X$  and  $P_X$  is supported on a finite set. For the second proof,  $\mathcal{X}$  can be an abstract alphabet. Note that  $R(P_X, D) = \mathbb{E}_{X \sim P_X} [j(X, D)]$ . We remark that for discrete memoryless sources, the  $D$ -tilted information  $j(x, D)$  coincides with the derivative of the rate-distortion function with respect to the source [25]

$$R'(x, D) = \frac{\partial}{\partial Q(x)} R(Q, D) \Big|_{Q=P_X}. \quad (134)$$

**C. Achievable Second-Order Coding Rates for the GP problem**

We conclude this section by stating and achievable second-order coding rate for the GP problem by presenting a lower

bound to the  $(n, \varepsilon, \Gamma)$ -capacity  $C_{\text{GP}}(n, \varepsilon, \Gamma)$  defined in (37). As in the previous two subsections, we start with definitions. For two finite sets  $\mathcal{U}$  and  $\mathcal{T}$ , define  $\tilde{\mathcal{P}}(W, P_S)$  to be the collection of all  $P_{UTSXY} \in \mathcal{P}(\mathcal{U} \times \mathcal{T} \times \mathcal{S} \times \mathcal{X} \times \mathcal{Y})$  such that the  $\mathcal{S}$ -marginal of  $P_{UTSXY}$  is  $P_S$ ,  $P_{Y|XS} = W$ ,  $U - (X, S, T) - Y$  forms a Markov chain and  $T$  is independent of  $S$ . Note that  $P_{UTSXY}$  does not necessarily have to satisfy the cost constraint in (45).

In addition, to facilitate the time-sharing for the cost function, we define

$$g(X|T) := g(X_T) \quad (135)$$

where  $X_t$  for any  $t \in \mathcal{T}$  has distribution  $P_{X|T=t}$ .

**Definition 16.** *The information-density-cost vector for the GP problem for  $P_{UTSXY} \in \tilde{\mathcal{P}}(W, P_S)$  is defined as*

$$\mathbf{j}(U, S, X, Y|T) := \begin{bmatrix} \log \frac{P_{Y|UT}(Y|U, T)}{P_{Y|T}(Y|T)} \\ -\log \frac{P_{S|UT}(S|U, T)}{P_S(S)} \\ -g(X|T) \end{bmatrix}. \quad (136)$$

Since  $\sum_t P_T(t) \mathbb{E}_{P_{X|T}}[g(X_T)|T = t] = \mathbb{E}[g(X)]$ , the expectation of this vector with respect to  $P_{UTSXY}$  is the vector of mutual informations and the negative cost, i.e.,

$$\mathbf{J}(P_{UTSXY}) := \mathbb{E}[\mathbf{j}(U, S, X, Y|T)] = \begin{bmatrix} I(U; Y|T) \\ -I(U; S|T) \\ -\mathbb{E}[g(X)] \end{bmatrix}. \quad (137)$$

**Definition 17.** *The information-dispersion matrix for the GP problem for  $P_{UTSXY} \in \tilde{\mathcal{P}}(W, P_S)$  is defined as*

$$\mathbf{V}(P_{UTSXY}) := \mathbb{E}_T[\text{Cov}(\mathbf{j}(U, S, X, Y|T))]. \quad (138)$$

**Definition 18.** *Let  $\mathbf{M}$  be the matrix defined in (124). If  $\mathbf{V}(P_{UTSXY}) \neq \mathbf{0}_{3 \times 3}$ , define the set  $\mathcal{R}_{\text{in}}(n, \varepsilon; P_{UTSXY})$  to be the set of all rate-cost pairs  $(R, \Gamma)$  satisfying*

$$\begin{bmatrix} R \\ -\Gamma \end{bmatrix} \in \mathbf{M} \left( \mathbf{J} - \frac{\mathcal{S}(\mathbf{V}, \varepsilon)}{\sqrt{n}} - \frac{2 \log n}{n} \mathbf{1}_3 \right) \quad (139)$$

where  $\mathbf{J} := \mathbf{J}(P_{UTSXY})$  and  $\mathbf{V} := \mathbf{V}(P_{UTSXY})$ . Else if  $\mathbf{V}(P_{UTSXY}, g) \neq \mathbf{0}_{3 \times 3}$ , define  $\mathcal{R}_{\text{in}}(n, \varepsilon; P_{UTSXY})$  to be the set of all rate-cost pairs  $(R, \Gamma)$  satisfying

$$\begin{bmatrix} R \\ -\Gamma \end{bmatrix} \in \mathbf{M} \left( \mathbf{J} - \frac{2 \log n}{n} \mathbf{1}_3 \right). \quad (140)$$

By leveraging on our finite blocklength CS-type bound for the GP problem in (71), we obtain the following:

**Theorem 21** (Inner Bound to the  $(n, \varepsilon)$ -GP Capacity-Cost Region). *For every  $0 < \varepsilon < 1$  and all  $n$  sufficiently large, the  $(n, \varepsilon)$ -GP capacity-cost region  $\mathcal{C}_{\text{GP}}(n, \varepsilon)$  satisfies*

$$\bigcup_{P_{UTSXY} \in \tilde{\mathcal{P}}(W, P_S)} \mathcal{R}_{\text{in}}(n, \varepsilon; P_{UTSXY}) \subset \mathcal{C}_{\text{GP}}(n, \varepsilon). \quad (141)$$

Furthermore, the union over  $P_{UTSXY}$  can be restricted to those distributions for which the supports  $\mathcal{U}$  and  $\mathcal{T}$  of auxiliary random variables  $U$  and  $T$  satisfy that  $|\mathcal{U}| \leq |\mathcal{S}||\mathcal{X}| + 6$  and  $|\mathcal{T}| \leq 9$  respectively.

The assumption that the encoding function  $P_{X|US}$  is stochastic appears to be necessary for establishing bounds



on  $|\mathcal{U}|$  and  $|\mathcal{T}|$ . See Remark 5. By projecting onto the first dimension (the rate) for a fixed cost  $\Gamma \geq 0$ , we obtain:

**Theorem 22** (Lower Bound to the  $(n, \varepsilon)$ -GP Capacity). *For every  $0 < \varepsilon < 1$  and all  $n$  sufficiently large, the  $(n, \varepsilon)$ -GP capacity-cost function  $C_{\text{GP}}(n, \varepsilon, \Gamma)$  satisfies*

$$C_{\text{GP}}(n, \varepsilon, \Gamma) \geq \sup \left\{ R : (R, \Gamma) \in \bigcup_{P_{UTSXY} \in \tilde{\mathcal{P}}(W, P_S)} \mathcal{R}_{\text{in}}(n, \varepsilon; P_{UTSXY}) \right\}. \quad (142)$$

The proof of Theorem 21 parallels that for the WZ case in Theorem 18 so it is omitted for brevity. The matrix  $\mathbf{M}$  serves to project the first two components of each element in the set  $\mathbf{J} + \mathcal{S}(\mathbf{V}, \varepsilon)/\sqrt{n}$  onto one dimension. Indeed, for a fixed  $P_{UTSXY} \in \tilde{\mathcal{P}}(W, P_S)$ , the first two components read  $I(U; Y|T) - I(U; S|T)$  which, if  $T = \emptyset$  and the random variables  $(U, S, X, Y)$  are capacity-achieving, reduces to the GP formula in (46). Hence, the set  $\mathbf{M}\mathcal{S}(\mathbf{V}, \varepsilon)/\sqrt{n} \subset \mathbb{R}$  quantifies all possible backoffs from the asymptotic GP capacity-cost region  $\mathcal{C}_{\text{GP}}$  (defined in (36)) at blocklength  $n$  and average error probability  $\varepsilon$  based on our CS-type finite blocklength bound for the GP problem in (71). The bound in (142) is clearly much tighter than the one provided in [10] which is based on the use of Wyner's PBL and Markov lemma.

Now by setting  $S = T = \emptyset$ ,  $U = X$  and  $\Gamma = \infty$  in Theorem 22, we recover the direct part of the second-order coding rate for channel coding without cost constraints [12], [22], [39].

**Corollary 23** (Achievable Second-Order Coding Rate for Channel Coding). *Fix a non-exotic [22] discrete memoryless channel  $W : \mathcal{X} \rightarrow \mathcal{Y}$  with channel capacity  $C(W) = \max_{P_X} I(X; Y)$ . Define the second-order coding rate for channel coding to be*

$$\sigma(W, \varepsilon) := \limsup_{n \rightarrow \infty} \sqrt{n}(C(W) - C_W(n, \varepsilon)) \quad (143)$$

where  $C_W(n, \varepsilon)$  is the maximal rate of transmission over the channel  $W$  at blocklength  $n$  and average error probability  $\varepsilon$ . Then,

$$\sigma(W, \varepsilon) \leq \min_{P_{X^*}} \sqrt{\text{Var} \left( \log \frac{W(Y^*|X^*)}{P_{Y^*}(Y^*)} \right)} Q^{-1}(\varepsilon) \quad (144)$$

where  $(X^*, Y^*) \sim P_{X^*} \times W$  and the minimization is over all capacity-achieving input distributions.

The bound in (144) has long been known to be an equality [39]. Note that the unconditional dispersion in (144)  $\text{Var} \left( \log \frac{W(Y^*|X^*)}{P_{Y^*}(Y^*)} \right)$  coincides with the conditional dispersion [22] since it is being evaluated at a capacity-achieving input distribution. As such, the converse can be proved using the meta-converse in [22] or an modification of the Verdú-Han converse [7, Lem. 3.2.2] with an judiciously chosen output distribution as was done in [12]. In fact, we can also derive a generalization of Corollary 23 with cost constraints incorporated [12, Thm. 3] using similar techniques as in the proof of Theorem 20. Namely, we use a uniform distribution

over a particular type class (constant composition codes) as the input distribution. The type is chosen to be close to the optimal input distribution (assuming it is unique).

## VII. NUMERICAL EXAMPLES

### A. Numerical Example for WAK Problem

In this section, we use an example to illustrate the inner bound on  $(n, \varepsilon)$ -optimal rate region for the WAK problem obtained in Theorem 15. We neglect the small  $O\left(\frac{\log n}{n}\right)$  term. The source is taken to be a discrete symmetric binary source DSBS( $\alpha$ ), i.e.,

$$P_{XY} = \frac{1}{2} \begin{bmatrix} 1 - \alpha & \alpha \\ \alpha & 1 - \alpha \end{bmatrix}. \quad (145)$$

In this case, the optimal rate region reduces to

$$\mathcal{R}_{\text{WAK}}^* = \left\{ (R_1, R_2) : R_1 \geq h(\beta * \alpha), \right. \\ \left. R_2 \geq 1 - h(\beta), 0 \leq \beta \leq \frac{1}{2} \right\}, \quad (146)$$

where  $h(\cdot)$  is the binary entropy function and  $\beta * \alpha := \beta(1 - \alpha) + (1 - \beta)\alpha$  is the binary convolution. The above region is attained by setting the backward test channel from  $U$  to  $Y$  to be a BSC with some crossover probability  $\beta$ . All the elements in the entropy-information dispersion matrix  $\mathbf{V}(\beta)$  can be evaluated in closed form in terms of  $\beta$ . Define  $\mathbf{J}(\beta) := [h(\beta * \alpha), 1 - h(\beta)]^T$ . In Fig. 5, we plot the second-order region

$$\tilde{\mathcal{R}}_{\text{in}}(n, \varepsilon) := \bigcup_{0 \leq \beta \leq \frac{1}{2}} \left\{ (R_1, R_2) : \mathbf{R} \in \mathbf{J}(\beta) + \frac{\mathcal{S}(\mathbf{V}(\beta), \varepsilon)}{\sqrt{n}} \right\}. \quad (147)$$

The first-order region  $\mathcal{R}_{\text{WAK}}^*$  and the second-order region with simple time-sharing ( $|\mathcal{T}| = 2$ ) are also shown for comparison. More precisely, the simple time-sharing is between  $\beta = 0$  and  $\beta = 1/2$ . As expected, as the block length increases, the  $(n, \varepsilon)$ -optimal rate region tends to the first-order one. Interestingly, at small block length, time-sharing makes the second-order  $(n, \varepsilon)$ -optimal rate region in (147) larger compared to that without time-sharing. Especially, the simple time-sharing is better than  $\tilde{\mathcal{R}}_{\text{in}}(n, \varepsilon)$  for  $n = 500$  because the rank of the entropy-information dispersion matrix  $\lambda \mathbf{V}(0) + (1 - \lambda) \mathbf{V}(1/2)$  for  $0 < \lambda \leq 1$  is one.<sup>6</sup>

We also consider the region  $\tilde{\mathcal{R}}_{\text{in}}^{\text{V}}(n, \varepsilon)$  which is the analogue of  $\tilde{\mathcal{R}}_{\text{in}}(n, \varepsilon)$  but derived from Verdú's bound in (8). In Fig. 6, we compare the second-order coefficients, namely that derived from our bound  $\mathcal{S}(\mathbf{V}(\beta), \varepsilon)$  and

$$\mathcal{S}^{\text{V}}(\mathbf{V}(\beta), \varepsilon) := \bigcup_{0 \leq \lambda \leq 1} \left\{ (z_1, z_2) : z_1 \geq \sqrt{V_H(\beta)} Q^{-1}(\lambda \varepsilon), \right. \\ \left. z_2 \geq \sqrt{V_I(\beta)} Q^{-1}((1 - \lambda) \varepsilon) \right\}. \quad (148)$$

Note that the difference between the two regions is quite small even for  $\varepsilon = 0.5$ . This is because, for this example, the covariance of the entropy- and information-density (off-diagonal in the dispersion matrix) is negative so the difference

<sup>6</sup>It should be noted that the rank of  $\mathbf{V}(1/2)$  is zero.

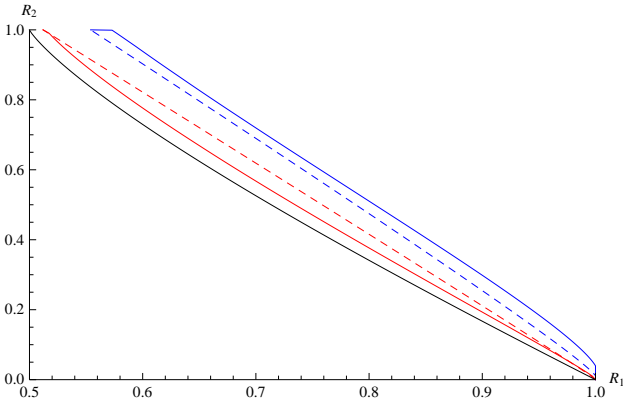


Fig. 5. A comparison between  $\tilde{\mathcal{R}}_{\text{in}}(n, \varepsilon)$  without time-sharing (solid line) and the time-sharing region (dashed line) for  $\varepsilon = 0.1$ . The regions are to the top right of the curves. The blue and red curves are for  $n = 500$  and  $n = 10,000$  respectively. The black curve is the first-order region (1).

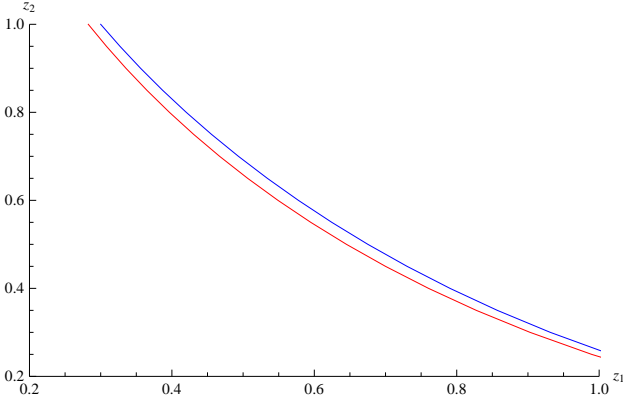


Fig. 6. A comparison between  $\mathcal{S}(\mathbf{V}(\beta), \varepsilon)$  (defined in (96)) and  $\mathcal{S}^V(\mathbf{V}(\beta), \varepsilon)$  (defined in (148)) for  $\beta = h^{-1}(0.5)$  and  $\varepsilon = 0.5$ . The red and blue curves are the boundaries of  $\mathcal{S}(\mathbf{V}(\beta), \varepsilon)$  and  $\mathcal{S}^V(\mathbf{V}(\beta), \varepsilon)$  respectively. The regions lie to the top right of the curves.

between  $\Pr(Z_1 \geq z_1 \text{ or } Z_2 \geq z_2)$  and  $\Pr(Z_1 \geq z_1) + \Pr(Z_2 \geq z_2)$  is small. In this case, the 2-dimensional Gaussian  $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{V}(\beta))$  has a negative covariance and hence the probability mass in the first and third quadrants are small. Hence, the union bound is not very loose in this case.

Next, we consider the binary joint source given by  $P_{X|Y}(1|0) = P_{X|Y}(0|1) = \alpha$  and  $P_Y(0) = p \leq \frac{1}{2}$ , which is a generalization of (145). This example was investigated in [53], and the optimal rate region reduces to

$$\mathcal{R}_{\text{WAK}}^* = \left\{ (R_1, R_2) : R_1 \geq h(\beta * \alpha), \right. \\ \left. R_2 \geq h(p) - h(\beta), 0 \leq \beta \leq p \right\}. \quad (149)$$

The above region is attained by setting the backward test channel from  $U$  to  $Y$  to be BSC with some crossover probability  $0 \leq \beta \leq p$ . All the elements in the entropy-information dispersion matrix  $\mathbf{V}(\beta)$  can be evaluated in closed form in terms of  $\beta$ . Define  $\mathbf{J}(\beta) := [h(\beta * \alpha), h(p) - h(\beta)]^T$ . In Fig. 7,

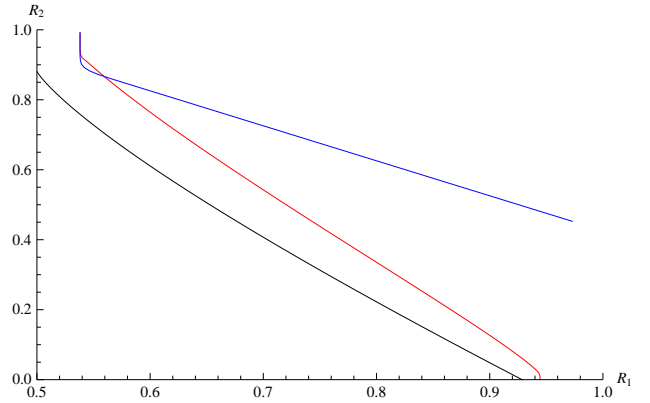


Fig. 7. A comparison between  $\tilde{\mathcal{R}}_{\text{in}}(n, \varepsilon)$  (red solid curve) and the bound from Remark 4 (blue solid curve) for  $\varepsilon = 0.1$  and  $n = 1000$ . The regions are to the top right of the curves.

we plot the second-order region

$$\tilde{\mathcal{R}}_{\text{in}}(n, \varepsilon) := \bigcup_{0 \leq \beta \leq p} \left\{ (R_1, R_2) : \mathbf{R} \in \mathbf{J}(\beta) + \frac{\mathcal{S}(\mathbf{V}(\beta), \varepsilon)}{\sqrt{n}} \right\}. \quad (150)$$

For comparison, we also plot the second-order region derived from Remark 4. Around the corner point defined by the entropies  $[H(X|Y), H(Y)]^T = [h(\beta), h(p)]^T$ , we find that the bound from Remark 4 is tighter than that given by (150).

### B. Numerical Example for GP Problem

In this section, we use an example to illustrate the inner bound on  $(n, \varepsilon)$ -optimal rate for the GP problem obtained in Theorem 21. We do not consider cost constraints here, i.e.,  $\Gamma = \infty$ . We also neglect the small  $O\left(\frac{\log n}{n}\right)$  term. We consider the *memory with stuck-at faults* example [54] (see also [1, Example 7.3]). The state  $S = 0$  correspond to a faculty memory cell that output 0 independent of the input value, the state  $S = 1$  corresponds to a faculty memory cell that outputs 1 independent of the input value, and the state  $S = 2$  corresponds to a binary symmetric channel with crossover probability  $\alpha$ . The probabilities of these states are  $\frac{p}{2}$ ,  $\frac{p}{2}$ , and  $1 - p$  respectively.

It is known [54] that the capacity is

$$C_{\text{GP}}^* = (1 - p)(1 - h(\alpha)). \quad (151)$$

The above capacity is attained by setting  $\mathcal{U} = \{0, 1\}$  and  $P_{U|X}(0|0) = P_{U|S}(1|1) = 1 - \alpha$ ,  $P_{U|S}(u|2) = \frac{1}{2}$ , and  $X = U$ . All the elements in the information dispersion matrix  $\mathbf{V}$  can be evaluated in closed form. In Fig. 8, we plot the second-order capacity

$$\tilde{R}_{\text{GP}}(n, \varepsilon; p, \alpha) := (1 - p)(1 - h(\alpha)) \\ - \frac{1}{\sqrt{n}} \min\{z_1 + z_2 : (z_1, z_2) \in \mathcal{S}(\mathbf{V}, \varepsilon)\}. \quad (152)$$

For comparison, let us consider the case in which the decoder, instead of the encoder, can access the state  $S$ . In this case, we can regard  $X$  as the channel input and  $(S, Y)$  as the channel output. It is known [54] that the capacity  $C(W)$

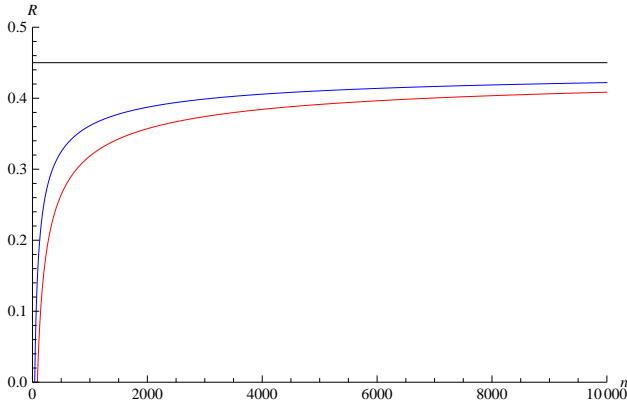


Fig. 8. A comparison between  $\tilde{R}_{\text{GP}}(n, \varepsilon; p, \alpha)$  (red solid line) and  $\tilde{C}(n, \varepsilon; p, \alpha)$  (blue solid line) for  $\varepsilon = 0.001$ ,  $p = 0.1$ , and  $\alpha = 0.11$ . The black solid line is the first-order capacity (151).

of this channel is the same as (151). The dispersion  $V$  can be evaluated in closed form by appealing to the law of total variance [55]. In Fig. 8, we also plot the second order capacity

$$\tilde{C}(n, \varepsilon; p, \alpha) := (1-p)(1-h(\alpha)) - \sqrt{\frac{V}{n}} Q^{-1}(\varepsilon). \quad (153)$$

From the figure, we can find that the lower bound  $\tilde{R}_{\text{GP}}(n, \varepsilon; p, \alpha)$  on the GP  $(n, \varepsilon)$ -optimal rate is smaller than the  $(n, \varepsilon)$ -optimal rate with decoder side-information though the first order rates coincide.

## VIII. CONCLUSION AND FURTHER WORK

### A. Summary

In this paper, we proved new non-asymptotic bounds on the error probability for side-information coding problems, including the WAK, WZ and GP problems. These bounds then yield known general formulas as simple corollaries. In addition, we used these bounds to provide achievable second-order coding rates for these three side-information problems. We argued that when evaluated using i.i.d. test channels, the second-order rates evaluated using our non-asymptotic bounds are the best known in the literature including [6].

### B. Further Work on Non-Asymptotic and Second-Order Achievability Bounds

Other challenging problems involving the derivation of non-asymptotic achievability bounds for multi-terminal problems include the Heegard-Berger [1, Sec. 11.4] problem, multiple description coding [1, Ch. 13], Marton's inner bound for the broadcast channel [1, Thm. 8.3], and hypothesis testing with multi-terminal data compression [56]. Achievable second-order coding rate regions for some of these problems have been derived independently and concurrently by Yassaee-Aref-Gohari [35], [38] using a completely different technique as discussed in the Introduction but it may be interesting to verify if the technique contained in this paper can be adapted to the above-mentioned coding problems.

### C. Further Work on Non-Asymptotic and Second-Order Converse Bounds

A natural question that arises from this work is whether one can derive non-asymptotic converse bounds that, when suitably specialized, coincide with the second-order achievability bounds in Section VI. Apart from the Slepian-Wolf problem [23], [44] and the Gaussian MAC with degraded message sets [57], this has not been done for other problems in network information theory. Because second-order converse bounds imply the strong converse, it appears that first establishing a strong converse provides intuition for establishing non-asymptotic converse bounds that are tight in the second-order sense after asymptotic evaluation.

To the best of the authors' knowledge, there are only three approaches that may be used to obtain second-order converses for network problems whose first-order (capacity region) characterization involve auxiliary random variables. The first is the information spectrum method. For example [58, Lem. 2] provides a non-asymptotic converse bound for the asymmetric broadcast channel. However, the evaluation is not efficiently computable for large (or even moderate)  $n$  as one has to perform an exhaustive search over the space of all  $n$ -letter auxiliary random variables (or equivalently  $n$ -letter joint distributions). The second is the entropy and image size characterization technique [29] based on the blowing-up lemma [28, Ch. 5]. This has been used to prove the strong converse for the WAK problem [29] and the GP problem [31]. However, the use of the blowing-up approach to obtain second-order converse bounds is not straightforward. The third method involves a non-standard change-of-measure argument and was used in the work of Kelly and Wagner [51, Thm. 2] to prove an upper bound on the error exponent for WAK coding. Again, it does not appear, at first glance, that this argument is amenable to second-order analysis.

## APPENDIX A

### PROOF OF PROPOSITION 1 (EXPURGATED CODE)

*Proof:* Let  $x_0 \in \mathcal{X}$  be a prescribed constant satisfying  $g(x_0) \leq \Gamma$ , and let  $P_X^*$  be the distribution such that  $P_X^*(x_0) = 1$ , i.e.,  $P_X^*(x) = \mathbf{1}[x = x_0]$ . Then, we define

$$\begin{aligned} \tilde{P}_{X|MS}(x|m, s) &:= P_{X|MS}(x|m, s) \mathbf{1}[g(x) \leq \Gamma] \\ &\quad + P_{X|MS}(\mathcal{T}_g^{\text{GP}}(\Gamma)^c|m, s) P_X^*(x). \end{aligned} \quad (154)$$

Then, it is obvious that  $\tilde{P}_X(\mathcal{T}_g^{\text{GP}}(\Gamma)) = 1$ . We also have

$$\begin{aligned} &\tilde{P}_{MSXY\hat{M}}[m \neq \hat{m}] \\ &= \sum_{\substack{m, \hat{m} \\ m \neq \hat{m}}} \sum_{s, x, y} P_M(m) P_S(s) \tilde{P}_{X|MS}(x|m, s) \\ &\quad \times W(y|x, s) P_{\hat{M}|Y}(\hat{m}|y) \\ &= \sum_{\substack{m, \hat{m} \\ m \neq \hat{m}}} \sum_{s, x, y} P_M(m) P_S(s) P_{X|MS}(x|m, s) \\ &\quad \times W(y|x, s) P_{\hat{M}|Y}(\hat{m}|y) \mathbf{1}[g(x) \leq \Gamma] \end{aligned} \quad (155)$$

$$+ \sum_{\substack{m, \hat{m} \\ m \neq \hat{m}}} \sum_{s, x, y} P_M(m) P_S(s) P_{X|MS}(\mathcal{T}_g^{\text{GP}}(\Gamma)^c | m, s) \\ \times P_X^*(x) W(y|x, s) P_{\hat{M}|Y}(\hat{m}|y) \quad (156)$$

$$\leq \sum_{\substack{m, \hat{m} \\ m \neq \hat{m}}} \sum_{s, x, y} P_M(m) P_S(s) P_{X|MS}(x|m, s) \\ \times W(y|x, s) P_{\hat{M}|Y}(\hat{m}|y) \mathbf{1}[\mathbf{g}(x) \leq \Gamma] \\ + \sum_{m, \hat{m}} \sum_{s, x, y} P_M(m) P_S(s) P_{X|MS}(\mathcal{T}_g^{\text{GP}}(\Gamma)^c | m, s) \\ \times P_X^*(x) W(y|x, s) P_{\hat{M}|Y}(\hat{m}|y) \quad (157)$$

$$= \sum_{\substack{m, \hat{m} \\ m \neq \hat{m}}} \sum_{s, x, y} P_M(m) P_S(s) P_{X|MS}(x|m, s) \\ \times W(y|x, s) P_{\hat{M}|Y}(\hat{m}|y) \mathbf{1}[\mathbf{g}(x) \leq \Gamma] \\ + \sum_{m, s} P_M(m) P_S(s) P_{X|MS}(\mathcal{T}_g^{\text{GP}}(\Gamma)^c | m, s) \quad (158)$$

$$= P_{MSXY\hat{M}}[\mathbf{g}(x) \leq \Gamma \cap m \neq \hat{m}] + P_{MSXY\hat{M}}[\mathbf{g}(x) > \Gamma] \quad (159)$$

$$= P_{MSXY\hat{M}}[\mathbf{g}(x) > \Gamma \cup m \neq \hat{m}] \quad (160)$$

as desired.  $\blacksquare$

## APPENDIX B CHANNEL RESOLVABILITY

In this appendix, we review notations and known results for channel resolvability [7, Ch. 6] [13] [14] [17].

As a start, we first review the properties of the variational distance. Let  $\mathcal{P}'(\mathcal{U})$  be the set of all sub-normalized non-negative functions (not necessarily probability distribution unless otherwise stated) on a finite set  $\mathcal{U}$ . Note that if  $P \in \mathcal{P}'(\mathcal{U})$  is normalized then  $P \in \mathcal{P}(\mathcal{U})$ , i.e.,  $P$  is a distribution on  $\mathcal{U}$ . For  $P, Q \in \mathcal{P}'(\mathcal{U})$ , we define the variational distance (divided by 2) as

$$d(P, Q) = \frac{1}{2} \sum_{u \in \mathcal{U}} |P(u) - Q(u)|. \quad (161)$$

For two sets  $\mathcal{U}$  and  $\mathcal{Z}$ , let  $\mathcal{P}'(\mathcal{Z}|\mathcal{U})$  be the set of all sub-normalized non-negative functions indexed by  $u \in \mathcal{U}$ . When  $W \in \mathcal{P}'(\mathcal{Z}|\mathcal{U})$  is normalized, it is a channel. In this section, we denote the joint distribution induced by  $P \in \mathcal{P}(\mathcal{U})$  and  $W \in \mathcal{P}'(\mathcal{Z}|\mathcal{U})$  as  $PW \in \mathcal{P}'(\mathcal{U} \times \mathcal{Z})$ . The following properties are useful in the proof of theorems. Since the proofs are almost the same as well known properties of the variational distance for normalized distributions, we omit the proofs.

**Lemma 24.** *The variational distance satisfies the following properties.*

- 1) *The monotonicity with respect to marginalization: For  $P, Q \in \mathcal{P}'(\mathcal{U})$  and  $W, V \in \mathcal{P}'(\mathcal{Z}|\mathcal{U})$ , let  $P', Q' \in \mathcal{P}'(\mathcal{Z})$  be*

$$P'(z) := \sum_{u \in \mathcal{U}} P(u) W(z|u), \quad Q'(z) := \sum_{u \in \mathcal{U}} Q(u) V(z|u). \quad (162)$$

Then,

$$d(P', Q') \leq d(PW, QV). \quad (163)$$

- 2) *The data-processing inequality: For  $P, Q \in \mathcal{P}'(\mathcal{U})$  and  $W \in \mathcal{P}'(\mathcal{Z}|\mathcal{U})$ ,*

$$d(PW, QW) \leq d(P, Q). \quad (164)$$

*In particular, when  $W \in \mathcal{P}(\mathcal{Z}|\mathcal{U})$ , the equality holds in (164).*

- 3) *For a distribution  $P \in \mathcal{P}(\mathcal{U})$ , a sub-normalized measure  $Q \in \mathcal{P}'(\mathcal{U})$ , and any subset  $\Gamma \subset \mathcal{U}$ ,*

$$P(\Gamma) \leq Q(\Gamma) + d(P, Q) + \frac{1 - Q(\mathcal{U})}{2}. \quad (165)$$

**Remark 6.** Combining (163) for  $V = W$  and (164), we have

$$d(P', Q') \leq d(P, Q). \quad (166)$$

Although the above inequality is usually referred as the data-processing inequality, we will use (164) in the proofs of non-asymptotic bounds.

Next, we introduce the concept of *smoothing* of a distribution [59]. For a distribution  $P \in \mathcal{P}(\mathcal{U})$  and a subset  $\mathcal{T} \subset \mathcal{U}$ , a smoothed sub-normalized function  $\bar{P}$  of  $P$  is derived by

$$\bar{P}(u) := P(u) \mathbf{1}[u \in \mathcal{T}]. \quad (167)$$

Note that the distance between the original distribution and a smoothed one is

$$d(P, \bar{P}) = \frac{P(\mathcal{T}^c)}{2}. \quad (168)$$

Similarly, for a channel  $W: \mathcal{U} \rightarrow \mathcal{Z}$  and a subset  $\mathcal{T} \subset \mathcal{U} \times \mathcal{Z}$ , a smoothed one  $\bar{W} \in \mathcal{P}'(\mathcal{Z}|\mathcal{U})$  is derived by

$$\bar{W}(z|u) := W(z|u) \mathbf{1}[(u, z) \in \mathcal{T}] \quad (169)$$

and it satisfies

$$d(PW, P\bar{W}) = \frac{PW(\mathcal{T}^c)}{2}, \quad (170)$$

where  $PW \in \mathcal{P}(\mathcal{U} \times \mathcal{Z})$  is the joint distribution induced by  $P$  and  $W$ .

Now, we consider the problem of channel resolvability. Let a channel  $P_{Z|U}: \mathcal{U} \rightarrow \mathcal{Z}$  and an input distribution  $P_U$  be given. We would like to approximate the output distribution

$$P_Z(z) = \sum_{u \in \mathcal{U}} P_U(u) P_{Z|U}(z|u) \quad (171)$$

by using  $P_{Z|U}$  and as small an amount of randomness as possible. This is done by means of a designing a deterministic map from a finite set  $\mathcal{I}$  to a codebook  $\mathcal{C} = \{u_i\}_{i \in \mathcal{I}} \subset \mathcal{U}$ . For a given resolvability code  $\mathcal{C}$ , let

$$P_{\bar{Z}}(z) = \sum_{i \in \mathcal{I}} \frac{1}{|\mathcal{I}|} P_{Z|U}(z|u_i) \quad (172)$$

be the simulated output distribution. The approximation error is evaluated by the distance  $d(P_{\bar{Z}}, P_Z)$ .

We consider using the random coding technique as follows. We randomly and independently generate codewords  $u_1, u_2, \dots, u_{|\mathcal{I}|}$  according to  $P_U$ . To derive an upper bound

on the averaged approximation error  $\mathbb{E}_C [d(P_{\bar{Z}}, P_Z)]$ , it is convenient to consider a smoothing operation defined as follows. For the set

$$\mathcal{T}_c(\gamma_c) := \left\{ (u, z) : \log \frac{P_{Z|U}(z|u)}{P_Z(z)} \leq \gamma_c \right\}, \quad (173)$$

let

$$\bar{P}_{Z|U}(z|u) := P_{Z|U}(z|u) \mathbf{1}[(u, z) \in \mathcal{T}_c(\gamma_c)]. \quad (174)$$

Moreover, for fixed resolvability code  $\mathcal{C} = \{u_1, \dots, u_{|\mathcal{I}|}\}$ , let

$$\bar{P}_{\bar{Z}}(z) := \sum_{i \in \mathcal{I}} \frac{1}{|\mathcal{I}|} \bar{P}_{Z|U}(z|u_i). \quad (175)$$

Then, we have the following lemma known as *soft covering*, which is an improvement of [14, Lemma 2].

**Lemma 25** (Corollary 7.2 of [17]). *For any  $\gamma_c \geq 0$ , we have*

$$\mathbb{E}_C [d(\bar{P}_{\bar{Z}}, \bar{P}_Z)] \leq \frac{\Delta(\gamma_c, P_{UZ})}{2\sqrt{|\mathcal{I}|}} \quad (176)$$

where  $\bar{P}_Z(z) = \sum_u P_U(u) \bar{P}_{Z|U}(z|u)$ .

**Remark 7.** *Although the statement of [17, Corollary 7.2] consists of two terms, the second term corresponds to the right hand side of (176). Since our target distribution  $\bar{P}_Z$  is smoothed, the first term of [17, Corollary 7.2] does not appear in (176).*

## APPENDIX C SIMULATION OF TEST CHANNEL

In this appendix, we develop two lemmas which form crucial components of the proof of all CS-type bounds. To do this, we consider the problem related to channel simulation [15]–[17], [60], [61]. Roughly speaking, the problem is described as follows. For a given message set  $\mathcal{L}$  and a code  $\mathcal{C} = \{u_1, \dots, u_{|\mathcal{L}|}\}$ , our goal is to construct a stochastic map  $\varphi : \mathcal{Z} \rightarrow \mathcal{L}$  such that the joint distribution  $P_{\hat{L}Z}$  of  $(\varphi(Z), Z)$  is indistinguishable from  $P_{L\bar{Z}}$ , where  $P_{L\bar{Z}}$  is the joint distribution such that  $u_L$  is sent over the channel  $P_{Z|U}$  for the uniform random number  $L$  on  $\mathcal{L}$ . This is done by the argument of the *likelihood encoder* [17] (see also [62]). However, we need to modify the argument in [17] since our goal is, in fact, to approximate a smoothed version of  $P_{L\bar{Z}}$ . We will use notations introduced in Appendix B.

**Remark 8.** *In the earlier version of this paper [63], we were considering exactly the problem of channel simulation, where we simulate the joint distribution  $P_{UZ}$  by the aid of the common randomness. However, simulating the marginal  $P_U$  is unnecessary to derive bounds on WAK, WZ, and GP problems. Thus, we consider approximation of  $P_{L\bar{Z}}$  in this paper, which enables us to remove a residual term in [63] that stems from the use of the common randomness.*

To construct a stochastic map from  $\mathcal{Z}$  to  $\mathcal{L}$ , we first consider the channel resolvability code as follows. Let us generate a codebook  $\mathcal{C} = \{u_1, \dots, u_{|\mathcal{L}|}\}$ , where each codeword  $u_l$  is randomly and independently generated from  $P_U$ , which is the marginal of  $P_{UZ}$ . Let  $L$  be the uniform random numbers on  $\mathcal{L}$ .

Moreover, let  $\bar{P}_{Z|U}$  be a smoothed version of  $P_{Z|U}$  defined in (174). Then,  $\mathcal{C}$ ,  $L$ , and  $\bar{P}_{Z|U}$  induce the sub-normalized measure

$$\bar{P}_{L\bar{Z}}(l, z) := \frac{1}{|\mathcal{L}|} \bar{P}_{Z|U}(z|u_l). \quad (177)$$

Marginal  $\bar{P}_{\bar{Z}}$  is also induced as

$$\bar{P}_{\bar{Z}}(z) = \sum_l \frac{1}{|\mathcal{L}|} \bar{P}_{Z|U}(z|u_l). \quad (178)$$

Now, we define a stochastic map  $\varphi_C : \mathcal{Z} \rightarrow \mathcal{L}$  as<sup>7</sup>

$$\varphi_C(l|z) = \frac{\bar{P}_{L\bar{Z}}(l, z)}{\bar{P}_{\bar{Z}}(z)}. \quad (179)$$

Let  $\hat{L}$  be the output of the stochastic map  $\varphi_C$  for the input  $Z$ . Then, the joint distribution of  $\hat{L}$  and  $Z$  is given by

$$P_{\hat{L}Z}(l, z) = P_Z(z) \varphi_C(l|z). \quad (180)$$

We also introduce a smoothed version of  $P_{\hat{L}Z}$  as follows:

$$\bar{P}_{\hat{L}Z}(l, z) = \bar{P}_Z(z) \varphi_C(l|z), \quad (181)$$

where  $\bar{P}_Z$  is the marginal of  $\bar{P}_{UZ} := P_U \bar{P}_{Z|U}$ ; i.e.  $\bar{P}_Z(z) := \sum_u P_U(u) \bar{P}_{Z|U}(z|u)$ .

Now, we prove two lemmas which can be used to evaluate the performance of the approximation of  $\bar{P}_{L\bar{Z}}$ .

**Lemma 26.** *We have*

$$d(P_{\hat{L}Z}, \bar{P}_{L\bar{Z}}) \leq \frac{P_{UZ}((u, z) \notin \mathcal{T}_c(\gamma_c))}{2} + d(\bar{P}_{\hat{L}Z}, \bar{P}_{L\bar{Z}}). \quad (182)$$

*Proof:* By the triangular inequality, we have

$$d(P_{\hat{L}Z}, \bar{P}_{L\bar{Z}}) \leq d(P_{\hat{L}Z}, \bar{P}_{\hat{L}Z}) + d(\bar{P}_{\hat{L}Z}, \bar{P}_{L\bar{Z}}). \quad (183)$$

Further, we can bound the first term of the right hand side of the above inequality as

$$d(P_{\hat{L}Z}, \bar{P}_{\hat{L}Z}) = d(P_Z \varphi_C, \bar{P}_Z \varphi_C) \quad (184)$$

$$= d(P_Z, \bar{P}_Z) \quad (185)$$

$$\leq d(P_{UZ}, \bar{P}_{UZ}) \quad (186)$$

$$= \frac{P_{UZ}((u, z) \in \mathcal{T}_c(\gamma_c)^c)}{2} \quad (187)$$

where (185) follows the data-processing inequality (164), (186) follows from the monotonicity property in (163), and (187) follows from (170). ■

**Lemma 27.** *We have*

$$\mathbb{E}_C [d(\bar{P}_{\hat{L}Z}, \bar{P}_{L\bar{Z}})] \leq \frac{\Delta(\gamma_c, P_{UZ})}{2\sqrt{|\mathcal{L}|}}. \quad (188)$$

*Proof:* By noting that the definition of  $\varphi_C$  in (179) implies  $\bar{P}_{L\bar{Z}} = \bar{P}_{\bar{Z}} \varphi_C$ , we have

$$d(\bar{P}_{\hat{L}Z}, \bar{P}_{L\bar{Z}}) = d(\bar{P}_Z \varphi_C, \bar{P}_{\bar{Z}} \varphi_C) \quad (189)$$

$$= d(\bar{P}_Z, \bar{P}_{\bar{Z}}). \quad (190)$$

Then, by taking the expectation with respect to the codebook  $\mathcal{C}$  and by invoking Lemma 25, we have the desired bound. ■

<sup>7</sup>When  $\bar{P}_{\bar{Z}}(z) = 0$ , we define  $\varphi_C(l|z)$  arbitrarily.

APPENDIX D  
PROOF OF THE FIRST NON-ASYMPTOTIC BOUND FOR  
WAK IN THEOREM 5

A. Code Construction

We construct a WAK code by using the stochastic map introduced in Appendix C. Let  $\mathcal{Z} = \mathcal{Y}$  and  $Z = Y$ , that is, let  $P_{UZ} = P_{UY}$ , where  $P_{UY}$  is the marginal of the given distribution  $P_{UXY} \in \mathcal{P}(P_{XY})$ . Also let  $\tilde{Z} = \tilde{Y}$  per (172). It should be noted here that, in this case,  $\mathcal{T}_c(\gamma_c)$  defined in (173) is equivalent to  $\mathcal{T}_c^{\text{WAK}}(\gamma_c)$  defined in (51). Now, let us consider the stochastic map  $\varphi_c$  constructed from the smoothed measure  $\bar{P}_{L\tilde{Y}}$  (cf. (179)).

By using  $\varphi_c$ , we construct a WAK code  $\Phi$  as follows. The main encoder uses a random bin coding  $f: \mathcal{X} \rightarrow \mathcal{M}$ . The helper uses the stochastic map  $\varphi_c: \mathcal{Y} \rightarrow \mathcal{L}$ . That is, when the side information is  $y \in \mathcal{Y}$ , the helper generates  $l \in \mathcal{L}$  according to  $\varphi_c(\cdot|y)$  and sends  $l$  to the decoder. For given  $m \in \mathcal{M}$  and  $l \in \mathcal{L}$ , the decoder outputs the unique  $\hat{x} \in \mathcal{X}$  such that  $f(\hat{x}) = m$  and

$$(u_l, \hat{x}) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b). \quad (191)$$

If no such unique  $\hat{x}$  exists, or if there is more than one such  $\hat{x}$ , then a decoding error is declared.

B. Analysis of Error Probability

Let  $\hat{L}$  be the random index chosen by the helper via the stochastic map  $\varphi_c(\cdot|Y)$ . Note that the joint distribution of  $\hat{L}$  and  $Y$  is given as follows; cf. (180)

$$P_{\hat{L}Y}(l, y) = P_Y(y)\varphi_c(l|y) \quad (192)$$

and then, the joint distribution of  $\hat{L}, Y$  and  $X$  is given as

$$P_{\hat{L}XY}(l, x, y) = P_{\hat{L}Y}(l, y)P_{X|Y}(x|y). \quad (193)$$

The smoothed versions  $\bar{P}_{\hat{L}Y}$  and  $\bar{P}_{\hat{L}XY}$  are given by substituting  $P_Y$  in (192) with  $\bar{P}_Y$ ; cf. (181).

If the decoding error occurs, at least one of the following events occurs:

$$\begin{aligned} \mathcal{E}_1 &:= \{(u_l, x) \notin \mathcal{T}_b^{\text{WAK}}(\gamma_b)\} \\ \mathcal{E}_2 &:= \{\exists \tilde{x} \neq x \text{ s.t. } f(\tilde{x}) = f(x), (u_l, \tilde{x}) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)\} \end{aligned}$$

Hence, the error probability averaged over random coding  $f$  and the random codebook  $\mathcal{C}$  can be bounded as

$$\mathbb{E}_f \mathbb{E}_{\mathcal{C}} [P_e(\Phi)] = \mathbb{E}_f \mathbb{E}_{\mathcal{C}} [P_{\hat{L}XY}(\mathcal{E}_1 \cup \mathcal{E}_2)]. \quad (194)$$

Let

$$\begin{aligned} \mathcal{E}_{12} &:= \left\{ (u, x) : (u, x) \notin \mathcal{T}_b^{\text{WAK}}(\gamma_b) \text{ or } \exists \tilde{x} \neq x \right. \\ &\quad \left. \text{s.t. } f(\tilde{x}) = f(x), (u, \tilde{x}) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b) \right\}. \end{aligned} \quad (195)$$

Then, for fixed  $f$  and  $\mathcal{C}$ , we have

$$\begin{aligned} P_{\hat{L}XY}(\mathcal{E}_1 \cup \mathcal{E}_2) &= P_{\hat{L}XY}((u_l, x) \in \mathcal{E}_{12}) \end{aligned} \quad (196)$$

$$\begin{aligned} &\leq \bar{P}_{LX\tilde{Y}}((u_l, x) \in \mathcal{E}_{12}) + \frac{1 - \bar{P}_{LX\tilde{Y}}(\mathcal{L} \times \mathcal{X} \times \mathcal{Y})}{2} \\ &\quad + d(P_{\hat{L}XY}, \bar{P}_{LX\tilde{Y}}) \end{aligned} \quad (197)$$

$$\begin{aligned} &= \bar{P}_{LX\tilde{Y}}((u_l, x) \in \mathcal{E}_{12}) + \frac{1 - \bar{P}_{LX\tilde{Y}}(\mathcal{L} \times \mathcal{X} \times \mathcal{Y})}{2} \\ &\quad + d(P_{\hat{L}Y}P_{X|Y}, \bar{P}_{L\tilde{Y}}P_{X|Y}) \end{aligned} \quad (198)$$

$$\begin{aligned} &\leq \bar{P}_{LX\tilde{Y}}((u_l, x) \in \mathcal{E}_{12}) + \frac{1 - \bar{P}_{LX\tilde{Y}}(\mathcal{L} \times \mathcal{X} \times \mathcal{Y})}{2} \\ &\quad + d(P_{\hat{L}Y}, \bar{P}_{L\tilde{Y}}) \end{aligned} \quad (199)$$

$$\begin{aligned} &\leq \bar{P}_{LX\tilde{Y}}((u_l, x) \notin \mathcal{T}_b^{\text{WAK}}(\gamma_b)) \\ &\quad + \bar{P}_{LX\tilde{Y}}[\exists \tilde{x} \neq x \text{ s.t. } f(\tilde{x}) = f(x), (u_l, \tilde{x}) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)] \\ &\quad + \frac{1 - \bar{P}_{LX\tilde{Y}}(\mathcal{L} \times \mathcal{X} \times \mathcal{Y})}{2} + d(P_{\hat{L}Y}, \bar{P}_{L\tilde{Y}}) \end{aligned} \quad (200)$$

$$\begin{aligned} &= P_{LX\tilde{Y}}((u_l, x) \notin \mathcal{T}_b^{\text{WAK}}(\gamma_b) \cap (u_l, y) \in \mathcal{T}_c^{\text{WAK}}(\gamma_c)) \\ &\quad + \bar{P}_{LX\tilde{Y}}[\exists \tilde{x} \neq x \text{ s.t. } f(\tilde{x}) = f(x), (u_l, \tilde{x}) \in \mathcal{T}_b(\gamma_b)] \\ &\quad + \frac{1 - \bar{P}_{LX\tilde{Y}}(\mathcal{L} \times \mathcal{X} \times \mathcal{Y})}{2} + d(P_{\hat{L}Y}, \bar{P}_{L\tilde{Y}}) \end{aligned} \quad (201)$$

where (197) follows from (165) for  $\bar{P}_{LX\tilde{Y}} = \bar{P}_{L\tilde{Y}}P_{X|Y}$  in the role of  $Q$ , and (199) follows from the data-processing inequality (164). By taking average over  $\mathcal{C}$ , the first term in (201) is given by

$$\begin{aligned} &\mathbb{E}_{\mathcal{C}} [P_{LX\tilde{Y}}((u_l, x) \notin \mathcal{T}_b^{\text{WAK}}(\gamma_b) \cap (u_l, y) \in \mathcal{T}_c^{\text{WAK}}(\gamma_c))] \\ &= \mathbb{E}_{\mathcal{C}} \left[ \sum_{u,x,y} \sum_l \frac{1}{|\mathcal{L}|} \mathbf{1}[u_l = u] P_{Y|U}(y|u) P_{X|Y}(x|y) \right. \\ &\quad \times \mathbf{1}[(u, x) \notin \mathcal{T}_b^{\text{WAK}}(\gamma_b) \cap (u, y) \in \mathcal{T}_c^{\text{WAK}}(\gamma_c)] \left. \right] \\ &= P_{UXY}((u, x) \notin \mathcal{T}_b^{\text{WAK}}(\gamma_b) \cap (u, y) \in \mathcal{T}_c^{\text{WAK}}(\gamma_c)), \end{aligned} \quad (202)$$

the third term in (201) is given by

$$\begin{aligned} &\mathbb{E}_{\mathcal{C}} [1 - \bar{P}_{LX\tilde{Y}}(\mathcal{L} \times \mathcal{X} \times \mathcal{Y})] \\ &= 1 - \mathbb{E}_{\mathcal{C}} \left[ \sum_{u,x,y} \sum_l \frac{1}{|\mathcal{L}|} \mathbf{1}[u_l = u] P_{Y|U}(y|u) P_{X|Y}(x|y) \right. \\ &\quad \times \mathbf{1}[(u, y) \in \mathcal{T}_c^{\text{WAK}}(\gamma_c)] \left. \right] \end{aligned} \quad (204)$$

$$= P_{UY}((u, y) \notin \mathcal{T}_c^{\text{WAK}}(\gamma_c)), \quad (205)$$

and the fourth term in (201) is upper bounded as

$$\begin{aligned} \mathbb{E}_{\mathcal{C}} [d(P_{\hat{L}Y}, \bar{P}_{L\tilde{Y}})] &\leq \frac{P_{UY}((u, y) \notin \mathcal{T}_c^{\text{WAK}}(\gamma_c))}{2} \\ &\quad + \frac{\Delta(\gamma_c, P_{UY})}{2\sqrt{|\mathcal{L}|}}, \end{aligned} \quad (206)$$

where we used Lemma 26 and Lemma 27. Furthermore, by taking average over  $f$  and  $\mathcal{C}$ , the second term in (201) is upper

bounded as

$$\begin{aligned} & \mathbb{E}_f \mathbb{E}_C \left[ \bar{P}_{LXY} [\exists \tilde{x} \neq x \text{ s.t.} \right. \\ & \quad \left. f(\tilde{x}) = f(x), (u_l, \tilde{x}) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)] \right] \\ &= \mathbb{E}_f \mathbb{E}_C \left[ \sum_{u,x,y} \sum_l \frac{1}{|\mathcal{L}|} \mathbf{1}[u_l = u] \bar{P}_{Y|U}(y|u) P_{X|Y}(x|y) \right. \\ & \quad \left. \times \mathbf{1}[\exists \tilde{x} \neq x \text{ s.t. } f(\tilde{x}) = f(x), (u, \tilde{x}) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)] \right] \end{aligned} \quad (207)$$

$$\begin{aligned} &= \mathbb{E}_f \left[ \sum_{u,x,y} \bar{P}_{UXY}(u, x, y) \right. \\ & \quad \left. \times \mathbf{1}[\exists \tilde{x} \neq x \text{ s.t. } f(\tilde{x}) = f(x), (u, \tilde{x}) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)] \right] \end{aligned} \quad (208)$$

$$\begin{aligned} &\leq \sum_{u,x,y} \bar{P}_{UXY}(u, x, y) \\ & \quad \times \sum_{\tilde{x} \neq x} \mathbb{E}_f [\mathbf{1}[f(\tilde{x}) = f(x)] \mathbf{1}[(u, \tilde{x}) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)]] \end{aligned} \quad (209)$$

$$\leq \frac{1}{|\mathcal{M}|} \sum_u P_U(u) \sum_{\tilde{x}} \mathbf{1}[(u, \tilde{x}) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)] \quad (210)$$

$$= \frac{1}{|\mathcal{M}|} \sum_{(u, \tilde{x}) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)} P_U(u) \quad (211)$$

where we used the fact  $\sum_{x,y} \bar{P}_{UXY}(u, x, y) \leq P_U(u)$  in (210). Hence, by (201), (203), (205), (206), and (211), we have

$$\begin{aligned} & \mathbb{E}_f \mathbb{E}_C [P_e(\Phi)] \\ &= \mathbb{E}_f \mathbb{E}_C [P_{K\hat{L}\hat{U}XY}(\mathcal{E}_1 \cup \mathcal{E}_2)] \end{aligned} \quad (212)$$

$$\begin{aligned} &\leq P_{UXY}((u, x) \notin \mathcal{T}_b^{\text{WAK}}(\gamma_b) \cup (u, y) \notin \mathcal{T}_c^{\text{WAK}}(\gamma_c)) \\ & \quad + \frac{\Delta(\gamma_c, P_{UY})}{2\sqrt{|\mathcal{L}|}} + \frac{1}{|\mathcal{M}|} \sum_{(u, \tilde{x}) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b)} P_U(u). \end{aligned} \quad (213)$$

Consequently, there exists at least one code  $(f, \mathcal{C})$  such that  $P_e(\Phi)$  is smaller than the right-hand-side of the inequality above. This completes the proof of Theorem 5.

#### APPENDIX E

##### PROOF OF THE SECOND NON-ASYMPTOTIC BOUND FOR WAK IN THEOREM 7

To prove Theorem 7, we modify the proof of Theorem 5 as follows. Since the analysis of error can be done in a similar manner as Appendix D, we only show the code construction.

First, we use  $\mathcal{J} = \{1, \dots, J\}$  instead of  $\mathcal{L}$  in the construction of  $\varphi_C$ , where  $J$  is the given integer. Then, the helper and the decoder are modified as follows. The helper first uses the stochastic map  $\varphi_C: \mathcal{Y} \rightarrow \mathcal{J}$ . That is, it generates  $j \in \mathcal{J}$  according to  $\varphi_C(\cdot|y)$  when the side information is  $y \in \mathcal{Y}$ . Then, the helper sends  $j$  by using random bin coding  $\kappa: \mathcal{J} \rightarrow \mathcal{L}$ . This means that to every  $j \in \mathcal{J}$ , it independently and uniformly assigns a random index  $l \in \mathcal{L}$ . For given  $m \in \mathcal{M}$  and  $l \in \mathcal{L}$ , the decoder outputs the unique  $\hat{x} \in \mathcal{X}$  such that  $f(\hat{x}) = m$  and

$$(u_j, \hat{x}) \in \mathcal{T}_b^{\text{WAK}}(\gamma_b) \quad (214)$$

for some  $j \in \mathcal{J}$  satisfying  $\kappa(j) = l$ . If no such unique  $\hat{x}$  exists, or if there is more than one such  $\hat{x}$ , then a decoding error is declared.

#### APPENDIX F

##### PROOF OF THE NON-ASYMPTOTIC BOUND FOR WZ IN THEOREM 8

###### A. Code Construction

Similar to WAK coding in the previous two sections, we use the stochastic map introduced in Appendix C. Also, the proof is rather similar to the WAK one so we just highlight the key steps, pointing the reader to various points of Appendix D for the details of the calculations.

In WZ coding, let  $\mathcal{Z} = \mathcal{X}$  and  $P_{UZ} = P_{UX}$ . Also let  $\tilde{Z} = \tilde{X}$  per (172). Note that  $\mathcal{T}_c(\gamma_c)$  defined in (173) is equivalent to  $\mathcal{T}_c^{\text{WZ}}(\gamma_c)$  defined in (62). Now, let us consider the stochastic map  $\varphi_C$  defined in (179).

By using  $\varphi_C$ , we construct a WZ code  $\Phi$  as follows. The encoder first uses the stochastic map  $\varphi_C: \mathcal{X} \rightarrow \mathcal{L}$ . That is, it generates  $l \in \mathcal{L}$  according to  $\varphi_C(\cdot|x)$  when the source output is  $x \in \mathcal{X}$ . Then, the encoder sends  $l$  by using random bin coding  $\kappa: \mathcal{L} \rightarrow \mathcal{M}$ . This means that to every  $l \in \mathcal{L}$ , it independently and uniformly assigns a random index  $m \in \mathcal{M}$ . For given  $m \in \mathcal{M}$ ,  $y \in \mathcal{Y}$ , the decoder finds the unique index  $l \in \mathcal{L}$  such that  $\kappa(l) = m$  and

$$(u_l, y) \in \mathcal{T}_p^{\text{WZ}}(\gamma_p). \quad (215)$$

Then, decoder outputs  $\hat{x} \in \hat{\mathcal{X}}$  according to  $P_{\hat{X}|UY}(\cdot|u_l, y)$ . We assume that we use the *stochastic* reproduction function  $P_{\hat{X}|UY}$  throughout. If the *deterministic* reproduction function  $g: \mathcal{U} \times \mathcal{Y} \rightarrow \hat{\mathcal{X}}$  is used, the decoder outputs  $\hat{x} = g(u_l, y)$ . If no unique  $l$  satisfying (215) exists, or if there is more than one such  $l$  satisfying (215), then a decoding error is declared.

###### B. Analysis of Probability of Excess Distortion

Let  $\hat{L}$  be the random index chosen by the encoder via the stochastic map  $\varphi_C(\cdot|X)$ . Note that the joint distribution of  $\hat{L}, X$  is given as follows; cf. (180)

$$P_{\hat{L}X}(l, x) = P_X(x) \varphi_C(l|x). \quad (216)$$

Next, the joint distribution of  $\hat{L}, X, Y, \hat{X}$  is given as

$$P_{\hat{L}XY\hat{X}}(l, x, y, \hat{x}) = P_{\hat{L}X}(l, x) P_{Y|X}(y|x) P_{\hat{X}|UY}(\hat{x}|u_l, y). \quad (217)$$

The smoothed versions  $\bar{P}_{\hat{L}X}$  and  $\bar{P}_{\hat{L}XY\hat{X}}$  are given by substituting  $P_X$  in (216) with  $\bar{P}_X$ ; cf. (181).

If the distortion exceeds  $D$ , at least one of the following events occurs:

$$\mathcal{E}_0 := \{(x, \hat{x}) \notin \mathcal{T}_{d, \text{st}}^{\text{WZ}}(D)\} \quad (218)$$

$$\mathcal{E}_1 := \{(u_l, y) \notin \mathcal{T}_p^{\text{WZ}}(\gamma_p)\} \quad (219)$$

$$\mathcal{E}_2 := \left\{ \exists \tilde{l} \neq l \text{ s.t. } \kappa(\tilde{l}) = \kappa(l), (u_{\tilde{l}}, y) \in \mathcal{T}_p^{\text{WZ}}(\gamma_p) \right\}. \quad (220)$$

Hence, the probability of excess distortion averaged over the random coding  $\kappa$  and the random codebook  $\mathcal{C}$  can be bounded

as

$$\begin{aligned} & \mathbb{E}_\kappa \mathbb{E}_\mathcal{C} [\mathbb{P}_e(\Phi; D)] \\ & \leq \mathbb{E}_\kappa \mathbb{E}_\mathcal{C} [P_{\hat{L}XY\hat{X}}(\mathcal{E}_0 \cup \mathcal{E}_1 \cup \mathcal{E}_2)] \end{aligned} \quad (221)$$

$$\leq \mathbb{E}_\mathcal{C} [P_{\hat{L}XY\hat{X}}(\mathcal{E}_0 \cup \mathcal{E}_1)] + \mathbb{E}_\kappa \mathbb{E}_\mathcal{C} [P_{\hat{L}XY}(\mathcal{E}_2)]. \quad (222)$$

At first, we evaluate the first term in (222). For fixed  $\mathcal{C}$ ,

$$\begin{aligned} & P_{\hat{L}XY\hat{X}}(\mathcal{E}_0 \cup \mathcal{E}_1) \\ & \leq \bar{P}_{\hat{L}\hat{X}Y\hat{X}}(\mathcal{E}_0 \cup \mathcal{E}_1) + \frac{1 - \bar{P}_{\hat{L}\hat{X}Y\hat{X}}(\mathcal{L} \times \mathcal{X} \times \mathcal{Y} \times \hat{\mathcal{X}})}{2} \\ & \quad + d(P_{\hat{L}XY\hat{X}}, \bar{P}_{\hat{L}\hat{X}Y\hat{X}}) \end{aligned} \quad (223)$$

$$\begin{aligned} & \leq \bar{P}_{\hat{L}\hat{X}Y\hat{X}}(\mathcal{E}_0 \cup \mathcal{E}_1) + \frac{1 - \bar{P}_{\hat{L}\hat{X}Y\hat{X}}(\mathcal{L} \times \mathcal{X} \times \mathcal{Y} \times \hat{\mathcal{X}})}{2} \\ & \quad + d(P_{\hat{L}X}, \bar{P}_{\hat{L}\hat{X}}) \end{aligned} \quad (224)$$

$$\begin{aligned} & = P_{\hat{L}XY\hat{X}}((u_l, x) \in \mathcal{T}_c^{\text{WZ}}(\gamma_c) \cup (x, \hat{x}) \notin \mathcal{T}_{\text{d,st}}^{\text{WZ}}(D) \\ & \quad \cup (u_l, y) \notin \mathcal{T}_p^{\text{WZ}}(\gamma_p)) \\ & \quad + \frac{1 - \bar{P}_{\hat{L}\hat{X}Y\hat{X}}(\mathcal{L} \times \mathcal{X} \times \mathcal{Y} \times \hat{\mathcal{X}})}{2} + d(P_{\hat{L}X}, \bar{P}_{\hat{L}\hat{X}}) \end{aligned} \quad (225)$$

where (223) follows from (165), (224) follows from the same reasoning that led to (199) and (225) from the same reasoning that led to (201).

By the same reasoning that led to (203) for the WAK problem, the expectation of the first term in (225) can be expressed as

$$\begin{aligned} & \mathbb{E}_\mathcal{C} [P_{\hat{L}XY\hat{X}}((u_l, x) \in \mathcal{T}_c^{\text{WZ}}(\gamma_c) \cup (x, \hat{x}) \notin \mathcal{T}_{\text{d,st}}^{\text{WZ}}(D) \\ & \quad \cup (u, y) \notin \mathcal{T}_p^{\text{WZ}}(\gamma_p))] \end{aligned} \quad (226)$$

$$\begin{aligned} & = P_{UXY\hat{X}}((u, x) \in \mathcal{T}_c^{\text{WZ}}(\gamma_c) \cup (x, \hat{x}) \notin \mathcal{T}_{\text{d,st}}^{\text{WZ}}(D) \\ & \quad \cup (u, y) \notin \mathcal{T}_p^{\text{WZ}}(\gamma_p)) \end{aligned} \quad (227)$$

By the same reasoning that led to (205) for the WAK problem, the expectation of the second term in (225) can be evaluated as

$$\mathbb{E}_\mathcal{C} [1 - \bar{P}_{\hat{L}\hat{X}Y\hat{X}}(\mathcal{L} \times \mathcal{X} \times \mathcal{Y} \times \hat{\mathcal{X}})] = P_{UX}((u, x) \notin \mathcal{T}_c^{\text{WZ}}(\gamma_c)). \quad (228)$$

Similarly to (206) for the WAK problem, the expectation of the third term in (225) can be bounded as

$$\mathbb{E}_\mathcal{C} [d(P_{\hat{L}X}, \bar{P}_{\hat{L}\hat{X}})] \leq \frac{P_{UX}((u, x) \notin \mathcal{T}_c^{\text{WZ}}(\gamma_c))}{2} + \frac{\Delta(\gamma_c, P_{UX})}{2\sqrt{|\mathcal{L}|}}. \quad (229)$$

Now we bound the final term in (222) using steps similar to

the ones leading to (211) for the WAK problem. We have

$$\begin{aligned} & \mathbb{E}_\kappa \mathbb{E}_\mathcal{C} [P_{\hat{L}XY}(\mathcal{E}_2)] \\ & = \mathbb{E}_\kappa \mathbb{E}_\mathcal{C} \left[ \sum_{u,x,y,l} \frac{1}{|\mathcal{L}|} \mathbf{1}[u_l = u] P_{\hat{L}XYU}(l, x, y, u) \mathbf{1}[\exists \tilde{l} \neq l \right. \\ & \quad \left. \text{s.t. } \kappa(\tilde{l}) = \kappa(l), (u_{\tilde{l}}, y) \in \mathcal{T}_p^{\text{WZ}}(\gamma_p)] \right] \end{aligned} \quad (230)$$

$$\begin{aligned} & \leq \mathbb{E}_\kappa \mathbb{E}_\mathcal{C} \left[ \sum_{u,x,y,l} \frac{1}{|\mathcal{L}|} \mathbf{1}[u_l = u] P_{\hat{L}XYU}(l, x, y, u) \right. \\ & \quad \left. \sum_{\tilde{l} \neq l} \mathbf{1}[\kappa(\tilde{l}) = \kappa(l)] \cdot \mathbf{1}[(u_{\tilde{l}}, y) \in \mathcal{T}_p^{\text{WZ}}(\gamma_p)] \right] \end{aligned} \quad (231)$$

$$\begin{aligned} & \leq \frac{1}{|\mathcal{M}|} \mathbb{E}_\mathcal{C} \left[ \sum_{u,x,y,l} \frac{1}{|\mathcal{L}|} \mathbf{1}[u_l = u] P_{\hat{L}XYU}(l, x, y, u) \right. \\ & \quad \left. \sum_{\tilde{l} \neq l} \mathbf{1}[(u_{\tilde{l}}, y) \in \mathcal{T}_p^{\text{WZ}}(\gamma_p)] \right] \end{aligned} \quad (232)$$

$$\leq \frac{|\mathcal{L}|}{|\mathcal{M}|} \sum_{u,y} P_U(u) P_Y(y) \mathbf{1}[(u, y) \in \mathcal{T}_p^{\text{WZ}}(\gamma_p)] \quad (233)$$

$$= \frac{|\mathcal{L}|}{|\mathcal{M}|} \sum_{(u,y) \in \mathcal{T}_p^{\text{WZ}}(\gamma_p)} P_U(u) P_Y(y). \quad (234)$$

By uniting (222), (225), (227), (228), (229) and (234), we obtain the final bound

$$\begin{aligned} & \mathbb{E}_\kappa \mathbb{E}_\mathcal{C} [\mathbb{P}_e(\Phi; D)] \\ & \leq P_{UXY\hat{X}}((u, x) \notin \mathcal{T}_c^{\text{WZ}}(\gamma_c) \cup (x, \hat{x}) \notin \mathcal{T}_{\text{d,st}}^{\text{WZ}}(D) \\ & \quad \cup (u, y) \notin \mathcal{T}_p^{\text{WZ}}(\gamma_p)) \\ & \quad + \frac{\Delta(\gamma_c, P_{UX})}{2\sqrt{|\mathcal{L}|}} + \frac{|\mathcal{L}|}{|\mathcal{M}|} \sum_{(u,y) \in \mathcal{T}_p^{\text{WZ}}(\gamma_p)} P_U(u) P_Y(y). \end{aligned} \quad (235)$$

This implies there is a deterministic code whose probability of excess distortion is no greater than the right-hand-side of (235). This completes the proof of Theorem 8.

## APPENDIX G

### PROOF OF THE NON-ASYMPTOTIC BOUND FOR GP IN THEOREM 10

Since the analysis of error probability can be done in an almost similar manner as those of WAK and WZ, we only show the code construction for GP.

#### A. Code Construction

As in WAK, we use the stochastic map introduced in Appendix C. In GP coding, let  $\mathcal{Z} = \mathcal{S}$  and  $P_{UZ} = P_{US}$ . Note that that,  $\mathcal{T}_c(\gamma_c)$  defined in (173) is equivalent to  $\mathcal{T}_c^{\text{GP}}(\gamma_c)$  defined in (69) in this case.

For GP coding, we construct  $|\mathcal{M}|$  stochastic maps. Each stochastic map corresponds to a message in  $\mathcal{M}$ . For each message  $m \in \mathcal{M}$ , generate a codebook  $\mathcal{C}^{(m)} = \{u_1^{(m)}, \dots, u_{|\mathcal{L}|}^{(m)}\}$  where each  $u_l^{(m)}$  is independently drawn according to  $P_U$ . Then, for each  $\mathcal{C}^{(m)}$  ( $m \in \mathcal{M}$ ), construct a stochastic map  $\varphi_{\mathcal{C}^{(m)}}$  as defined in (179).



By using  $\{\varphi_{\mathcal{C}^{(m)}}\}_{m \in \mathcal{M}}$ , we construct a GP code  $\Phi$  as follows. Given the message  $m \in \mathcal{M}$  and the channel state  $s \in \mathcal{S}$ , the encoder first generates  $l \in \mathcal{L}$  according to  $\varphi_{\mathcal{C}^{(m)}}(\cdot|s)$ . Then, the encoder generates  $x \in \mathcal{X}$  according to  $P_{X|US}(\cdot|u_l^{(m)}, s)$  and inputs  $x$  into the channel. If the randomly generated  $x$  results in  $g(x) > \Gamma$  (i.e., the channel input does not satisfy the cost constraint), declare a cost-constraint violation error.<sup>8</sup> Given the channel output  $y \in \mathcal{Y}$ , the decoder finds the unique index  $\hat{m} \in \mathcal{M}$  such that

$$(u_l^{(\hat{m})}, y) \in \mathcal{T}_p^{\text{GP}}(\gamma_p) \quad (236)$$

for some  $l \in \mathcal{L}$ . If there is no unique index  $\hat{m} \in \mathcal{M}$  or more than one, declare a decoding error. This is a Feinstein-like decoder [7] for average probability of error. If no such unique  $\hat{m}$  exists, or if there exists more than one such  $\hat{m}$ , then a decoding error is declared.

## APPENDIX H

### PRELIMINARIES FOR PROOFS OF THE SECOND-ORDER CODING RATE

In this appendix, we provide some technical results that will be used in Appendices I and K. More specifically, we will use the following multidimensional Berry-Esséen theorem and its corollary.

**Theorem 28** (Göetze [21]). *Let  $\mathbf{U}_1, \dots, \mathbf{U}_n$  be independent random vectors in  $\mathbb{R}^k$  with zero mean. Let  $\mathbf{S}_n = \frac{1}{\sqrt{n}}(\mathbf{U}_1 + \dots + \mathbf{U}_n)$ ,  $\text{Cov}(\mathbf{S}_n) = \mathbf{I}$ , and  $\xi = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\|\mathbf{U}_i\|_2^3]$ . Let the standard Gaussian random vector  $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ . Then, for all  $n \in \mathbb{N}$ , we have*

$$\sup_{\mathcal{C} \in \mathcal{C}_k} |\Pr\{\mathbf{S}_n \in \mathcal{C}\} - \Pr\{\mathbf{Z} \in \mathcal{C}\}| \leq \frac{C_k \xi}{\sqrt{n}}, \quad (237)$$

where  $\mathcal{C}_k$  is the family of all convex, Borel measurable subsets of  $\mathbb{R}^k$ , and where  $C_k$  is a constant that depends on the dimension  $k$ .

It should be noted that Theorem 28 can be applied for random vectors that are independent but not necessarily identical.

We will frequently encounter random vectors with non-identity covariance matrices. Thus, we slightly modify Theorem 28 in a similar manner as [23, Corollary 7] as follows.

**Corollary 29.** *Let  $\mathbf{U}_1, \dots, \mathbf{U}_n$  be independent random vectors in  $\mathbb{R}^k$  with zero mean. Let  $\mathbf{S}_n = \frac{1}{\sqrt{n}}(\mathbf{U}_1 + \dots + \mathbf{U}_n)$ ,  $\text{Cov}(\mathbf{S}_n) = \mathbf{V} \succ 0$ , and  $\xi = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\|\mathbf{U}_i\|_2^3]$ . Let the Gaussian random vector  $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{V})$ . Then, for all  $n \in \mathbb{N}$ ,*

$$\sup_{\mathcal{C} \in \mathcal{C}_k} |\Pr\{\mathbf{S}_n \in \mathcal{C}\} - \Pr\{\mathbf{Z} \in \mathcal{C}\}| \leq \frac{C_k \xi}{\lambda_{\min}(\mathbf{V})^{3/2} \sqrt{n}}, \quad (238)$$

where  $\mathcal{C}_k$  is the family of all convex, Borel measurable subsets of  $\mathbb{R}^k$ , where  $C_k$  is a constant that depends on the dimension  $k$ , and where  $\lambda_{\min}(\mathbf{V})$  is the smallest eigenvalue of  $\mathbf{V}$ .

<sup>8</sup>Even if  $g(x) > \Gamma$  occurs, we still send  $x$  through the channel. The error event for this occurrence must be taken into account in the error analysis.

## APPENDIX I

### ACHIEVABILITY PROOF OF THE SECOND-ORDER CODING RATE FOR WAK IN THEOREM 15

*Proof:* It suffices to show the inclusion  $\mathcal{R}_{\text{in}}(n, \varepsilon; P_{UTXY}) \subset \mathcal{R}_{\text{WAK}}(n, \varepsilon)$  for fixed  $P_{UTXY} \in \tilde{\mathcal{P}}(P_{XY})$ .

We first consider the case such that  $\mathbf{V} = \mathbf{V}(P_{UTXY}) \succ 0$ . First, note that  $\mathbf{R} \in \mathcal{R}_{\text{in}}(n, \varepsilon; P_{UTXY})$  implies

$$\tilde{\mathbf{z}} := \sqrt{n} \left( \mathbf{R} - \mathbf{J} - \frac{2 \log n}{n} \mathbf{1}_2 \right) \in \mathcal{S}(\mathbf{V}, \varepsilon). \quad (239)$$

We fix a time-sharing sequence  $t^n \in \mathcal{T}^n$  with type  $P_{t^n} \in \mathcal{P}_n(\mathcal{T})$  such that

$$|P_{t^n}(t) - P_T(t)| \leq \frac{1}{n} \quad (240)$$

for every  $t \in \mathcal{T}$  [42]. Then, we consider the test channel given by  $P_{U^n|Y^n}(u^n|y^n) = P_{U|TY}^n(u^n|t^n, y^n)$ , and we use Corollary 6 for  $P_{U^n X^n Y^n} = P_{XY}^n P_{U^n|Y^n}$  by setting  $\gamma_b = \log |\mathcal{M}_n| - \log n$ ,  $\gamma_c = \log |\mathcal{L}_n| - \log n$ , and  $\delta = \frac{1}{n}$ . Then, there exists a WAK code  $\Phi_n$  such that

$$\begin{aligned} 1 - P_e(\Phi_n) &\geq \Pr \left\{ \sum_{i=1}^n \mathbf{j}(U_i, X_i, Y_i|t_i) \leq n\mathbf{R} - \log n \mathbf{1}_2 \right\} - \frac{2}{n} - \sqrt{\frac{1}{n}} \\ &= \Pr \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^n (\mathbf{j}(U_i, X_i, Y_i|t_i) - \mathbf{J}) \leq \tilde{\mathbf{z}} + \frac{\log n}{\sqrt{n}} \mathbf{1}_2 \right\} - \frac{2}{n} - \sqrt{\frac{1}{n}}. \end{aligned} \quad (241)$$

$$= \Pr \left\{ \mathbf{Z} \leq \tilde{\mathbf{z}} + \frac{\log n}{\sqrt{n}} \mathbf{1}_2 \right\} - \frac{2}{n} - \sqrt{\frac{1}{n}}. \quad (242)$$

By using Corollary 29 to the first term of (242), we have

$$1 - P_e(\Phi_n) \geq \Pr \left\{ \mathbf{Z} \leq \tilde{\mathbf{z}} + \frac{\log n}{\sqrt{n}} \mathbf{1}_2 \right\} - O \left( \frac{1}{\sqrt{n}} \right) \quad (243)$$

$$= \Pr\{\mathbf{Z} \leq \tilde{\mathbf{z}}\} + O \left( \frac{\log n}{\sqrt{n}} \right) \quad (244)$$

$$\geq 1 - \varepsilon \quad (245)$$

for sufficiently large  $n$ , where (244) follows from the Taylor's approximation, and (245) follows from (239).

Next, we consider the case with  $\mathbf{V}$  is singular but not 0. In this case, we cannot apply Corollary 29 because  $\lambda_{\min}(\mathbf{V}) = 0$ . Since  $\text{rank}(\mathbf{V}) = 1$ , we can write  $\mathbf{V} = \mathbf{v}\mathbf{v}^T$  by using the vector  $\mathbf{v}$ . Let  $\mathbf{A}_i = \mathbf{j}(U_i, X_i, Y_i|t_i) - \mathbf{J}$ . Then we can write  $\mathbf{A}_i = \mathbf{v}B_i$  by using the scalar independent random variables  $\{B_i\}_{i=1}^n$ . Thus, by using the ordinary Berry-Esséen theorem [64, Ch. XVI] for  $\{B_i\}_{i=1}^n$ , we can derive (245).

Finally, we consider the case where  $\mathbf{V} = \mathbf{0}$ . In this case, by setting  $\tilde{\mathbf{z}} = \mathbf{0}$  in (242), we can find that the right hand side converges to 1.

For the bounds on the cardinalities of auxiliary random variables, see Appendix M. ■

## APPENDIX J

## ACHIEVABILITY PROOF OF THE SECOND-ORDER CODING RATE FOR WAK IN THEOREM 16

*Proof:* We only provide a sketch of the proof because most of the steps are the same as Appendix I. The only modification is that we use Theorem 7 instead of Corollary 6 by setting  $\gamma_b = \log |\mathcal{M}_n| - \rho\sqrt{n} - \log n$ ,  $\gamma_c = \log |\mathcal{L}_n| + \rho\sqrt{n} - \log n$ ,  $J_n = |\mathcal{L}_n|2^{\rho\sqrt{n}}$ , and  $\delta = \frac{1}{n}$ . ■

## APPENDIX K

## ACHIEVABILITY PROOF OF THE SECOND-ORDER CODING RATE FOR WZ IN THEOREM 18

*Proof:* It suffices to show the inclusion  $\mathcal{R}_{\text{in}}(n, \varepsilon; P_{UTXY}, P_{\hat{X}|UYT}) \subset \mathcal{R}_{\text{WZ}}(n, \varepsilon)$  for fixed pair  $(P_{UTXY}, P_{\hat{X}|UYT})$  of  $P_{UTXY} \in \tilde{\mathcal{P}}(P_{XY})$  and  $P_{\hat{X}|UYT}$ . We assume that  $\mathbf{V} = \mathbf{V}(P_{UTXY}, P_{\hat{X}|UYT}) \succ 0$ , since the case where  $\mathbf{V}$  is singular can be handled in a similar manner as Appendix I (see also [23, Proof of Theorem 5]).

First, note that  $[R, D]^T \in \mathcal{R}_{\text{in}}(n, \varepsilon; P_{UTXY}, P_{\hat{X}|UYT})$  implies

$$\tilde{\mathbf{z}} := \sqrt{n} \left( \begin{bmatrix} -\frac{1}{n} \log \frac{L_n}{|\mathcal{M}_n|} \\ \frac{1}{n} \log L_n \\ D \end{bmatrix} - \mathbf{J} - \frac{2 \log n}{n} \mathbf{1}_3 \right) \in \mathcal{S}(\mathbf{V}, \varepsilon) \quad (246)$$

for some positive integer  $L_n$ . We fix a sequence  $t^n \in \mathcal{T}^n$  satisfying (240) for every  $t \in \mathcal{T}$ . Then, we consider the test channel given by  $P_{U^n|X^n}(u^n|x^n) = P_{U|TX}^n(u^n|t^n, x^n)$  and the reproduction channel given by  $P_{\hat{X}^n|U^n Y^n}(\hat{x}^n|u^n, y^n) = P_{\hat{X}|UYT}^n(\hat{x}^n|u^n, t^n)$ . Then, Corollary 9 for  $P_{U^n X^n Y^n \hat{X}^n} = P_{XY}^n P_{U^n|X^n} P_{\hat{X}^n|U^n Y^n}$  with  $\gamma_p = \log \frac{L_n}{|\mathcal{M}_n|} + \log n$ ,  $\gamma_c = \log L_n - \log n$ , and  $\delta = \frac{1}{n}$  shows that there exists a WZ code such that

$$1 - P_e(\Phi_n; D) \geq \Pr \left\{ \sum_{i=1}^n \mathbf{j}(U_i, X_i, Y_i, \hat{X}_i | t_i) \leq \begin{bmatrix} -\log \frac{L_n}{|\mathcal{M}_n|} \\ \log L_n \\ nD \end{bmatrix} - \log n \mathbf{1}_3 \right\} - \frac{2}{n} - \sqrt{\frac{1}{n}} \quad (247)$$

$$= \Pr \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^n \left( \mathbf{j}(U_i, X_i, Y_i, \hat{X}_i | t_i) - \mathbf{J} \right) \leq \tilde{\mathbf{z}} + \frac{\log n}{\sqrt{n}} \mathbf{1}_3 \right\} - \frac{2}{n} - \sqrt{\frac{1}{n}}. \quad (248)$$

Now the rest of the proof proceeds by using the multidimensional Berry-Esséen theorem as in (243) to (245) for the WAK problem.

For the bounds on the cardinalities of auxiliary random variables, see Appendix M. ■

## APPENDIX L

## ACHIEVABILITY PROOF OF THE SECOND-ORDER CODING RATE FOR LOSSY SOURCE CODING IN THEOREM 20

We slightly modify a special case of Corollary 9 as follows, which will be used in both Appendices L-A and L-B.

**Corollary 30.** For arbitrary distribution  $Q_{\hat{X}} \in \mathcal{P}(\hat{\mathcal{X}})$ , and for arbitrary constants  $\gamma_c, \nu \geq 0$  and  $\delta, \tilde{\delta} > 0$ , there exists a lossy source code  $\Phi$  with probability of excess distortion satisfying

$$P_e(\Phi; D) \leq P_{\hat{X}X} \left[ \log \frac{P_{\hat{X}|X}(\hat{x}|x)}{Q_{\hat{X}}(\hat{x})} > \gamma_c - \nu \text{ or } d(x, \hat{x}) > D \right] + \tilde{\delta} + \sqrt{\frac{2^{\gamma_c}}{\tilde{\delta}|\mathcal{M}|}} + \delta + 2^{-\nu}. \quad (249)$$

*Proof:* As a special case of Corollary 9, we have

$$P_e(\Phi; D) \leq P_{\hat{X}X} \left[ \log \frac{P_{\hat{X}|X}(\hat{x}|x)}{P_{\hat{X}}(\hat{x})} > \gamma_c \text{ or } d(x, \hat{x}) > D \right] + \tilde{\delta} + \sqrt{\frac{2^{\gamma_c}}{\tilde{\delta}|\mathcal{M}|}} + \delta, \quad (250)$$

where we set  $\gamma_p = 0$  and  $L = \tilde{\delta}|\mathcal{M}|$ . We can further upper bound the first term of (250) as

$$P_{\hat{X}X} \left[ \log \frac{P_{\hat{X}|X}(\hat{x}|x)}{P_{\hat{X}}(\hat{x})} > \gamma_c \text{ or } d(x, \hat{x}) > D \right] \quad (251)$$

$$= P_{\hat{X}X} \left[ \log \frac{P_{\hat{X}|X}(\hat{x}|x)}{Q_{\hat{X}}(\hat{x})} + \log \frac{Q_{\hat{X}}(\hat{x})}{P_{\hat{X}}(\hat{x})} > \gamma_c \text{ or } d(x, \hat{x}) > D \right] \quad (252)$$

$$\leq P_{\hat{X}X} \left[ \log \frac{P_{\hat{X}|X}(\hat{x}|x)}{Q_{\hat{X}}(\hat{x})} > \gamma_c - \nu \text{ or } \log \frac{Q_{\hat{X}}(\hat{x})}{P_{\hat{X}}(\hat{x})} > \nu \text{ or } d(x, \hat{x}) > D \right] \quad (253)$$

$$\leq P_{\hat{X}X} \left[ \log \frac{P_{\hat{X}|X}(\hat{x}|x)}{Q_{\hat{X}}(\hat{x})} > \gamma_c - \nu \text{ or } d(x, \hat{x}) > D \right] + P_{\hat{X}X} \left[ \log \frac{Q_{\hat{X}}(\hat{x})}{P_{\hat{X}}(\hat{x})} > \nu \right] \quad (254)$$

$$= P_{\hat{X}X} \left[ \log \frac{P_{\hat{X}|X}(\hat{x}|x)}{Q_{\hat{X}}(\hat{x})} > \gamma_c - \nu \text{ or } d(x, \hat{x}) > D \right] + P_{\hat{X}} \left[ \log \frac{Q_{\hat{X}}(\hat{x})}{P_{\hat{X}}(\hat{x})} > \nu \right] \quad (255)$$

$$\leq P_{\hat{X}X} \left[ \log \frac{P_{\hat{X}|X}(\hat{x}|x)}{Q_{\hat{X}}(\hat{x})} > \gamma_c - \nu \text{ or } d(x, \hat{x}) > D \right] + 2^{-\nu}. \quad (256)$$

This completes the proof. ■

**Remark 9.** By showing Corollary 30 directly instead of via Corollary 9, we can eliminate the residual term  $\tilde{\delta}$ .

## A. Proof Based on the Method of Types

To prove Theorem 20 by the method of types, we use the following lemma.

**Lemma 31** (Rate-Redundancy [25]). Suppose that  $R(P_X, D)$  is differentiable w.r.t.  $D$  and twice differentiable w.r.t.  $P_X$  at

some neighbourhood of  $(P_X, D)$ . Let  $\varepsilon$  be given probability and let  $\Delta R$  be any quantity chosen such that

$$P_X^n [R(P_{x^n}, D) - R(P_X, D) > \Delta R] = \varepsilon + g_n, \quad (257)$$

where  $g_n = O\left(\frac{\log n}{\sqrt{n}}\right)$ . Then, as  $n$  grows,

$$\Delta R = \sqrt{\frac{\text{Var}(j(X, D))}{n}} Q^{-1}(\varepsilon) + O\left(\frac{\log n}{n}\right). \quad (258)$$

Note that the quantity  $j(x, D)$  has an alternative representation as the derivative of  $Q \mapsto R(Q, D)$  with respect to  $Q(x)$  evaluated at  $P_X(x)$ ; cf. (134).

We also use the following lemma, which is a consequence of the argument right after [65, Theorem 1].

**Lemma 32.** For a type  $q \in \mathcal{P}_n(\mathcal{X})$ , suppose that  $\left|\frac{\partial R(q, D)}{\partial D}\right| < C$  for a constant  $C > 0$  in some neighbourhood of  $q$ . Then, there exists a test channel  $V \in \mathcal{V}_n(\mathcal{Y}; q)$  such that

$$\sum_{x, \hat{x}} q(x) V(\hat{x}|x) d(x, \hat{x}) \leq D \quad (259)$$

and

$$I(q, V) \leq R(q, D) + \frac{\tau}{n}, \quad (260)$$

where  $\tau$  is a constant depending on  $C$ ,  $|\mathcal{X}|$ ,  $|\hat{\mathcal{X}}|$ , and  $D_{\max}$ .

Using Lemmas 31 and 32, we prove Theorem 20.

*Proof:* We construct a test channel  $P_{\hat{X}^n|X^n}$  as follows. For a fixed constant  $\tilde{\tau} > 0$ , we set

$$\Omega_n = \left\{ q \in \mathcal{P}_n(\mathcal{X}) : \|P_x - q\|^2 \leq \frac{\tilde{\tau} \log n}{n} \right\}. \quad (261)$$

Since we assumed that  $R(P_X, D)$  is differentiable w.r.t.  $D$  at  $P_X$ , the derivative is bounded over any small enough neighbourhood of  $P_X$ . In particular, it is bounded by some constant  $C$  over  $\Omega_n$  for sufficiently large  $n$ . For each  $q \in \Omega_n$ , we choose test channel  $V_q \in \mathcal{V}_n(\mathcal{Y}; q)$  satisfying the statement of Lemma 32. Then, we define the test channel

$$P_{\hat{X}^n|X^n}(\hat{x}^n|x^n) = \begin{cases} \frac{1}{|\mathcal{T}_{V_{P_{x^n}}}(x^n)|} & \text{if } \hat{x}^n \in \mathcal{T}_{V_{P_{x^n}}}(x^n) \\ 0 & \text{else} \end{cases} \quad (262)$$

for  $x^n$  satisfying  $P_{x^n} \in \Omega_n$ , and otherwise we define  $P_{\hat{X}^n|X^n}(\hat{x}^n|x^n)$  arbitrarily as long as the channel only outputs  $\hat{x}^n$  satisfying  $d_n(x^n, \hat{x}^n) \leq D$ . Let  $P_q \in \mathcal{P}_n(\hat{\mathcal{X}})$  be such that

$$P_q(\hat{x}) = \sum_x q(x) V_q(\hat{x}|x). \quad (263)$$

Then, let  $\tilde{P}_q^n \in \mathcal{P}(\hat{\mathcal{X}}^n)$  be the uniform distribution on  $\mathcal{T}_{P_q}$ . Furthermore, let  $Q_{\hat{X}^n} \in \mathcal{P}(\hat{\mathcal{X}}^n)$  be the distribution given by

$$Q_{\hat{X}^n}(\hat{x}^n) = \sum_{q \in \Omega_n} \frac{1}{|\Omega_n|} \tilde{P}_q^n(\hat{x}^n). \quad (264)$$

We now use Corollary 30 for  $P_X = P_X^n$ ,  $P_{\hat{X}|X} = P_{\hat{X}^n|X^n}$ , and  $Q_{\hat{X}} = Q_{\hat{X}^n}$ . Then, by noting that

$$d_n(x^n, \hat{x}^n) = \sum_{x, \hat{x}} P_{x^n}(x) V_{P_{x^n}}(\hat{x}|x) d(x, \hat{x}) > D \quad (265)$$

never occurs for the test channel  $P_{\hat{X}^n|X^n}$ , we have

$$\begin{aligned} P_e(\Phi_n; D) &\leq P_{\hat{X}^n X^n} \left[ \log \frac{P_{\hat{X}^n|X^n}(\hat{x}^n|x^n)}{Q_{\hat{X}^n}(\hat{x}^n)} > \gamma_c - \nu \right] \\ &\quad + \tilde{\delta} + \sqrt{\frac{2\gamma_c}{|\tilde{\delta}| |\mathcal{M}_n|}} + \delta + 2^{-\nu} \end{aligned} \quad (266)$$

$$\begin{aligned} &= P_{\hat{X}^n X^n} \left[ \frac{1}{n} \log \frac{P_{\hat{X}^n|X^n}(\hat{x}^n|x^n)}{Q_{\hat{X}^n}(\hat{x}^n)} > \tilde{\gamma} - \frac{\log n}{n} \right] \\ &\quad + \sqrt{\frac{n 2^{\tilde{\gamma} n}}{|\mathcal{M}_n|}} + \frac{3}{n}, \end{aligned} \quad (267)$$

where we set  $\gamma_c = \tilde{\gamma} n$ ,  $\tilde{\delta} = \delta = \frac{1}{n}$ , and  $\nu = \log n$ . Furthermore, by noting that

$$Q_{\hat{X}^n}(\hat{x}^n) \geq \frac{1}{|\Omega_n|} \tilde{P}_q^n(\hat{x}^n) \quad (268)$$

for any  $q \in \Omega_n$ , we have

$$P_{\hat{X}^n X^n} \left[ \frac{1}{n} \log \frac{P_{\hat{X}^n|X^n}(\hat{x}^n|x^n)}{Q_{\hat{X}^n}(\hat{x}^n)} > \tilde{\gamma} - \frac{\log n}{n} \right] \quad (269)$$

$$\begin{aligned} &\leq P_{\hat{X}^n X^n} \left[ \frac{1}{n} \log \frac{P_{\hat{X}^n|X^n}(\hat{x}^n|x^n)}{Q_{\hat{X}^n}(\hat{x}^n)} > \tilde{\gamma} - \frac{\log n}{n}, P_{x^n} \in \Omega_n \right] \\ &\quad + P_{X^n} [P_{x^n} \notin \Omega_n] \end{aligned} \quad (270)$$

$$\begin{aligned} &\leq P_{\hat{X}^n X^n} \left[ \frac{1}{n} \log \frac{P_{\hat{X}^n|X^n}(\hat{x}^n|x^n)}{Q_{\hat{X}^n}(\hat{x}^n)} > \tilde{\gamma} - \frac{\log n}{n}, P_{x^n} \in \Omega_n \right] \\ &\quad + \frac{2\tilde{\tau}}{n^2} \end{aligned} \quad (271)$$

$$\begin{aligned} &\leq P_{\hat{X}^n X^n} \left[ \frac{1}{n} \log \frac{P_{\hat{X}^n|X^n}(\hat{x}^n|x^n)}{\tilde{P}_{P_{x^n}}^n(\hat{x}^n)} \right. \\ &\quad \left. > \tilde{\gamma} - \frac{\log n}{n} - \frac{|\mathcal{X}| \log(n+1)}{n}, P_{x^n} \in \Omega_n \right] + \frac{2\tilde{\tau}}{n^2}, \end{aligned} \quad (272)$$

where (271) follows from [25, Lemma 2] and (272) follows from (268) and the fact that  $|\Omega_n| \leq |\mathcal{P}_n(\mathcal{X})| \leq (n+1)^{|\mathcal{X}|}$ .

Furthermore, we also have

$$\log \frac{P_{\hat{X}^n|X^n}(\hat{x}^n|x^n)}{\tilde{P}_{P_{x^n}}^n(\hat{x}^n)} = \log \frac{|\mathcal{T}_{P_{x^n}}|}{|\mathcal{T}_{V_{P_{x^n}}}(x^n)|} \quad (273)$$

$$= nI(P_{x^n}, V_{P_{x^n}}) + O(\log n). \quad (274)$$

Thus, for  $\mu_n = O\left(\frac{\log n}{n}\right)$ , we have

$$\begin{aligned} P_e(\Phi_n; D) &\leq P_{\hat{X}^n X^n} [I(P_{x^n}, V_{P_{x^n}}) > \tilde{\gamma} - \mu_n, P_{x^n} \in \Omega_n] \\ &\quad + O\left(\frac{1}{n}\right) + \sqrt{\frac{n 2^{\tilde{\gamma} n}}{|\mathcal{M}_n|}} \end{aligned} \quad (275)$$

$$\begin{aligned} &\leq P_{\hat{X}^n X^n} \left[ R(P_{x^n}, D) > \tilde{\gamma} - \mu_n - \frac{\tau}{n}, P_{x^n} \in \Omega_n \right] \\ &\quad + O\left(\frac{1}{n}\right) + \sqrt{\frac{n 2^{\tilde{\gamma} n}}{|\mathcal{M}_n|}} \end{aligned} \quad (276)$$

$$\leq P_{\hat{X}^n X^n} \left[ R(P_{x^n}, D) > \tilde{\gamma} - \mu_n - \frac{\tau}{n} \right]$$

$$+ O\left(\frac{1}{n}\right) + \sqrt{\frac{n2^{\tilde{\gamma}n}}{|\mathcal{M}_n|}} \quad (277)$$

$$\leq P_{X^n} \left[ R(P_{x^n}, D) > \tilde{\gamma} - \mu_n - \frac{\tau}{n} \right] + O\left(\frac{1}{n}\right) + \sqrt{\frac{n2^{\tilde{\gamma}n}}{|\mathcal{M}_n|}}. \quad (278)$$

Thus, by setting  $\tilde{\gamma} = R(P_X, D) + \Delta R$ ,  $\frac{1}{n} \log |\mathcal{M}_n| = \tilde{\gamma} + \frac{2 \log n}{n}$  and by using Lemma 31 (with  $g_n = O\left(\frac{\log n}{\sqrt{n}}\right)$ ) being the residual terms in (278)), we have

$$R(n, \varepsilon; D) \leq R(P_X, D) + \sqrt{\frac{\text{Var}(j(X, D))}{n}} Q^{-1}(\varepsilon) + O\left(\frac{\log n}{n}\right) \quad (279)$$

for sufficiently large  $n$ , which implies the statement of the theorem. ■

### B. Proof Based on the $D$ -tilted Information

Let

$$\mathcal{B}_D(x^n) := \{\hat{x}^n : d_n(x^n, \hat{x}^n) \leq D\} \quad (280)$$

be the  $D$ -sphere, and let  $P_{\hat{X}^*}$  be the output distribution of the optimal test channel of

$$\min_{P_{\hat{X}^*|X}} I(X; \hat{X}). \quad (281)$$

$\mathbb{E}[d(X, \hat{X})] \leq D$

To prove Theorem 20 by the  $D$ -tilted information, we use the following lemma.

**Lemma 33** (Lemma 2 of [26]). *Under some regularity conditions, which are explicitly given in [26, Lemma 2] and satisfied by discrete memoryless sources, there exists constants  $n_0, c, K > 0$  such that*

$$P_X^n \left[ \log \frac{1}{P_{\hat{X}^*}^n(\mathcal{B}_D(x^n))} \leq \sum_{i=1}^n j(x_i, D) + C \log n + c \right] \geq 1 - \frac{K}{\sqrt{n}} \quad (282)$$

for all  $n \geq n_0$ , where  $C > 0$  is a constant given by [26, Equation (86)].

*Proof:* We construct test channel  $P_{\hat{X}^n|X^n}$  as

$$P_{\hat{X}^n|X^n}(\hat{x}^n|x^n) = \begin{cases} \frac{P_{\hat{X}^*}^n(\hat{x}^n)}{P_{\hat{X}^*}^n(\mathcal{B}_D(x^n))} & \text{if } \hat{x}^n \in \mathcal{B}_D(x^n) \\ 0 & \text{else} \end{cases} \quad (283)$$

We now use Corollary 30 for  $P_X = P_X^n$ ,  $P_{\hat{X}|X} = P_{\hat{X}^n|X^n}$ ,  $Q_{\hat{X}} = P_{\hat{X}^*}^n$ ,  $\gamma_c = \tilde{\gamma}n$ ,  $\tilde{\delta} = \delta = \frac{1}{n}$  and  $\nu = \log n$ . Then, by noting that  $d_n(x^n, \hat{x}^n) > D$  never occur for the test channel

$P_{\hat{X}^n|X^n}$ , we have

$$\begin{aligned} & P_e(\Phi_n; D) \\ & \leq P_{\hat{X}^n X^n} \left[ \log \frac{P_{\hat{X}^n|X^n}(\hat{x}^n|x^n)}{P_{\hat{X}^*}^n(\hat{x}^n)} > \tilde{\gamma}n - \log n \right] \\ & \quad + \sqrt{\frac{n2^{\tilde{\gamma}n}}{|\mathcal{M}_n|}} + \frac{3}{n} \end{aligned} \quad (284)$$

$$\begin{aligned} & \leq P_{\hat{X}^n X^n} \left[ \log \frac{1}{P_{\hat{X}^*}^n(\mathcal{B}_D(x^n))} > \tilde{\gamma}n - \log n \right] \\ & \quad + \sqrt{\frac{n2^{\tilde{\gamma}n}}{|\mathcal{M}_n|}} + \frac{3}{n} \end{aligned} \quad (285)$$

$$= P_X^n \left[ \log \frac{1}{P_{\hat{X}^*}^n(\mathcal{B}_D(x^n))} > \tilde{\gamma}n - \log n \right] + \sqrt{\frac{n2^{\tilde{\gamma}n}}{|\mathcal{M}_n|}} + \frac{3}{n} \quad (286)$$

$$\begin{aligned} & \leq P_X^n \left[ \sum_{i=1}^n j(x_i, D) > \tilde{\gamma}n - (C+1) \log n - c \right] \\ & \quad + P_X^n \left[ \log \frac{1}{P_{\hat{X}^*}^n(\mathcal{B}_D(x^n))} > \sum_{i=1}^n j(x_i, D) + C \log n + c \right] \\ & \quad + \sqrt{\frac{n2^{\tilde{\gamma}n}}{|\mathcal{M}_n|}} + \frac{3}{n} \end{aligned} \quad (287)$$

$$\begin{aligned} & \leq P_X^n \left[ \sum_{i=1}^n j(x_i, D) > \tilde{\gamma}n - (C+1) \log n - c \right] \\ & \quad + \frac{K}{\sqrt{n}} + \sqrt{\frac{n2^{\tilde{\gamma}n}}{|\mathcal{M}_n|}} + \frac{3}{n}, \end{aligned} \quad (288)$$

where (288) follows from Lemma 33. Thus, by setting  $\tilde{\gamma} = \frac{1}{n} \log |\mathcal{M}_n| - \frac{2 \log n}{n}$  and by applying the Berry-Esséen theorem [64], we have (279) for sufficiently large  $n$ , which implies the statement of the theorem. ■

## APPENDIX M

### CARDINALITY BOUND FOR SECOND-ORDER CODING THEOREMS

The following three theorems allow us to restrict the cardinalities of auxiliary random variables in second-order coding theorems.

**Theorem 34** (Cardinality Bound for WAK). *For any  $P_{UTXY} \in \tilde{\mathcal{P}}(P_{XY})$ , where  $\tilde{\mathcal{P}}(P_{XY})$  is defined in VI-A, there exists  $P_{U'T'XY}$  with  $|\mathcal{U}'| \leq |\mathcal{Y}| + 4$  and  $|\mathcal{T}'| \leq 5$  such that (i)  $\mathcal{X} \times \mathcal{Y}$ -marginal of  $P_{U'T'XY}$  is  $P_{XY}$ , (ii)  $U' - (Y, T') - X$  forms a Markov chain, (iii)  $T'$  is independent of  $(X, Y)$ , and (iv)  $P_{U'T'XY}$  preserves the mean  $\mathbf{J}$  of the entropy-information density vector and the entropy-information dispersion matrix  $\mathbf{V}$ , i.e.,*

$$\mathbf{J}(P_{UTXY}) = \mathbf{J}(P_{U'T'XY}) \quad (289)$$

$$\mathbf{V}(P_{UTXY}) = \mathbf{V}(P_{U'T'XY}). \quad (290)$$

**Theorem 35** (Cardinality Bound for WZ). *For any pair of  $P_{UTXY} \in \tilde{\mathcal{P}}(P_{XY})$  and  $P_{\hat{X}|UYT}$ , where  $\tilde{\mathcal{P}}(P_{XY})$  is defined in VI-B, there exist  $P_{U'T'XY}$  and  $P_{\hat{X}'|U'YT'} : \mathcal{U}' \times \mathcal{Y} \times \mathcal{T}' \rightarrow$*

$\hat{\mathcal{X}}$  with  $|\mathcal{U}'| \leq |\mathcal{Y}| + 8$  and  $|\mathcal{T}'| \leq 9$  such that (i)  $\mathcal{X} \times \mathcal{Y}$ -marginal of  $P_{U'T'XY}$  is  $P_{XY}$ , (ii)  $U' - (X, T') - Y$  forms a Markov chain, (iii)  $T'$  is independent of  $(X, Y)$ , and (iv)  $P_{U'T'XY}$  and  $P_{\hat{\mathcal{X}}'|U'YT'}$  preserve  $\mathbf{J}$  and  $\mathbf{V}$ , i.e.,

$$\mathbf{J}(P_{UTXY}, P_{\hat{\mathcal{X}}|UYT}) = \mathbf{J}(P_{U'T'XY}, P_{\hat{\mathcal{X}}'|U'YT'}) \quad (291)$$

$$\mathbf{V}(P_{UTXY}, P_{\hat{\mathcal{X}}|UYT}) = \mathbf{V}(P_{U'T'XY}, P_{\hat{\mathcal{X}}'|U'YT'}). \quad (292)$$

**Theorem 36** (Cardinality Bound for GP). *For any  $P_{UTSXY} \in \tilde{\mathcal{P}}(W, P_S)$ , where  $\tilde{\mathcal{P}}(W, P_S)$  is defined in VI-C, there exists  $P_{U'T'SXY}$  with  $|\mathcal{U}'| \leq |\mathcal{Y}| + 6$  and  $|\mathcal{T}'| \leq 9$  such that (i)  $\mathcal{S} \times \mathcal{X} \times \mathcal{Y}$ -marginal of  $P_{U'T'SXY}$  is  $P_{SXY}$ , (ii)  $U' - (X, S, T') - Y$  forms a Markov chain, (iii)  $T'$  is independent of  $S$ , and (iv)  $P_{U'T'SXY}$  preserves  $\mathbf{J}$  and  $\mathbf{V}$ , i.e.,*

$$\mathbf{J}(P_{UTSXY}) = \mathbf{J}(P_{U'T'SXY}) \quad (293)$$

$$\mathbf{V}(P_{UTSXY}) = \mathbf{V}(P_{U'T'SXY}). \quad (294)$$

We can prove all of the three theorems in the same manner. Because the proof for Wyner-Ziv problem is most complicated, we prove Theorem 35 in M-A, and then, give proof sketches for Theorems 34 and 36 in M-B.

#### A. Proof of Cardinality Bound for WZ problem

To prove Theorem 35, we use variations of the support lemma. Note that we can identify  $\mathcal{P}(\mathcal{X}) \times \mathcal{P}(\hat{\mathcal{X}}|\mathcal{Y})$  with a connected compact subset of  $|\mathcal{X}||\hat{\mathcal{X}}||\mathcal{Y}|$ -dimensional Euclidean space. Hence, as a consequence of the Fenchel-Eggleston-Carathéodory theorem (see, e.g. [1, Appendix A]), we have the following lemma.

**Lemma 37.** *Let  $f_j$  ( $j = 1, 2, \dots, k$ ) be real-valued continuous functions on  $\mathcal{P}(\mathcal{X}) \times \mathcal{P}(\hat{\mathcal{X}}|\mathcal{Y})$ . Then, for any  $P_U \in \mathcal{P}(\mathcal{U})$  and any collection  $\{(P_{X|U}(\cdot|u), P_{\hat{\mathcal{X}}|YU}(\cdot|u)) : u \in \mathcal{U}\} \subset \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\hat{\mathcal{X}}|\mathcal{Y})$ , there exist a distribution  $P_{U'} \in \mathcal{P}(\mathcal{U}')$  with  $|\mathcal{U}'| \leq k$  and a collection  $\{(P_{X'|U'}(\cdot|u'), P_{\hat{\mathcal{X}}'|Y'U'}(\cdot|u')) : u' \in \mathcal{U}'\} \subset \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\hat{\mathcal{X}}|\mathcal{Y})$  such that for  $j = 1, 2, \dots, k$ ,*

$$\begin{aligned} & \int_{\mathcal{U}} f_j \left( P_{X|U}(\cdot|u), P_{\hat{\mathcal{X}}|YU}(\cdot|u) \right) dP_U(u) \\ &= \sum_{u' \in \mathcal{U}'} f_j \left( P_{X'|U'}(\cdot|u'), P_{\hat{\mathcal{X}}'|Y'U'}(\cdot|u') \right) P_{U'}(u'). \end{aligned} \quad (295)$$

**Remark 10.** *Let us consider applying Lemma 37 to a case where  $P_{\hat{\mathcal{X}}|YU}$  is a deterministic function. In this case,  $P_U$  appearing in the left hand side of (295) satisfies  $P_U(u) > 0$  only if  $P_{\hat{\mathcal{X}}|YU}(\cdot|u)$  is deterministic, i.e., for each  $y$  there exists  $\hat{x}$  satisfying  $P_{\hat{\mathcal{X}}|YU}(\hat{x}|y, u) = 1$ . On the other hand, Lemma 37 does not guarantee that we can choose  $\mathcal{U}'$  and a collection of distributions so that  $P_{\hat{\mathcal{X}}'|Y'U'}(\cdot|u') \in \mathcal{P}(\mathcal{Y}|\mathcal{X})$  is deterministic for all  $u' \in \mathcal{U}'$ . That is why we use a stochastic reproduction function to establish bounds on the cardinalities of the auxiliary random variables.*

Similarly, by identifying  $\mathcal{P}(\mathcal{U}|\mathcal{X}) \times \mathcal{P}(\hat{\mathcal{X}}|\mathcal{U} \times \mathcal{Y})$  with a connected compact subset of Euclidean space, we have another variation of the support lemma.

**Lemma 38.** *Let  $f_j$  ( $j = 1, 2, \dots, k$ ) be real-valued continuous functions on  $\mathcal{P}(\mathcal{U}|\mathcal{X}) \times \mathcal{P}(\hat{\mathcal{X}}|\mathcal{U} \times \mathcal{Y})$ . Then, for any  $P_T \in \mathcal{P}(\mathcal{T})$  and any collection  $\{(P_{U|XT}(\cdot|t), P_{\hat{\mathcal{X}}|UYT}(\cdot|t, \cdot, t)) : t \in \mathcal{T}\} \subset \mathcal{P}(\mathcal{U}|\mathcal{X}) \times \mathcal{P}(\hat{\mathcal{X}}|\mathcal{U} \times \mathcal{Y})$ , there exist a distribution  $P_{T'} \in \mathcal{P}(\mathcal{T}')$  with  $|\mathcal{T}'| \leq k$  and a collection  $\{(P_{U'|X'T'}(\cdot|t'), P_{\hat{\mathcal{X}}'|U'Y'T'}(\cdot|t', \cdot, t')) : t' \in \mathcal{T}'\} \subset \mathcal{P}(\mathcal{U}|\mathcal{X}) \times \mathcal{P}(\hat{\mathcal{X}}|\mathcal{U} \times \mathcal{Y})$  such that for  $j = 1, 2, \dots, k$ ,*

$$\begin{aligned} & \int_{\mathcal{T}} f_j \left( P_{U|XT}(\cdot|t), P_{\hat{\mathcal{X}}|UYT}(\cdot|t, \cdot, t) \right) dP_T(t) \\ &= \sum_{t' \in \mathcal{T}'} f_j \left( P_{U'|X'T'}(\cdot|t'), P_{\hat{\mathcal{X}}'|U'Y'T'}(\cdot|t', \cdot, t') \right) P_{T'}(t'). \end{aligned} \quad (296)$$

*Proof of Theorem 35:*

1) Bound on  $|\mathcal{U}'|$ : Fix  $P_{UTXY} \in \tilde{\mathcal{P}}(P_{XY})$ . Without loss of generality, we assume that  $\mathcal{X} = \{1, 2, \dots, |\mathcal{X}|\}$ . Let us consider the following  $|\mathcal{X}|+8$  functions: For  $(Q, q) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\hat{\mathcal{X}}|\mathcal{Y})$ ,

$$f_j(Q, q) := Q(j), \quad j = 1, 2, \dots, |\mathcal{X}| - 1 \quad (297)$$

$$\begin{aligned} f_{|\mathcal{X}|}(Q, q) &:= - \sum_{y \in \mathcal{Y}} \left[ \sum_{x \in \mathcal{X}} P_{Y|X}(y|x) Q(x) \right] \\ &\quad \times \log \left[ \sum_{x \in \mathcal{X}} P_{Y|X}(y|x) Q(x) \right] \end{aligned} \quad (298)$$

$$f_{|\mathcal{X}|+1}(Q, q) := - \sum_{x \in \mathcal{X}} Q(x) \log Q(x) \quad (299)$$

$$f_{|\mathcal{Y}|+2}(Q, q) := \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \sum_{\hat{x} \in \hat{\mathcal{X}}} Q(x) P_{Y|X}(y|x) q(\hat{x}|y) d(x, \hat{x}) \quad (300)$$

$$\begin{aligned} f_{|\mathcal{X}|+3}(Q, q) &:= \sum_{y \in \mathcal{Y}} \left[ \sum_{x \in \mathcal{X}} P_{Y|X}(y|x) Q(x) \right] \\ &\quad \times \left\{ \log \frac{\left[ \sum_{x \in \mathcal{X}} P_{Y|X}(y|x) Q(x) \right]}{P_Y(y)} \right\}^2 \end{aligned} \quad (301)$$

$$f_{|\mathcal{X}|+4}(Q, q) := \sum_{x \in \mathcal{X}} Q(x) \left\{ \log \frac{Q(x)}{P_X(x)} \right\}^2 \quad (302)$$

$$\begin{aligned} f_{|\mathcal{X}|+5}(Q, q) &:= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \sum_{\hat{x} \in \hat{\mathcal{X}}} Q(x) P_{Y|X}(y|x) q(\hat{x}|y) \\ &\quad \times \{d(x, \hat{x})\}^2 \end{aligned} \quad (303)$$

$$\begin{aligned} f_{|\mathcal{X}|+6}(Q, q) &:= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} Q(x) P_{Y|X}(y|x) \\ &\quad \times \left( \log \frac{P_Y(y)}{\sum_{\bar{x} \in \mathcal{X}} Q(\bar{x}) P_{Y|X}(y|\bar{x})} \right) \\ &\quad \times \left( \log \frac{Q(x)}{P_X(x)} \right) \end{aligned} \quad (304)$$

$$\begin{aligned} f_{|\mathcal{X}|+7}(Q, q) &:= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \sum_{\hat{x} \in \hat{\mathcal{X}}} Q(x) P_{Y|X}(y|x) q(\hat{x}|y) \\ &\quad \times \left( \log \frac{P_Y(y)}{\sum_{\bar{x} \in \mathcal{X}} Q(\bar{x}) P_{Y|X}(y|\bar{x})} \right) d(x, \hat{x}) \end{aligned} \quad (305)$$

$$f_{|\mathcal{X}|+8}(Q, q) := \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \sum_{\hat{x} \in \hat{\mathcal{X}}} Q(x) P_{Y|X}(y|x) q(\hat{x}|y) \times \left( \log \frac{Q(x)}{P_X(x)} \right) d(x, \hat{x}). \quad (306)$$

Fix  $t \in \mathcal{T}$ . Then, Lemma 37 guarantees that there exist  $P_{U'|T}(\cdot|t) \in \mathcal{P}(\mathcal{U}')$  with  $|\mathcal{U}'| \leq |\mathcal{X}| + 8$  and a collection  $\{(P_{X'|U'T}(\cdot|u', t), P_{\hat{X}'|Y'U'T}(\cdot|u', t)) : u' \in \mathcal{U}'\} \subset \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\hat{\mathcal{X}}|\mathcal{Y})$  such that for all  $j = 1, 2, \dots, |\mathcal{X}| + 8$ ,

$$\begin{aligned} & \sum_{u \in \mathcal{U}} f_j \left( P_{X|UT}(\cdot|u, t), P_{\hat{X}|YUT}(\cdot|u, t) \right) P_{U|T}(u|t) \\ &= \sum_{u' \in \mathcal{U}'} f_j \left( P_{X'|U'T}(\cdot|u', t), P_{\hat{X}'|Y'U'T}(\cdot|u', t) \right) P_{U'|T}(u'|t). \end{aligned} \quad (307)$$

Now, we have  $P_{U'|T}, P_{X'|U'T}, P_{\hat{X}'|Y'U'T}$  satisfying (307) for each  $t \in \mathcal{T}$ . Let  $U', T, X', Y', \hat{X}'$  be random variables induced by  $P_{U'|T}, P_{X'|U'T}, P_{\hat{X}'|Y'U'T}$ , and  $P_{Y|X}, P_T$ , i.e., for each  $(u', t, x, y, \hat{x}) \in \mathcal{U}' \times \mathcal{T} \times \mathcal{X} \times \mathcal{Y} \times \hat{\mathcal{X}}$ ,

$$\begin{aligned} & P_{U'TX'Y'\hat{X}'}(u', t, x, y, \hat{x}) \\ &:= P_T(t) P_{U'|T}(u'|t) P_{X'|U'T}(x|u', t) P_{Y|X}(y|x) \\ & \quad \times P_{\hat{X}'|Y'U'T}(\hat{x}|y, u', t). \end{aligned} \quad (308)$$

Observe that  $U' - (T', Y') - X'$  forms a Markov chain and that  $T$  is independent of  $(X', Y')$ . Further, (307) with  $j = 1, \dots, |\mathcal{X}| - 1$  guarantees that  $P_{X'Y'} = P_{XY}$ . Hence, we have  $P_{TX'Y'} = P_T P_{XY}$ , and thus, we can write  $P_{U'TX'Y'\hat{X}'} = P_{U'TXY\hat{X}'}$ .

On the other hand, some calculations show that, for each  $t \in \mathcal{T}$ ,

$$\begin{aligned} & H(Y|U, T = t) \\ &= \sum_{u \in \mathcal{U}} f_{|\mathcal{X}|}(P_{X|UT}(\cdot|u, t), P_{\hat{X}|YUT}(\cdot|u, t)) P_{U|T}(u|t) \end{aligned} \quad (309)$$

$$\begin{aligned} & H(X|U, T = t) \\ &= \sum_{u \in \mathcal{U}} f_{|\mathcal{X}|+1}(P_{X|UT}(\cdot|u, t), P_{\hat{X}|YUT}(\cdot|u, t)) P_{U|T}(u|t) \end{aligned} \quad (310)$$

$$\begin{aligned} & \mathbb{E}[d(X, \hat{X}|t)] \\ &= \sum_{u \in \mathcal{U}} f_{|\mathcal{X}|+2}(P_{X|UT}(\cdot|u, t), P_{\hat{X}|YUT}(\cdot|u, t)) P_{U|T}(u|t) \end{aligned} \quad (311)$$

$$\begin{aligned} & \text{Var} \left( -\log \frac{P_{Y|UT}(Y|U, t)}{P_Y(Y)} \right) \\ &= \sum_{u \in \mathcal{U}} f_{|\mathcal{X}|+3}(P_{X|UT}(\cdot|u, t), P_{\hat{X}|YUT}(\cdot|u, t)) P_{U|T}(u|t) \\ & \quad - \{H(Y) - H(Y|U, T = t)\}^2 \end{aligned} \quad (312)$$

$$\begin{aligned} & \text{Var} \left( \log \frac{P_{X|UT}(X|U, t)}{P_X(X)} \right) \\ &= \sum_{u \in \mathcal{U}} f_{|\mathcal{X}|+4}(P_{X|UT}(\cdot|u, t), P_{\hat{X}|YUT}(\cdot|u, t)) P_{U|T}(u|t) \\ & \quad - \{H(X) - H(X|U, T = t)\}^2 \end{aligned} \quad (313)$$

$$\begin{aligned} & \text{Var} \left( d(X, \hat{X}|t) \right) \\ &= \sum_{u \in \mathcal{U}} f_{|\mathcal{X}|+5}(P_{X|UT}(\cdot|u, t), P_{\hat{X}|YUT}(\cdot|u, t)) P_{U|T}(u|t) \\ & \quad - \mathbb{E}[d(X, \hat{X}|t)]^2 \end{aligned} \quad (314)$$

and

$$\begin{aligned} & \text{Cov} \left( -\log \frac{P_{Y|UT}(Y|U, t)}{P_Y(Y)}, \log \frac{P_{X|UT}(X|U, t)}{P_X(X)} \right) \\ &= \sum_{u \in \mathcal{U}} f_{|\mathcal{X}|+6}(P_{X|UT}(\cdot|u, t), P_{\hat{X}|YUT}(\cdot|u, t)) P_{U|T}(u|t) \\ & \quad + \{H(Y) - H(Y|U, T = t)\} \{H(X) - H(X|U, T = t)\}, \end{aligned} \quad (315)$$

$$\begin{aligned} & \text{Cov} \left( -\log \frac{P_{Y|UT}(Y|U, t)}{P_Y(Y)}, d(X, \hat{X}|t) \right) \\ &= \sum_{u \in \mathcal{U}} f_{|\mathcal{X}|+7}(P_{X|UT}(\cdot|u, t), P_{\hat{X}|YUT}(\cdot|u, t)) P_{U|T}(u|t) \\ & \quad + \{H(Y) - H(Y|U, T = t)\} \mathbb{E}[d(X, \hat{X}|t)], \end{aligned} \quad (316)$$

$$\begin{aligned} & \text{Cov} \left( \log \frac{P_{X|UT}(X|U, t)}{P_X(X)}, d(X, \hat{X}|t) \right) \\ &= \sum_{u \in \mathcal{U}} f_{|\mathcal{X}|+8}(P_{X|UT}(\cdot|u, t), P_{\hat{X}|YUT}(\cdot|u, t)) P_{U|T}(u|t) \\ & \quad - \{H(X) - H(X|U, T = t)\} \mathbb{E}[d(X, \hat{X}|t)]. \end{aligned} \quad (317)$$

Thus, equations (307) and (310)–(317) guarantee that a pair  $P_{U'XY\hat{X}'|T=t}$  preserves all components of  $\mathbf{J}$  and  $\mathbf{V}$  for each  $t \in \mathcal{T}$ . By taking the average with respect to  $T$ , we can show that the pair  $(P_{U'TXY}, P_{\hat{X}'|Y'YT})$  satisfies the all conditions of the theorem except the cardinality of  $T$ .

2) Bound on  $|\mathcal{T}'|$ : Fix  $P_{UTXY\hat{X}} \in \tilde{\mathcal{P}}(P_{XY})$  and  $P_{\hat{X}|UYT}$ . By the first part of the proof, we can assume that  $\mathcal{U} = \mathcal{U}'$  and  $|\mathcal{U}| = |\mathcal{U}'| \leq |\mathcal{X}| + 8$ . Let us consider the following 9 functions on  $\mathcal{P}(\mathcal{U} \times \mathcal{X} \times \mathcal{Y} \times \hat{\mathcal{X}})$ :

$$F_1(P_{UXY\hat{X}}) := I(Y; U) \quad (318)$$

$$F_2(P_{UXY\hat{X}}) := I(X; U) \quad (319)$$

$$F_3(P_{UXY\hat{X}}) := \mathbb{E}[d(X, \hat{X})] \quad (320)$$

$$F_4(P_{UXY\hat{X}}) := \text{Var} \left( -\log \frac{P_{Y|U}(Y|U)}{P_Y(Y)} \right) \quad (321)$$

$$F_5(P_{UXY\hat{X}}) := \text{Var} \left( \log \frac{P_{X|U}(X|U)}{P_X(X)} \right) \quad (322)$$

$$F_6(P_{UXY\hat{X}}) := \text{Var} \left( d(X, \hat{X}) \right) \quad (323)$$

$$F_7(P_{UXY\hat{X}}) := \text{Cov} \left( -\log \frac{P_{Y|U}(Y|U)}{P_Y(Y)}, \log \frac{P_{X|U}(X|U)}{P_X(X)} \right) \quad (324)$$

$$F_8(P_{UXY\hat{X}}) := \text{Cov} \left( -\log \frac{P_{Y|U}(Y|U)}{P_Y(Y)}, d(X, \hat{X}) \right) \quad (325)$$

$$F_9(P_{UXY\hat{X}}) := \text{Cov} \left( \log \frac{P_{X|U}(X|U)}{P_X(X)}, d(X, \hat{X}) \right) \quad (326)$$

and a function  $F: \mathcal{P}(\mathcal{U}|\mathcal{X}) \times \mathcal{P}(\hat{\mathcal{X}}|\mathcal{U} \times \mathcal{Y}) \rightarrow \mathcal{P}(\mathcal{U} \times \mathcal{X} \times \mathcal{Y} \times \hat{\mathcal{X}})$  such as  $P_{UXY\hat{X}} = F(P_{U|X}, P_{\hat{X}|UY})$  satisfies

$$P_{UXY\hat{X}}(u, x, y, \hat{x}) = P_{XY}(x, y) P_{U|X}(u|x) P_{\hat{X}|YU}(\hat{x}|y, u). \quad (327)$$

Then, by applying Lemma 38 to  $f_j(\cdot) := F_j(F(\cdot))$  ( $j = 1, 2, \dots, 9$ ), we have  $P_{T'} \in \mathcal{P}(\mathcal{T}')$  with  $|\mathcal{T}'| \leq 9$  and  $\{(P_{U'|X'T'}(\cdot|\cdot, t'), P_{\hat{X}'|U'Y'T'}(\cdot|\cdot, t')) : t' \in \mathcal{T}'\} \subset \mathcal{P}(\mathcal{U}|\mathcal{X}) \times \mathcal{P}(\hat{\mathcal{X}}|\mathcal{U} \times \mathcal{Y})$  satisfying (296). By  $P_{T'}$ ,  $(P_{U'|X'T'}, P_{\hat{X}'|U'Y'T'})$  and  $P_{XY}$ , let us define  $P_{U'T'XY\hat{X}'} = P_{U'T'XY\hat{X}'}$  as

$$\begin{aligned} P_{U'T'XY\hat{X}'}(u', t', x, y, \hat{x}') \\ = P_{XY}(x, y) P_{T'}(t') P_{U'|X'T'}(u'|x, t') P_{\hat{X}'|U'Y'T'}(\hat{x}'|u', y, t'). \end{aligned} \quad (328)$$

We can verify that the pair  $(P_{U'T'XY}, P_{\hat{X}'|U'Y'T'})$  derived from  $P_{U'T'XY\hat{X}'}$  satisfies the conditions of the theorem. ■

### B. Proof Sketches of Cardinality Bounds for WAK and GP problems

*Proof of Theorem 34:* We fix  $t \in \mathcal{T}$  and then consider the following  $|\mathcal{Y}|+4$  quantities:  $|\mathcal{Y}|-1$  elements  $P_Y(y)$  ( $y = 1, 2, \dots, |\mathcal{Y}|-1$ ) of  $P_Y$ , the conditional entropy  $H(X|U, T=t)$ , the mutual information  $I(U; Y|T=t)$ , two variances on the diagonals of  $\text{Cov}(\mathbf{j}(U, X, Y|t))$ , and the covariance in the upper part of  $\text{Cov}(\mathbf{j}(U, X, Y|t))$ . Then, in the same manner as the first part of the proof for Wyner-Ziv problem, we can choose a random variable  $U' \sim P_{U'|T=t} \in \mathcal{P}(\mathcal{U})$  with  $|\mathcal{U}'| \leq |\mathcal{Y}|+4$  which preserves the marginal distribution  $P_{XY|T=t}$ ,  $\mathbb{E}[\mathbf{j}(U, X, Y|t)]$ , and  $\text{Cov}(\mathbf{j}(U, X, Y|t))$ . By taking the average with respect to  $T$ , we can show that  $U'$  satisfies the conditions of the theorem. Further, in the same way as the second part of the proof for Wyner-Ziv problem, we can show that  $T'$  with  $|\mathcal{T}'| \leq 5$  preserves the following five quantities: two elements of  $\mathbf{J}$ , two variances along the diagonals of  $\mathbf{V}$ , and the covariance in the upper part of  $\mathbf{V}$ . ■

*Proof of Theorem 36:* We fix  $t \in \mathcal{T}$  and then consider the following  $|\mathcal{S}||\mathcal{X}|+6$  quantities:  $|\mathcal{S}||\mathcal{X}|-1$  elements  $P_{SX}(s, x)$  of  $P_{SX}$ , two mutual informations  $I(U; Y|t)$ ,  $I(U; S|t)$ , two variances  $\text{Var}(\log P_{Y|UT}(Y|U, t)/P_{Y|T}(Y|t))$ ,  $\text{Var}(-\log P_{S|UT}(S|U, t)/P_S(S))$ , and three covariances in the strict upper triangular part of  $\text{Cov}(\mathbf{j}(U, S, X, Y|t))$ . Note that, if the marginal distribution  $P_{SXY|T=t}$  is preserved then the average  $\mathbb{E}[\mathbf{g}(X_T)|T=t]$  and the variance  $\text{Var}(\mathbf{g}(X_T)|T=t)$  of  $\mathbf{g}(X_T)$  with respect to the distribution  $P_{X|T=t}$  is automatically preserved. Hence, in the same manner as the first part of the proof for Wyner-Ziv problem, we can choose a random variable  $U' \sim P_{U'|T=t} \in \mathcal{P}(\mathcal{U})$  with  $|\mathcal{U}'| \leq |\mathcal{S}||\mathcal{X}|+6$  which preserves the marginal distribution  $P_{SX|T=t}$ ,  $\mathbb{E}[\mathbf{j}(U, S, X, Y|t)]$ , and  $\text{Cov}(\mathbf{j}(U, S, X, Y|t))$ . By taking the average with respect to  $T$ , we can show that  $U'$  satisfies the conditions of the theorem. Further, in the same way as the second part of the proof for Wyner-Ziv problem, we can show that  $T'$  with  $|\mathcal{T}'| \leq 5$  preserves the following nine quantities: three elements of  $\mathbf{J}$ , three variances along the diagonals of  $\mathbf{V}$ , and three covariances in the strict upper triangular part of  $\mathbf{V}$ . ■

### Acknowledgements

The authors would like to thank J. Scarlett for pointing out an error of the numerical calculation of the GP problem in

an earlier version of the paper. The authors also appreciate anonymous reviewers for valuable comments, in particular for pointing out Remark 8. The work of first author is supported in part by JSPS Postdoctoral Fellowships for Research Abroad. The work of third author is supported in part by NUS startup grant R-263-000-A98-750/133 and in part by A\*STAR, Singapore.

### REFERENCES

- [1] A. El Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge, U.K.: Cambridge University Press, 2012.
- [2] A. D. Wyner, "On source coding with side information at the decoder," *IEEE Trans. on Inf. Th.*, vol. 21, no. 3, pp. 294–300, 1975.
- [3] R. Ahlswede and J. Körner, "Source coding with side information and a converse for the degraded broadcast channel," *IEEE Trans. on Inf. Th.*, vol. 21, no. 6, pp. 629–637, 1975.
- [4] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. on Inf. Th.*, vol. 22, no. 1, pp. 1–10, Jan 1976.
- [5] S. Gelfand and M. Pinsker, "Coding for channel with random parameters," *Prob. of Control and Inf. Th.*, vol. 9, no. 1, pp. 19–31, 1980.
- [6] S. Verdú, "Non-asymptotic achievability bounds in multiuser information theory," in *Allerton Conference*, 2012.
- [7] T. S. Han, *Information-Spectrum Methods in Information Theory*. Springer Berlin Heidelberg, Feb 2003.
- [8] S. Miyake and F. Kanaya, "Coding theorems on correlated general sources," *IEICE Trans. on Fundamentals of Electronics, Communications and Computer*, vol. E78-A, no. 9, pp. 1063–70, 1995.
- [9] K.-I. Iwata and J. Muramatsu, "An information-spectrum approach to rate-distortion function with side information," *IEICE Trans. on Fundamentals of Electronics, Communications and Computer*, vol. E85-A, no. 6, pp. 1387–95, 2002.
- [10] V. Y. F. Tan, "A formula for the capacity of the general Gel'fand-Pinsker channel," in *Int. Symp. Inf. Th.*, Istanbul, Turkey, 2013.
- [11] M. Hayashi, "Second-order asymptotics in fixed-length source coding and intrinsic randomness," *IEEE Trans. on Inf. Th.*, vol. 54, pp. 4619–37, Oct 2008.
- [12] —, "Information spectrum approach to second-order coding rate in channel coding," *IEEE Trans. on Inf. Th.*, vol. 55, pp. 4947–66, Nov 2009.
- [13] T. S. Han and S. Verdú, "Approximation theory of output statistics," *IEEE Trans. on Inf. Th.*, vol. 39, no. 3, pp. 752–72, Mar 1993.
- [14] M. Hayashi, "General nonasymptotic and asymptotic formulas in channel resolvability and identification capacity and their application to the wiretap channel," *IEEE Trans. on Inf. Th.*, vol. 52, no. 4, pp. 1562–75, Apr 2006.
- [15] C. H. Bennett, P. W. Shor, J. A. Smolin, and A. V. Thapliyal, "Entanglement-assisted capacity of a quantum channel and the reverse Shannon theorem," *IEEE Trans. on Inf. Th.*, vol. 48, no. 10, pp. 2637–2655, Oct 2002.
- [16] A. Winter, "Compression of sources of probability distributions and density operators," *arXiv:quant-ph/0208131*, 2002.
- [17] P. Cuff, "Distributed channel synthesis," *IEEE Trans. on Inf. Th.*, vol. 59, no. 11, pp. 7071–7096, Nov. 2013.
- [18] T. M. Cover, "A proof of the data compression theorem of Slepian and Wolf for ergodic sources," *IEEE Trans. Inf. Th.*, vol. 21, pp. 226–228, Mar. 1975.
- [19] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. on Inf. Th.*, vol. 19, pp. 471–80, 1973.
- [20] S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Trans. on Inf. Th.*, vol. 40, no. 4, pp. 1147–57, Apr 1994.
- [21] F. Götzte, "On the rate of convergence in the multivariate CLT," *The Annals of Probability*, vol. 19, no. 2, pp. 721–739, 1991.
- [22] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding in the finite blocklength regime," *IEEE Trans. on Inf. Th.*, vol. 56, pp. 2307–59, May 2010.
- [23] V. Y. F. Tan and O. Kosut, "On the dispersions of three network information theory problems," *IEEE Trans. on Inf. Th.*, vol. 60, no. 2, pp. 881–903, Feb 2014.
- [24] D. Wang, A. Ingber, and Y. Kochman, "The dispersion of joint source-channel coding," in *Allerton Conference*, 2011, arXiv:1109.6310.
- [25] A. Ingber and Y. Kochman, "The dispersion of lossy source coding," in *Data Compression Conference (DCC)*, 2011.

- [26] V. Kostina and S. Verdú, "Fixed-length lossy compression in the finite blocklength regime," *IEEE Trans. on Inf. Th.*, vol. 58, no. 6, pp. 3309–38, Jun 2012.
- [27] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," *IRE Int. Conv. Rec.*, vol. 7, pp. 142–163, 1959.
- [28] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2011.
- [29] R. Ahlswede and P. Gács and J. Körner, "Bounds on conditional probabilities with applications in multi-user communication," *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, vol. 34, no. 3, pp. 157–177, 1976.
- [30] S. Kuzuoka, "A simple technique for bounding the redundancy of source coding with side information," in *Int. Symp. Inf. Th.*, Boston, MA, 2012.
- [31] H. Tyagi and P. Narayan, "The Gelfand-Pinsker channel: Strong converse and upper bound for the reliability function," in *Proc. of IEEE Intl. Symp. on Info. Theory*, Seoul, Korea, 2009.
- [32] Y. Steinberg and S. Verdú, "Simulation of random processes and rate-distortion theory," *IEEE Trans. on Inf. Th.*, vol. 42, no. 1, pp. 63–86, Jan 1996.
- [33] C. H. Bennett, I. Devetak, A. W. Harrow, P. W. Shor, and A. Winter, "The quantum reverse shannon theorem," *arXiv:0912.5537*, 2009.
- [34] M. H. Yassaee, M. R. Aref, and A. Gohari, "Achievability proof via output statistics of random binning," *IEEE Trans. on Inf. Th.*, vol. 60, no. 11, pp. 6760–6786, Nov. 2014.
- [35] —, "Non-asymptotic output statistics of random binning and its applications," *arXiv:1303.0695*, Mar 2013.
- [36] K. Marton, "A coding theorem for the discrete memoryless broadcast channel," *IEEE Trans. on Inf. Th.*, vol. 25, pp. 306–311, Mar 1979.
- [37] A. D. Wyner, "The wire-tap channel," *The Bell Systems Technical Journal*, vol. 54, pp. 1355–1387, 1975.
- [38] M. H. Yassaee, M. R. Aref, and A. Gohari, "A technique for deriving one-shot achievability results in network information theory," *arXiv:1303.0696*, Mar 2013.
- [39] V. Strassen, "Asymptotische Abschätzungen in Shannons Informations-theorie," in *Trans. Third. Prague Conf. Inf. Th.*, 1962, pp. 689–723.
- [40] I. Kontoyiannis, "Second-order noiseless source coding theorems," *IEEE Trans. on Inf. Th.*, pp. 1339–41, Jul 1997.
- [41] D. Baron, M. A. Khojastepour, and R. G. Baraniuk, "How quickly can we approach channel capacity?" in *Asilomar Conf.*, 2004.
- [42] Y.-W. Huang and P. Moulin, "Finite blocklength coding for multiple access channels," in *Int. Symp. Inf. Th.*, 2012.
- [43] E. MolavianJazi and J. N. Laneman, "Simpler achievable rate regions for multiaccess with finite blocklength," in *Int. Symp. Inf. Th.*, Boston, MA, 2012.
- [44] R. Nomura and T. S. Han, "Second-order Slepian-Wolf coding theorems for non-mixed and mixed sources," *IEEE Trans. on Inf. Th.*, vol. 60, no. 9, pp. 5553–5572, Sep 2014.
- [45] E. Haim, Y. Kochman, and U. Erez, "A note on the dispersion of network problems," in *Convention of Electrical and Electronics Engineers in Israel (IEEEI)*, 2012.
- [46] A. Gupta and S. Verdú, "Operational duality between Gelfand-Pinsker and Wyner-Ziv coding," in *Intl. Symp. Inf. Th.*, Austin, TX, 2010.
- [47] M. Costa, "Writing on dirty paper," *IEEE Trans. on Inf. Th.*, vol. 29, no. 3, pp. 439–441, Mar 1983.
- [48] A. Feinstein, "A new basic theorem of information theory," *IEEE Trans. on Inf. Th.*, vol. 4, no. 4, pp. 2–22, 1954.
- [49] P. Moulin and Y. Wang, "Capacity and random-coding exponents for channel coding with side information," *IEEE Trans. on Inf. Th.*, vol. 53, no. 4, pp. 1326–1347, Apr 2007.
- [50] V. Bentkus, "On the dependence of the Berry-Esseen bound on dimension," *J. Stat. Planning and Inference*, vol. 113, pp. 385–402, 2003.
- [51] B. Kelly and A. Wagner, "Reliability in source coding with side information," *IEEE Trans. on Inf. Th.*, vol. 58, no. 8, pp. 5086–5111, Aug 2012.
- [52] M. Tomamichel and V. Y. F. Tan, "A tight upper bound for the third-order asymptotics of discrete memoryless channels," *IEEE Trans. on Inf. Th.*, vol. 59, no. 11, pp. 7041–7051, Nov 2013.
- [53] W. Gu, R. Koetter, M. Effros, and T. Ho, "On source coding with coded side information for a binary source with binary side information," in *Intl. Symp. Info. Th.*, Nice, France, July 2007.
- [54] C. Heegard and A. El Gamal, "On the capacity of computer memory with defects," *IEEE Trans. on Inf. Th.*, vol. 29, no. 5, pp. 731–739, May 1983.
- [55] A. Ingber and M. Feder, "Finite blocklength coding for channels with side information at the receiver," in *Convention of Electrical and Electronics Engineers in Israel (IEEEI)*, 2010.
- [56] T. S. Han, "Hypothesis testing with multiterminal data compression," *IEEE Trans. on Inf. Th.*, vol. 33, no. 6, pp. 759–772, Jun 1987.
- [57] J. Scarlett and V. Y. F. Tan, "Second-order asymptotics for the Gaussian MAC with degraded message sets," Oct 2013. *arXiv:1310.1197v2*.
- [58] S. Boucheron and M. R. Salamatian, "About priority encoding transmission," *IEEE Trans. on Inf. Th.*, vol. 46, no. 2, pp. 699–705, 2000.
- [59] R. Renner and S. Wolf, "Simple and tight bounds for information reconciliation and privacy amplification," in *Advances in Cryptology—ASIACRYPT 2005, Lecture Notes in Computer Science, Springer-Verlag*, vol. 3788, Dec 2005, pp. 199–216.
- [60] Z. Luo and I. Devetak, "Channel simulation with quantum side information," *IEEE Trans. on Inf. Th.*, vol. 55, no. 3, pp. 1331–1342, 2009.
- [61] C. H. Bennett, P. W. Shor, J. A. Smolin, and A. V. Thapliyal, "Entanglement-assisted classical capacity of noisy quantum channels," *Phys. Rev. Lett.*, vol. 83, no. 15, pp. 3081–3084, Oct 1999.
- [62] P. Cuff and E. C. Song, "The likelihood encoder for source coding," in *2013 IEEE Information Theory Workshop*, 2013, pp. 1–2.
- [63] S. Watanabe, S. Kuzuoka, and V. Y. F. Tan, "Non-Asymptotic and Second-Order Achievability Bounds for Source Coding With Side-Information," in *2013 IEEE International Symposium on Information Theory*, 2013, pp. 3055–3059.
- [64] W. Feller, *An Introduction to Probability Theory and Its Applications*, 2nd ed. John Wiley and Sons, 1971.
- [65] B. Yu and T. P. Speed, "A rate of convergence result for a universal d-semifaithful code," *IEEE Trans. on Inf. Th.*, vol. 39, no. 3, pp. 813–820, Mar 1997.

**Shun Watanabe** (M'09) received the B.E., M.E., and Ph.D. degrees from the Tokyo Institute of Technology in 2005, 2007, and 2009, respectively. Since April 2009, he has been an Assistant Professor in the Department of Information Science and Intelligent Systems at the University of Tokushima. Since April 2013, he has also been a visiting Assistant Professor in the Institute for Systems Research at the University of Maryland, College Park. His current research interests are in the areas of information theory, quantum information theory, and quantum cryptography.

**Shigeaki Kuzuoka** (S'05-M'07) received the B.E., M.E., and Ph.D. degrees from Tokyo Institute of Technology in 2002, 2004, and 2007 respectively. He was an assistant professor from 2007 to 2009, and has been a lecturer since 2009 in the Department of Computer and Communication Sciences, Wakayama University. His current research interests are in the areas of information theory, especially Shannon theory, source coding, and multi-terminal information theory.

**Vincent Y. F. Tan** (S'07-M'11) is an Assistant Professor in the Department of Electrical and Computer Engineering (ECE) and the Department of Mathematics at the National University of Singapore (NUS). He received the B.A. and M.Eng. degrees in Electrical and Information Sciences from Cambridge University in 2005. He received the Ph.D. degree in Electrical Engineering and Computer Science (EECS) from the Massachusetts Institute of Technology in 2011. He was a postdoctoral researcher in the Department of ECE at the University of Wisconsin-Madison and following that, a research scientist at the Institute for Infocomm (I<sup>2</sup>R) Research, A\*STAR, Singapore. His research interests include information theory, machine learning and signal processing.

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