

New Second-Order Achievability Bounds for Coding with Side Information via Type Deviation Convergence

Xiang Li and Cheuk Ting Li

Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong, China

Email: lx024@ie.cuhk.edu.hk, ctli@ie.cuhk.edu.hk

Abstract

We propose a framework for second-order achievability, called type deviation convergence, that is generally applicable to settings in network information theory, and is especially suitable for lossy source coding and channel coding with cost. We give a second-order achievability bound for lossy source coding with side information at the decoder (Wyner-Ziv problem) that improves upon all known bounds (e.g., Watanabe-Kuzuoka-Tan, Yassaee-Aref-Gohari and Li-Anantharam). We also give second-order achievability bounds for lossy compression where side information may be absent (Heegard-Berger problem) and channels with noncausal state information at the encoder and cost constraint (Gelfand-Pinsker problem with cost) that improve upon previous bounds.

Index Terms

Channel dispersion, lossy source coding, Wyner-Ziv problem, Gelfand-Pinsker problem, method of types.

I. INTRODUCTION

In information theory, first-order analysis concerns the characterization of the limit R^* of the optimal coding rate $R^*(n, \epsilon)$ (e.g., the maximal message rate for channel coding, or the minimal compression rate for lossy source coding) as a function of the blocklength n and the error probability ϵ , as $n \rightarrow \infty$; whereas second-order analysis concerns a more refined characterization of how $R^*(n, \epsilon)$ approaches R^* . Most second-order results are in the form

$$R^*(n, \epsilon) = R^* + W/\sqrt{n} + o(1/\sqrt{n}),$$

where W is the coefficient of the second-order term that may depend on ϵ . For example, in channel coding, we have $W = -\sqrt{V}Q^{-1}(\epsilon)$, where $V = \text{Var}[\iota(X; Y)]$ is the channel dispersion, and Q^{-1} is the inverse of the Q -function [1], [2], [3]. In lossy source coding, we have $W = \sqrt{V}Q^{-1}(\epsilon)$, where V is the dispersion of lossy source coding [4], [5], [6] (see Section V).

Second-order analysis has also been performed on more complex settings where side information is present. An example is the Wyner-Ziv problem [7], [8], where the encoder compresses a source X^n into a description M so that the decoder who observes M and a side information Y^n (correlated with X^n , unknown to the encoder) can output a lossy reconstruction of X^n . Second-order achievability results have been given in [9], [10], [11], [12], [13]. For the lossless case, the Wyner-Ziv problem reduces to a special case of the Slepian-Wolf problem [14], where the dispersion has been characterized [15]. Nevertheless, the optimal dispersion for the general lossy Wyner-Ziv problem remains an open problem. The Wyner-Ziv problem can be generalized to the scenario where the side information may be absent [16], [17], where a second-order result was given in [10]. Also see [18], [19], [20], [21], [22], [23] for related results.

Another example is the Gelfand-Pinsker problem [24], [25], [26] about a channel coding setting where the channel depends on a state sequence known noncausally to the encoder. The current best second-order results were given in [27] (for discrete and Gaussian channels) and [12] (for general channels). A second-order result for Gelfand-Pinsker problem with a cost constraint has been derived in [11].

The method of types [28], [29], which concerns the type (or the empirical distribution) of sequences in the coding setting, is a common tool in second-order analysis. For example, it has been used in [6] for the second-order result for lossy source coding, and in [15] for the Slepian-Wolf problem, multiple-access channels and asymmetric broadcast channels. See [30], [31] and references therein for discussions on more settings in network information theory. Constant-composition codes, where each codeword has the same type, have been applied to prove second-order achievability results in multiple access channels [32], [33] and the Gelfand-Pinsker problem [27]. However, to the best of the authors' knowledge, the method of types had not been successful for lossy source coding problems with side information (e.g., Wyner-Ziv), likely due to the complexity of the analysis.

Another useful technique is the Poisson matching lemma [12] and the Poisson functional representation [34], which have been applied to derive various finite-blocklength and refined asymptotic results about coding with side information [20], [35], [36], [22], [37], [23]. Nevertheless, [12] mostly focused on i.i.d. codebooks, and it was unclear whether the Poisson matching lemma would still be useful for code constructions with a more precise control over the type. In particular, the second-order

achievability bound for the Wyner-Ziv problem in [12], which was already one of the best bounds at that time, fails to subsume the dispersion of lossy source coding without side information [4], [5], [6] as a special case.

The purpose of this paper is to propose a new definition of “well-behaved” random sequences, called *type deviation convergent sequences*, where the type deviates from its limit by an amount that has a convergent distribution. This includes i.i.d. sources (where the deviation is asymptotically Gaussian by the central limit theorem), constant-composition codes (the deviation is zero), as well as sequences with a non-Gaussian deviation. We give a small, essential toolbox of properties of type deviation convergent sequences that, together with the Poisson matching lemma, can be applied to a wide range of problems to give simple second-order achievability proofs that recover and sometimes improve upon the state of the art. We list some advantages of our approach:

- **Improving upon best known bounds.** Our approach yields second-order achievability results that improves upon the state of the art for the following problems:
 - For the Wyner-Ziv problem, our approach gives a dispersion term that significantly improves upon [9], [11], [10], [12], [35], and subsumes the achievability of the dispersion of lossy source coding without side information [4], [5], [6] as a special case¹ (Section VI). This improvement is partly due to the use of codes with non-Gaussian deviations, which differ from the usual Gaussian deviation from the central limit theorem.
 - For the indirect Wyner-Ziv problem [38], [39], our bound improves upon [22] (Section VII).
 - For lossy compression where side information may be absent (Heegard-Berger) [16], [17], our bound improves upon [10] (Section VIII).
 - For Gelfand-Pinsker coding with cost constraint, our bound improves upon [11] and [12] (after a straightforward generalization to include cost) (Section X).
- **Recovering best known bounds.** For other settings, our approach can recover state-of-the-art second-order results. This includes channel coding (with or without cost constraint) [40], [2], [3], [41], lossy source coding [4], [42], [6], indirect or noisy lossy source coding [43], [44], Gelfand-Pinsker coding [27], [12] and broadcast channels [45].
- **Simplicity.** The main novelty of our approach is to couple the source and channel input/output sequences for different blocklengths n together in the same probability space. This allows the type of a sequence to be approximated by a single random vector that does not keep changing as n increases. As a result, the proofs are shorter than conventional proofs using the method of types. See Remark 9.²
- **A unified workflow.** The proofs in this paper all follow the same unified workflow using the same sequence of tools, which can accomodate i.i.d. sequences and constant-composition codes under the same framework. This eliminates the need of devising ad-hoc arguments for each coding settings, and significantly reduces the difficulty of deriving second-order achievability results for new settings. This is in the same spirit as previous unified workflows for refined asymptotical analysis in network information theory such as [9], [10], [46], [45], [12], [35], though our approach yields better second-order bounds than previous unified approaches for settings with distortion or cost constraints.

Since the purpose of this paper is to present a new technique for achievability results, we will not discuss converse results which require completely different techniques, and hence are out of the scope of this paper. For refined asymptotical converse bounds for the Wyner-Ziv problem, see [47], [23].

II. PRELIMINARIES

A. Notations

Throughout this paper, we assume all coding settings concern discrete sources and channels with finite alphabets. Entropy is in bits, and logarithms are to the base 2. We use upper-case serif letters (e.g., X, U) for random variables, bold letters (e.g., $\mathbf{X}, \mathbf{U}$) for random processes, and san-serif letters (e.g., R, D) for non-random parameters. Write $[n] := \{1, \dots, n\}$. For a discrete random variable $X \in \mathcal{X}$, write $P_X : \mathcal{X} \rightarrow \mathbb{R}$ for its probability mass function. For random variables X, Y , write $\iota_X(x) := -\log P_X(x)$, $\iota_{Y|X}(x) := -\log P_{Y|X}(y|x)$, $\iota_{X;Y}(x; y) := \iota_Y(y) - \iota_{Y|X}(y|x)$ (we sometimes omit the subscripts and simply write $\iota(x; y)$). For a random sequence $X_1, \dots, X_n$, we say that it has an exchangeable distribution if $P_{X_1, \dots, X_n}(x_1, \dots, x_n) = P_{X_1, \dots, X_n}(x_{\pi(1)}, \dots, x_{\pi(n)})$ for any permutation π over $[n]$. For a random vector X , write $\text{Var}[X]$ for its covariance matrix. Write the Q-function as $\mathcal{Q}(x) = \mathbb{P}(X \geq x)$ where $X \sim \mathcal{N}(0, 1)$, and its inverse as $\mathcal{Q}^{-1}(t)$.

For a finite set $\mathcal{X}$, write $\mathbb{R}^{\mathcal{X}}$ for the space of real vectors with entries indexed by $\mathcal{X}$, or equivalently, the space of all functions $\mathcal{X} \rightarrow \mathbb{R}$ (we use $\mathbb{R}^{\mathcal{X}}$ and $\mathcal{X} \rightarrow \mathbb{R}$ interchangeably). For function $f : \mathcal{X} \rightarrow \mathbb{R}$, write $\|f\| := \sqrt{\sum_{x \in \mathcal{X}} (f(x))^2}$,

¹To the best of the authors’ knowledge, this is the first second-order achievability result for the Wyner-Ziv problem that subsumes the achievability of the dispersion of lossy source coding as a special case. Although the finite-blocklength bound in [11] subsumes the dispersion of lossy source coding, the second-order bound stated in [11] does not subsume the dispersion of lossy source coding.

²While the proof of the second-order result for Wyner-Ziv (Theorem 15) is short by itself, it is no longer short if we also count the proofs of the basic properties of type deviation convergent sequences in Sections III, IV. This should not be considered as an argument against the simplicity of our approach, since “factoring out” the complicated parts of the proofs to reusable basic properties can simplify subsequent proofs of new theorems. This is similar to typicality and asymptotic equipartition property, which are considered to be valuable tools for simplifying proofs in information theory, even though the basic properties of typicality often have complicated proofs.

$\|f\|_\infty := \max_{x \in \mathcal{X}} |f(x)|$. For $f, g \in \mathbb{R}^{\mathcal{X}}$, we say that f is dominated by g , denoted as $f \ll g$, if $g(x) = 0$ implies $f(x) = 0$ for $x \in \mathcal{X}$.

For a vector $x^n \in \mathcal{X}^n$, write $\hat{P}_n(x^n) \in \mathbb{R}^{\mathcal{X}}$ to be its type or empirical distribution, i.e., $\hat{P}_n(x^n)(x') = |\{i \in [n] : x_i = x'\}|/n$. Note that when $n = 1$, $\hat{P}_1(x)$ is the one-hot encoding of $x \in \mathcal{X}$. Its range $\hat{P}_n(\mathcal{X}) := \{\hat{P}_n(x^n) : x^n \in \mathcal{X}^n\} \subseteq \mathbb{R}^{\mathcal{X}}$ is the set of all probability mass functions over $\mathcal{X}$ where each entry is a multiple of $1/n$. We sometimes omit the subscript and simply write $\hat{P}$ if n is clear in the context.

For sets $\mathcal{A}_1, \mathcal{A}_2 \subseteq \mathbb{R}^{\mathcal{X}}$ and $c \in \mathbb{R}$, write the Minkowski sum as $\mathcal{A}_1 + \mathcal{A}_2 := \{a_1 + a_2 : a_1 \in \mathcal{A}_1, a_2 \in \mathcal{A}_2\}$, and $c\mathcal{A}_1 := \{ca_1 : a_1 \in \mathcal{A}_1\}$. Write $\mathcal{B}^{\mathcal{X}} := \{x \in \mathbb{R}^{\mathcal{X}} : \|x\| < 1\}$ for the open ℓ_2 ball. For two probability distributions P, Q over a vector space $\mathbb{R}^{\mathcal{X}}$, their *Lévy-Prokhorov distance* is [48]

$$d_{\Pi}(P, Q) := \inf \left\{ \epsilon : P(\mathcal{A}) \leq Q(\mathcal{A} + \epsilon \mathcal{B}^{\mathcal{X}}) + \epsilon, \right. \\ \left. \forall \text{ closed } \mathcal{A} \subseteq \mathbb{R}^{\mathcal{X}} \right\}. \quad (1)$$

For two random vectors $X, Y \in \mathbb{R}^{\mathcal{X}}$, their *Ky Fan distance* [49] is defined as

$$d_{\text{KF}}(X, Y) := \inf \{ \epsilon : \mathbb{P}(\|X - Y\| > \epsilon) \leq \epsilon \}. \quad (2)$$

These two metrics are related via the *Strassen-Dudley theorem* [49], [50, Theorem 6.9], which states that $d_{\Pi}(P, Q) = \inf_{X \sim P, Y \sim Q} d_{\text{KF}}(X, Y)$ (the infimum is over couplings (X, Y) of P and Q).

B. Distributions and Perturbations

We now define some notations that are useful for studying discrete distributions and their perturbations. Write $\Delta(\mathcal{X}) := \{p \in \mathbb{R}^{\mathcal{X}} : \min_x p(x) \geq 0, \sum_x p(x) = 1\}$ for the set of probability mass functions over $\mathcal{X}$. Write $\Delta(\mathcal{Y}|\mathcal{X}) := \{p(\cdot|\cdot) \in \mathbb{R}^{\mathcal{X} \times \mathcal{Y}} : p(\cdot|x) \in \Delta(\mathcal{Y}), \forall x\}$ for the set of conditional probability mass functions from $\mathcal{X}$ to $\mathcal{Y}$, where $p(\cdot|x)$ denotes the function $y \mapsto p(y|x)$. Note that if we write $p(\cdot|\cdot) \in \mathbb{R}^{\mathcal{X} \times \mathcal{Y}}$, then the function $p : \mathbb{R}^{\mathcal{X} \times \mathcal{Y}} \rightarrow \mathbb{R}$ is being used as $p(y|x)$ (instead of $p(x, y)$) to conform with the usual notation of conditional distributions (we use $\mathbb{R}^{\mathcal{X} \times \mathcal{Y}}$ instead of $\mathbb{R}^{\mathcal{Y} \times \mathcal{X}}$, despite y being written before x in $p(y|x)$, because x is usually generated before y). For $p \in \Delta(\mathcal{X})$, write

$$\text{Tan}(p) := \{v \in \mathbb{R}^{\mathcal{X}} : v \ll p, \sum_x v(x) = 0\}.$$

Note that $\text{Tan}(p)$ is the tangent space of $\Delta(\mathcal{X})$ at p , in the sense that there exists $\epsilon > 0$ such that $p + v \in \Delta(\mathcal{X})$ for every $v \in \text{Tan}(p)$ with $\|v\| < \epsilon$. Hence, we can understand $\text{Tan}(p)$ as the “space of small perturbations” of the distribution p . Write $\text{Tan}(\mathcal{X}) := \{v \in \mathbb{R}^{\mathcal{X}} : \sum_x v(x) = 0\}$. Similarly, for $p \in \Delta(\mathcal{Y}|\mathcal{X})$, write $\text{Tan}(p) := \{v \in \mathbb{R}^{\mathcal{X} \times \mathcal{Y}} : v \ll p, \sum_y v(y|x) = 0, \forall x\}$.

For a function $f : \mathcal{X} \rightarrow [0, \infty)$, write $\sqrt{f} : \mathcal{X} \rightarrow [0, \infty)$ for the function $x \mapsto \sqrt{f(x)}$. For functions $f : \mathcal{X} \rightarrow \mathbb{R}$ and $g : \mathcal{Y} \rightarrow \mathbb{R}$, denote their product as $f \times g : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$, $(f \times g)(x, y) := f(x)g(y)$. For functions $f : \mathcal{X} \rightarrow \mathbb{R}$ and $g : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$, denote their *semidirect product* as $f \circ g : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$,

$$(f \circ g)(x, y) := f(x)g(x, y).$$

For example, for random variables $(X, Y) \sim P_{X,Y}$, we have $P_X \circ P_{Y|X} = P_{X,Y}$. If some parts of the domain of f is not present in the domain of g , then the domain of g is suitably extended to include those parts. For example, if we have $f : \mathcal{X} \times \mathcal{Z} \rightarrow \mathbb{R}$ instead where the part $\mathcal{Z}$ is not present in the domain of $g : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$, then we extend g to $g : \mathcal{X} \times \mathcal{Z} \times \mathcal{Y} \rightarrow \mathbb{R}$, $g(x, z, y) = g(x, y)$, so $(f \circ g)(x, z, y) = f(x, z)g(x, y)$. Note that $P_{X,Z} \circ P_{Y|X} = P_{X,Z,Y}$ if $Z \leftrightarrow X \leftrightarrow Y$ forms a Markov chain. For functions $f, g : \mathcal{X} \rightarrow \mathbb{R}$ write

$$\langle f, g \rangle := \sum_{x \in \mathcal{X}} f(x)g(x).$$

If the domain of f is larger than the domain of g , then the domain of g is suitably extended to match the domain of f (e.g., if we have $f : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$, then we extend g to $g(x, y) = g(x)$ so $\langle f, g \rangle = \sum_{x,y} f(x, y)g(x)$). Note that if $(X, Y) \sim P_{X,Y}$, then $\langle P_{X,Y}, g \rangle = \mathbb{E}[g(X)]$, so $\langle f, g \rangle$ is a generalization of expectation where f may not be a probability mass function.³

To demonstrate the use of these notations, note that the derivative of entropy $H(P_X)$ along the direction $V \in \text{Tan}(P_X)$ is

$$\left. \frac{dH(P_X + tV)}{dt} \right|_{t=0} = \langle V, \iota_X \rangle, \quad (3)$$

where $\iota_X : \mathcal{X} \rightarrow \mathbb{R}$, $\iota_X(x) = -\log P_X(x)$ is the self-information. As a result, the derivative of $I(X; Y)$ when $(X, Y) \sim P_{X,Y} + tV$ along the direction $V \in \text{Tan}(P_{X,Y})$ is $\langle V, \iota_{X;Y} \rangle$.

³Similar notations have appeared, for example, in [51].

C. Gaussian-Multinomial Distribution

The main technique in this paper is to approximate the distribution of the type $\hat{P}(X^{(n)})$ of a random sequence $X^{(n)} \in \mathcal{X}^n$ by a Gaussian distribution. We first consider a simple example where $X^{(n)} \sim P_X^n$ is an i.i.d. sequence. Since $\hat{P}(X^{(n)}) = n^{-1} \sum_{i=1}^n \hat{P}_1(X_i^{(n)})$, where $\hat{P}_1(X_i^{(n)}) \in \mathbb{R}^{\mathcal{X}}$ is the one-hot encoding of $X_i^{(n)} \in \mathcal{X}$, we know that $n\hat{P}(X^{(n)})$ follows a multinomial distribution, and $\sqrt{n}(\hat{P}(X^{(n)}) - P_X)$ is approximately Gaussian with mean 0 and covariance matrix $\text{Var}[\hat{P}_1(X)]$ where $X \sim P_X$, by the central limit theorem. This distribution is defined as follows. It is a special case of the Gaussian-multinomial distribution in [52], and hence we adopt this name.

Definition 1 (Gaussian-multinomial distribution). For a probability mass function $P_X \in \Delta(\mathcal{X})$, denote

$$\text{NM}(P_X) := \text{N}(\mathbf{0}, \Sigma)$$

to be the multivariate Gaussian distribution over $\text{Tan}(P_X) \subseteq \mathbb{R}^{\mathcal{X}}$ with covariance matrix $\Sigma(x, x) = P_X(x)(1 - P_X(x))$, $\Sigma(x, x') = -P_X(x)P_X(x')$ for $x \neq x'$. Equivalently, $\Sigma = \text{Var}[\hat{P}_1(X)]$ where $X \sim P_X$, and $\hat{P}_1(X) \in \mathbb{R}^{\mathcal{X}}$ is the one-hot encoding of X .

We now define a conditional version of this distribution, which consists of independent Gaussian-multinomial vectors stacked together to form a matrix.

Definition 2 (Conditional Gaussian-multinomial distribution). For a conditional probability mass function $P_{Y|X} \in \Delta(\mathcal{Y}|\mathcal{X})$, denote $\text{NM}(P_{Y|X})$ to be the distribution of $(V_x)_{x \in \mathcal{X}} \in \text{Tan}(P_{Y|X}) \subseteq \mathbb{R}^{\mathcal{X} \times \mathcal{Y}}$ (i.e., stacking V_x for $x \in \mathcal{X}$ together), where $V_x \in \mathbb{R}^{\mathcal{Y}}$, $V_x \sim \text{NM}(P_{Y|X}(\cdot|x))$ are independent across $x \in \mathcal{X}$.

We now state some simple facts about $\text{NM}(P_X)$ and $\text{NM}(P_{Y|X})$. The proofs are immediate and hence omitted.

Proposition 3. Let $(X, Y) \sim P_{X,Y} = P_X \circ P_{Y|X}$. Let $G_X \sim \text{NM}(P_X)$ and $G_{Y|X} \sim \text{NM}(P_{Y|X})$ be independent. We have the following:

- For $f : \mathcal{X} \rightarrow \mathbb{R}$,

$$\langle G_X, f \rangle \sim \text{N}(0, \text{Var}[f(X)]). \quad (4)$$

-

$$G_X \circ P_{Y|X} + \sqrt{P_X} \circ G_{Y|X} \sim \text{NM}(P_{X,Y}).$$

- For $f : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$,

$$\langle G_X \circ P_{Y|X}, f \rangle \sim \text{N}(0, \text{Var}[\mathbb{E}[f(X, Y)|X]]),$$

$$\langle \sqrt{P_X} \circ G_{Y|X}, f \rangle \sim \text{N}(0, \mathbb{E}[\text{Var}[f(X, Y)|X]]).$$

Note that $\text{Var}[\mathbb{E}[f(X, Y)|X]] + \mathbb{E}[\text{Var}[f(X, Y)|X]] = \text{Var}[f(X, Y)]$ by the law of total variance. Hence, $G_X \circ P_{Y|X} + \sqrt{P_X} \circ G_{Y|X}$ decomposes the randomness in $\text{NM}(P_{X,Y})$ into two parts: the part $G_X \circ P_{Y|X}$ that comes from the randomness in X , and the part $\sqrt{P_X} \circ G_{Y|X}$ from the randomness in Y .

III. TYPE DEVIATION CONVERGENCE

We now introduce a framework, called *type deviation convergence*, which is a collection of notations and results that simplifies second-order analyses. We consider families of random sequences in the form $\mathbf{X} = (X^{(n)})_{n \in \mathbb{N}}$, where $X^{(n)} \in \mathcal{X}^n$ is a random sequence and $\mathcal{X}$ is finite. Such an $\mathbf{X}$ is called a *general source* [53]. Note that we do not require $X^{(n)}$ to be the prefix of $X^{(n+1)}$, i.e., $X_i^{(n)}$ (the i -th entry of $X^{(n)}$) may not equal $X_i^{(n+1)}$. This is the reason we use the notation $X^{(n)}$ instead of X^n .

We focus on general sources with a type $\hat{P}(X^{(n)})$ that can be approximated by a distribution P_X in the sense that $\hat{P}(X^{(n)}) - P_X$ is $O(n^{-1/2})$, and $\sqrt{n}(\hat{P}(X^{(n)}) - P_X)$ converges to a subgaussian random vector. This is captured by the following definition.

Definition 4 (Type deviation convergence). For a general source $\mathbf{X} = (X^{(n)})_{n \in \mathbb{N}}$, we say that it is *type deviation convergent* if $X^{(n)}$ has an exchangeable distribution, and there exists a probability vector $P_X \in \Delta(\mathcal{X})$ and a subgaussian⁴ random vector $G_{\mathbf{X}} \in \text{Tan}(P_X)$ such that $\hat{P}(X^{(n)}) \ll P_X$ almost surely, and $G_{\mathbf{X}}^{(n)} := \sqrt{n}(\hat{P}(X^{(n)}) - P_X)$ approaches $G_{\mathbf{X}}$ in the order $O(n^{-1/2})$ with respect to the Ky Fan distance (2), that is, there exists $c > 0$ such that

$$\mathbb{P}\left(\|G_{\mathbf{X}}^{(n)} - G_{\mathbf{X}}\| > \frac{c}{\sqrt{n}}\right) < \frac{c}{\sqrt{n}},$$

⁴A random vector $G \in \mathbb{R}^{\mathcal{X}}$ is subgaussian if there exists $\eta > 0$ such that $\mathbb{P}(\|G\| \geq t) \leq 2 \exp(-t^2/\eta^2)$ for all $t \geq 0$.

for all n . If this is satisfied, we say that $\mathbf{X}$ is type deviation convergent with *center* P_X and *asymptotic deviation* $G_{\mathbf{X}}$.

For two general sources $\mathbf{X}, \mathbf{Y}$, we say that $(\mathbf{X}, \mathbf{Y})$ is (jointly) type deviation convergent if $((X_i^{(n)}, Y_i^{(n)})_{i \in [n]})_{n \in \mathbb{N}}$ is type deviation convergent (which implies that $\mathbf{X}$ and $\mathbf{Y}$ are both type deviation convergent).

Note that we do not require $G_{\mathbf{X}}$ to be Gaussian or have zero mean. Definition 4 allows many interesting statistics of $\mathbf{X}$ to be summarized by the first-order term P_X and the second-order term $G_{\mathbf{X}}$. We remark that the $X^{(n)}$'s for different n 's are assumed to be coupled together (i.e., defined in the same probability space), so we can talk about the limit $G_{\mathbf{X}}$. This is a main novelty in our approach which significantly simplifies the proofs, compared to previous achievability proofs in information theory where the source sequences for different blocklengths are effectively in separate probability spaces.⁵ Refer to Remark 9 for further discussions on why a coupling is desirable.

A. Examples

The “prototypical example” of a type deviation convergent source is the i.i.d. memoryless source in which the entries of $X^{(n)}$ are i.i.d. following P_X , where $G_{\mathbf{X}} \sim \text{NM}(P_X)$ is a Gaussian-multinomial random variable. Nevertheless, the memoryless source is not technically type deviation convergent since the $X^{(n)}$'s for different n 's are not defined in the same probability space.⁶ Therefore, it is usually more reasonable to ask whether there exists a coupling of the $X^{(n)}$'s that is type deviation convergent. We say that $\tilde{\mathbf{X}} = (\tilde{X}^{(n)})_n$ is a coupling of $\mathbf{X} = (X^{(n)})_n$ if the $\tilde{X}^{(n)}$'s for $n \in \mathbb{N}$ are defined in the same probability space, and $\tilde{X}^{(n)}$ has the same marginal distribution as $X^{(n)}$. The following proposition gives the condition for a type deviation convergent coupling to exist.

Proposition 5. *For a general source $\mathbf{X} = (X^{(n)})_n$, there exists a type deviation convergent coupling if and only if $X^{(n)}$ has an exchangeable distribution, and there exists $P_X \in \Delta(\mathcal{X})$ and a subgaussian random vector $G_{\mathbf{X}} \in \text{Tan}(P_X)$ such that $\hat{P}(X^{(n)}) \ll P_X$ almost surely, and the distribution of $G_{\mathbf{X}}^{(n)} = \sqrt{n}(\hat{P}(X^{(n)}) - P_X)$ approaches the distribution of $G_{\mathbf{X}}$ in the order $O(n^{-1/2})$ with respect to the Lévy-Prokhorov distance (1), i.e., there exists $c > 0$ such that*

$$d_{\Pi}(G_{\mathbf{X}}^{(n)}, G_{\mathbf{X}}) \leq cn^{-1/2}.$$

Proof: For the “only if” part, if $\tilde{\mathbf{X}} = (\tilde{X}^{(n)})_n$ is a type deviation convergent coupling, then for any closed $\mathcal{A} \subseteq \text{Tan}(P_X)$, we have

$$\mathbb{P}(G_{\tilde{\mathbf{X}}} \in \mathcal{A}) \leq \mathbb{P}(G_{\tilde{\mathbf{X}}}^{(n)} \in \mathcal{A} + cn^{-1/2}\mathcal{B}) + cn^{-1/2},$$

where $\mathcal{B}$ is the open unit ball in $\text{Tan}(P_X)$, and hence $d_{\Pi}(G_{\mathbf{X}}^{(n)}, G_{\tilde{\mathbf{X}}}) = d_{\Pi}(G_{\tilde{\mathbf{X}}}^{(n)}, G_{\tilde{\mathbf{X}}}) \leq cn^{-1/2}$. For the “if” part, by the Strassen-Dudley theorem [50, Theorem 6.9], there exists a coupling of $G_{\mathbf{X}}$ and $G_{\mathbf{X}}^{(n)}$ with a Ky Fan distance (2) $d_{\text{KF}}(G_{\mathbf{X}}^{(n)}, G_{\mathbf{X}}) \leq 2cn^{-1/2}$. Hence, $(X^{(n)})_n$ can be coupled with $G_{\mathbf{X}}$ such that $d_{\text{KF}}(G_{\mathbf{X}}^{(n)}, G_{\mathbf{X}}) \leq 2cn^{-1/2}$ holds for all n . ■

We can now show that a type deviation convergent coupling exists for an i.i.d. source $\mathbf{X}$.

Proposition 6. *If $X^{(n)} \sim P_X^n$ (the n -fold i.i.d. distribution), then $\mathbf{X}$ has a type deviation convergent coupling with center P_X and asymptotic deviation being a Gaussian-multinomial random variable $G_{\mathbf{X}} \sim \text{NM}(P_X)$ (Definition 1).*

Proof: Recall that $G_{\mathbf{X}}^{(n)} = \sqrt{n}(\hat{P}(X^{(n)}) - P_X)$. We have $\hat{P}(X^{(n)}) = n^{-1} \sum_{i=1}^n \hat{P}_1(X_i)$, where $\hat{P}_1(X_i)$ are i.i.d. with $\mathbb{E}[\hat{P}_1(X_i)] = P_X$. By the result in [54] on the $O(n^{-1/2})$ convergence in the central limit theorem with respect to d_{Π} , there exists a constant c such that $d_{\Pi}(G_{\mathbf{X}}^{(n)}, G_{\mathbf{X}}) \leq cn^{-1/2}$. The result follows from Proposition 5. ■

More generally, if we pass a type deviation convergent source through a memoryless channel, then the source and the output can be coupled to be jointly type deviation convergent. The proof is in Appendix B.

Proposition 7 (Type deviation convergence of memoryless channels). *Assume $\mathbf{X}$ is type deviation convergent with center P_X and asymptotic deviation $G_{\mathbf{X}}$, and $P_{Y|X}$ is a conditional distribution from $\mathcal{X}$ to $\mathcal{Y}$. Then there exists a source $\mathbf{Y}$ such that $Y^{(n)}|X^{(n)} \sim P_{Y|X}^n$ (i.e., the conditional marginal distribution of $Y^{(n)}$ given $X^{(n)}$ is as if $Y^{(n)}$ is the output when $X^{(n)}$ is*

⁵In a conventional proof of the source coding theorem, the source $X^{(n)} \in \mathcal{X}^n$ is either assumed to be in different probability spaces for different n 's, or assumed to be the first n symbols of the same infinite sequences $X_1, X_2, \dots$ (and hence $X^{(n)}$'s are “coupled” together). It makes no difference which of these two definitions is used, and hence the $X^{(n)}$'s are “effectively in separate probability spaces”. In our approach, the coupling of $X^{(n)}$ is important, and the aforementioned infinite sequence “coupling” will not work.

⁶If we put them in the same probability space simply by taking $X^{(n)} = (X_1, \dots, X_n)$ where $X_i \stackrel{\text{iid}}{\sim} P_X$, then it is not type deviation convergent since $G_{\mathbf{X}}^{(n)} = \sqrt{n}(\hat{P}(X^{(n)}) - P_X)$ does not converge by Donsker's theorem.

passed through the memoryless channel $P_{Y|X}$, and $(\mathbf{X}, \mathbf{Y})$ is type deviation convergent with center $P_X \circ P_{Y|X}$ and asymptotic deviation

$$G_{\mathbf{X}, \mathbf{Y}} = G_{\mathbf{X}} \circ P_{Y|X} + \sqrt{P_X} \circ G_{\mathbf{Y}|\mathbf{X}},$$

where $G_{\mathbf{Y}|\mathbf{X}} \sim \text{NM}(P_{Y|X})$ is independent of $G_{\mathbf{X}}$.

B. Average Properties

To see why the interesting statistics of a type deviation convergent source $\mathbf{X}$ are summarized by P_X and $G_{\mathbf{X}}$, first we note that for any function $f : \mathcal{X} \rightarrow \mathbb{R}$, we have

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n f(X_i^{(n)}) &= \langle \hat{P}(X^{(n)}), f \rangle \\ &= \langle P_X, f \rangle + \frac{1}{\sqrt{n}} \langle G_{\mathbf{X}}, \iota_X \rangle + O\left(\frac{1}{n}\right) \end{aligned} \quad (5)$$

with probability $1 - O(n^{-1/2})$,⁷ which follows directly from the definition. Next, we show that the self-information $\iota_{X^{(n)}}(X^{(n)}) = -\log P_{X^{(n)}}(X^{(n)})$ can be approximated as

$$\iota_{X^{(n)}}(X^{(n)}) = nH(X) + \sqrt{n} \langle G_{\mathbf{X}}, \iota_X \rangle + O(\log n) \quad (6)$$

with probability $1 - O(n^{-1/2})$. The proof is in Appendix A.

Proposition 8. Assume $\mathbf{X}$ is type deviation convergent with center P_X and asymptotic deviation $G_{\mathbf{X}}$. Then

$$\mathbb{P} \left(\left| nH(X) + \sqrt{n} \langle G_{\mathbf{X}}, \iota_X \rangle - \iota_{X^{(n)}}(X^{(n)}) \right| > c \log n \right) \leq \frac{c}{\sqrt{n}},$$

for some constant c that only depends on the distribution of $\mathbf{X}$, where $H(X)$ and ι_X are computed using P_X .

As a result, if $(\mathbf{X}, \mathbf{Y})$ is type deviation convergent with center $P_{X,Y}$ and asymptotic deviation $G_{\mathbf{X}, \mathbf{Y}}$, applying Proposition 8 on $\mathbf{X}, \mathbf{Y}$ and $(\mathbf{X}, \mathbf{Y})$, we have

$$\begin{aligned} \iota_{X^{(n)}; Y^{(n)}}(X^{(n)}; Y^{(n)}) \\ = nI(X; Y) + \sqrt{n} \langle G_{\mathbf{X}, \mathbf{Y}}, \iota_{X; Y} \rangle + O(\log n) \end{aligned} \quad (7)$$

with probability $1 - O(n^{-1/2})$. As we will see, for many settings, sample averages $n^{-1} \sum_{i=1}^n f(X_i^{(n)})$ (e.g., cost and distortion constraints) and combinations of $\iota_{X^{(n)}}(X^{(n)})$ are all the statistics we need to prove coding theorems.

Remark 9. The use of coupling in Definition 4 is a main novelty of our technique. Alternatively, one might simply take Proposition 5 as the definition of type deviation convergent sources, which will not require $X^{(n)}$'s to be coupled together. This alternative definition is more similar to conventional proofs based on the method of types. However, it comes with a significant downside that we can no longer have simple statements like (5) and (6), since it is no longer reasonable to talk about convergence of random variables as $n \rightarrow \infty$ (we can only talk about convergence of distributions). We argue that this is a main source of complexity in previous proofs, and might have been a reason why the method of types has not been applied to more complicated settings, e.g., Heegard-Berger and the broadcast channel, where our approach can be applied with relative ease (Theorems 20 and 23). The proofs in this paper (e.g., Theorems 14 and 15) are comparatively simpler (and also stronger).

IV. GENERAL CONSTANT-COMPOSITION CODES

Constant-composition code is a common technique for proving second-order achievability results [32], [33], [27]. In this paper, we consider a general definition of constant-composition codes, where the conditional type of the output $U^{(n)}$ can depend on the type of the source $X^{(n)}$ in a general manner through a “deviation function” $\zeta : \text{Tan}(P_X) \rightarrow \text{Tan}(P_{U|X})$. Intuitively, when $\mathbf{X}$ is type deviation convergent, we want the conditional type of $U^{(n)}$ given $X^{(n)}$ to be approximately $\zeta(G_{\mathbf{X}})$, which can depend on $G_{\mathbf{X}}$. Note that constructions where the type of $U^{(n)}$ depend of the type of $X^{(n)}$ have appeared in previous proofs (e.g., [51]), though the level of generality in Definition 10, which allows ζ to be nonlinear, appears to be novel.

Definition 10 (General constant-composition (GCC) channel). Given a joint distribution $P_{X,U} = P_X \circ P_{U|X}$ and a Lipschitz continuous function $\zeta : \text{Tan}(P_X) \rightarrow \text{Tan}(P_{U|X})$, we say that a family of conditional distributions $(P_{U^{(n)}|X^{(n)}})_n$ (from $\mathcal{X}^n$ to $\mathcal{U}^n$) for $n \in \mathbb{N}$ is *general constant-composition* with center $P_X \circ P_{U|X}$ and deviation function ζ if $P_{U^{(n)}|X^{(n)}}$ is jointly

⁷“ $X = Y + O(1/n)$ with probability $1 - O(n^{-1/2})$ ” means that there exists $c > 0$ such that $\mathbb{P}(|X - Y| \leq c/n) \geq 1 - cn^{-1/2}$.

exchangeable (i.e., $P_{U^{(n)}|X^{(n)}}(u^n|x^n) = P_{U^{(n)}|X^{(n)}}((u_{\pi(i)})_{i \in [n]}|(x_{\pi(i)})_{i \in [n]})$ for every permutation $\pi : [n] \rightarrow [n]$), and there exists $c > 0$ such that for every $n \geq c$ and $x^n \in \mathcal{X}^n$ with $\hat{P}(x^n) \ll P_X$ and $\|\hat{P}(x^n) - P_X\| \leq 1/c$, and for every u^n with $P_{U^{(n)}|X^{(n)}}(u^n|x^n) > 0$, we must have $\hat{P}(x^n, u^n) \ll P_{X,U}$ and

$$\|G_{x^n}^{(n)} \circ P_{U|X} + P_X \circ \zeta(G_{x^n}^{(n)}) - G_{x^n, u^n}^{(n)}\| \leq cn^{-1/2},$$

where

$$\begin{aligned} G_{x^n}^{(n)} &:= \sqrt{n}(\hat{P}(x^n) - P_X), \\ G_{x^n, u^n}^{(n)} &:= \sqrt{n}(\hat{P}(x^n, u^n) - P_{X,U}). \end{aligned}$$

We first prove the existence of a GCC channel. The proof is in Appendix C.

Proposition 11. *For every $P_{X,U}$ and Lipschitz continuous function $\zeta : \text{Tan}(P_X) \rightarrow \text{Tan}(P_{U|X})$, there exists a GCC channel $(P_{U^{(n)}|X^{(n)}})_n$.*

Next, we show that passing a type deviation convergent source as the input to a GCC channel results in an output that is jointly type deviation convergent with the input.

Proposition 12 (Type deviation convergence of GCC channels). *Assume $\mathbf{X}$ is type deviation convergent with center P_X and asymptotic deviation $G_{\mathbf{X}}$, and $(P_{U^{(n)}|X^{(n)}})_n$ is a GCC channel with center $P_{X,U} = P_X \circ P_{U|X}$ and deviation function ζ . Let $\mathbf{U} = (U^{(n)})_n$ be random sequences such that $U^{(n)}$ follows $P_{U^{(n)}|X^{(n)}}$ given $X^{(n)}$. Then $(\mathbf{X}, \mathbf{U})$ is type deviation convergent with center $P_{X,U}$ and asymptotic deviation*

$$G_{\mathbf{X}, \mathbf{U}} = G_{\mathbf{X}} \circ P_{U|X} + P_X \circ \zeta(G_{\mathbf{X}}).$$

Proof: Recall that $G_{\mathbf{X}}^{(n)} = \sqrt{n}(\hat{P}(X^{(n)}) - P_X)$. Let c be the maximum of the constant in Definition 4 for $\mathbf{X}$, and the constant in Definition 10 for $(P_{U^{(n)}|X^{(n)}})_n$. Since $\mathbf{X}$ is type deviation convergent, we have $\|\hat{P}(X^{(n)}) - P_X\| \leq 1/c$ with probability $1 - O(n^{-1/2})$. Consider

$$\begin{aligned} G_{\mathbf{X}, \mathbf{U}}^{(n)} &:= \sqrt{n}(\hat{P}(X^{(n)}, U^{(n)}) - P_{X,U}), \\ \hat{G}_{\mathbf{X}, \mathbf{U}}^{(n)} &:= G_{\mathbf{X}}^{(n)} \circ P_{U|X} + P_X \circ \zeta(G_{\mathbf{X}}^{(n)}). \end{aligned}$$

By Definition 10, $\|G_{\mathbf{X}, \mathbf{U}}^{(n)} - \hat{G}_{\mathbf{X}, \mathbf{U}}^{(n)}\| \leq cn^{-1/2}$ with probability $1 - O(n^{-1/2})$. We also have $\|\hat{G}_{\mathbf{X}, \mathbf{U}}^{(n)} - G_{\mathbf{X}, \mathbf{U}}\| = O(n^{-1/2})$ with probability $1 - O(n^{-1/2})$ since ζ is Lipschitz and $\|G_{\mathbf{X}}^{(n)} - G_{\mathbf{X}}\| \leq cn^{-1/2}$ with probability $1 - O(n^{-1/2})$. The result follows. $\blacksquare$

We remark that if $G_{\mathbf{X}}$ is Gaussian and ζ is an affine function, then $G_{\mathbf{X}, \mathbf{U}}$ is also Gaussian. Nevertheless, we do not require ζ to be affine, so $G_{\mathbf{X}, \mathbf{U}}$ may not be Gaussian even if $G_{\mathbf{X}}$ is Gaussian. Non-affine ζ is useful for the Wyner-Ziv problem in Theorem 15.

V. SECOND-ORDER LOSSY SOURCE CODING

We will now utilize type deviation convergence to prove coding theorems. Before we proceed to the theorems, we state the Poisson matching lemma introduced in [12], which is another basic tool for constructing coding schemes via exponential random variables that we will use throughout this paper.⁸

Lemma 13 (Poisson matching lemma on X, U, Y [12]). *Consider a joint distribution $P_{X,U,Y} = P_X \circ P_{U|X} \circ P_{Y|X,U}$. Let $T_u \sim \text{Exp}(1)$, i.i.d. across $u \in \mathcal{U}$. Assume $X \sim P_X$, $U = \arg\min_u T_u / P_{U|X}(u|X)$, $Y|(X, U) \sim P_{Y|X,U}$, and $\hat{U} = \arg\min_u T_u / P_{U|Y}(u|Y)$. Then $(X, U, Y) \sim P_{X,U,Y}$, and*

$$\mathbb{P}(U \neq \hat{U} | X, U, Y) \leq 2^{\iota_{U;X}(U;X) - \iota_{U;Y}(U;Y)} \quad (8)$$

holds almost surely. As a result, for any error event E that depends only on (X, U, Y) and any $\gamma \in \mathbb{R}$,

$$\begin{aligned} &\mathbb{P}(E \text{ or } U \neq \hat{U}) \\ &\leq \mathbb{P}(E \text{ or } \iota_{U;X}(U;X) - \iota_{U;Y}(U;Y) > -\gamma) + 2^{-\gamma}. \end{aligned}$$

⁸[12] proved the Poisson matching lemma for the general case where X, U, Y may not be discrete, which involve Poisson processes. Here, we focus on the discrete case, which has a simpler statement and proof in [12].

We first demonstrate the use of type deviation convergence and the Poisson matching lemma, by recovering the second-order result on lossy source coding for a discrete memoryless source [4], [6], [42]. In this setting, the encoder encodes a discrete memoryless source $X^{(n)} \sim P_X^n$ into a message $M \in [[2^{nR}]]$ with rate $R > 0$. The decoder observes M and recovers $\hat{Y}^{(n)} \in \mathcal{Y}^n$. The goal is to minimize the probability of excess distortion

$$P_e := \mathbb{P}(d(X^{(n)}, \hat{Y}^{(n)}) > D),$$

where $d : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ is a distortion function, $d(X^{(n)}, \hat{Y}^{(n)}) := n^{-1} \sum_{i=1}^n d(X_i^{(n)}, \hat{Y}_i^{(n)})$, and $D \in \mathbb{R}$ is the allowed distortion level. The optimal asymptotic rate needed to have $P_e \rightarrow 0$ is given by the rate-distortion function $R(D) := \min_{P_{Y|X} : \mathbb{E}[d(X, Y)] \leq D} I(X; Y)$ [55]. We now recover the second-order result in [4], [42], [6], given in terms of the d -tilted information [4], [42]

$$\begin{aligned} J_{X,D}(x) &:= -\log \mathbb{E} \left[2^{-\lambda(d(x, Y^*) - D)} \right] \\ &\stackrel{(a)}{=} \iota_{X;Y}(x; y) + \lambda(d(x, y) - D) \\ &\stackrel{(b)}{=} \mathbb{E}[\iota_{X;Y}(X; Y) + \lambda(d(X, Y) - D) \mid X = x], \end{aligned} \quad (9)$$

where $\lambda := -dR(D)/dD$, we assume $P_{Y|X}$ is the distribution attaining $R(D)$, and $Y^* \sim P_Y$ (the marginal distribution of Y in $P_X \circ P_{Y|X}$), (a) holds for P_Y -almost all y 's [56], and (b) is assuming $(X, Y) \sim P_X \circ P_{Y|X}$.

Theorem 14. *For discrete lossy source coding, if $D > \min_{P_{Y|X}} \mathbb{E}[d(X, Y)]$, then for any $0 < \epsilon < 1$ and any large enough n , there is a scheme achieving a probability of excess distortion $P_e \leq \epsilon$, with rate*

$$R = R(D) + \sqrt{\frac{V}{n}} \mathcal{Q}^{-1}(\epsilon) + O\left(\frac{\log n}{n}\right),$$

where $V := \text{Var}[J_{X,D}(X)]$, and the constant in $O((\log n)/n)$ depends on P_X , d , D and ϵ .

We give an intuitive explanation of Theorem 14. Loosely speaking, there are two error events in lossy source coding: excess distortion $d(X^{(n)}, Y^{(n)}) > D$, and excess information $n^{-1} \iota(X^{(n)}; Y^{(n)}) > R$ (the nR -bit message is insufficient to store $Y^{(n)}$). In case if $(X^{(n)}, Y^{(n)}) \stackrel{\text{iid}}{\sim} P_X \circ P_{Y|X}$, then by the central limit theorem, the vector $[n^{-1} \iota(X^{(n)}; Y^{(n)}), d(X^{(n)}, Y^{(n)})]^\top$ will be approximately Gaussian with covariance matrix

$$\begin{aligned} &\frac{1}{n} \text{Var} \left[\begin{bmatrix} \iota \\ d \end{bmatrix} \right] \\ &= \frac{1}{n} \text{Var} \left[\mathbb{E} \left[\begin{bmatrix} \iota \\ d \end{bmatrix} \mid X \right] \right] + \frac{1}{n} \mathbb{E} \left[\text{Var} \left[\begin{bmatrix} \iota \\ d \end{bmatrix} \mid X \right] \right], \end{aligned} \quad (10)$$

due to the law of total covariance, where $\iota := \iota(X; Y)$, $d := d(X, Y)$. The first term and the second term above are the contribution of the randomness in $X^{(n)}$ (an observable but uncontrollable component of the deviation) and $Y^{(n)}$ (a controllable component of the deviation), respectively. Error occurs if one of the two coordinates of the random vector is too large; see Figure 1 (left). The second term is rank-one since $\iota(X; Y) + \lambda d(X, Y)$ depends only on X by 9. The intuitive reason is that if changing Y can affect $\iota + \lambda d$, then we can perturb $P_{Y|X}$ to reduce $I(X; Y) + \lambda \mathbb{E}[d(X, Y)]$, contradicting the optimality of $P_{Y|X}$.

To reduce the error probability, note that the encoder has full control of $Y^{(n)}$, and there is no reason to randomize $Y^{(n)}$ and increase the error probability. The second term in (10) can be eliminated via a constant composition code for $Y^{(n)}$. Moreover, we can control the deviation of the type of $Y^{(n)}$ according to the type of $X^{(n)}$ in order to combine the two error events into one. Since $\iota(X; Y) + \lambda d(X, Y)$ depends only on X , controlling Y can only allow trading off $\iota(X; Y)$ and $d(X, Y)$ along a diagonal line with slope $-1/\lambda$. We can move the point $[\iota, d]^\top$ along this diagonal line to push it out of the error region as much as possible, for example, by moving it to the blue line in Figure 1 (right).

We now prove Theorem 14 using type deviation convergence.

Proof: The proof is divided into five steps.

1) Code construction. Consider the $P_{Y|X}$ that attains $R(D)$. Consider a GCC channel $(P_{Y^{(n)}|X^{(n)}})_n$ with center $P_X \circ P_{Y|X}$ and Lipschitz deviation function $\zeta_{Y|X} : \text{Tan}(P_X) \rightarrow \text{Tan}(P_{Y|X})$ to be specified later. Let $T_{m,y^n} \sim \text{Exp}(1)$, i.i.d. across $m \in [[2^{nR}]]$, $y^n \in \mathcal{Y}^n$, which serves as a random codebook available to the encoder and the decoder. Consider the joint distribution

$$(M, X^{(n)}, Y^{(n)}) \sim \text{Unif}([2^{nR}]) \times (P_X^n \circ P_{Y^{(n)}|X^{(n)}}),$$

i.e., $M \sim \text{Unif}([2^{nR}])$ is independent of $(X^{(n)}, Y^{(n)}) \sim P_X^n \circ P_{Y^{(n)}|X^{(n)}}$. The encoder observes $X^{(n)}$, finds

$$(M, Y^{(n)}) = \text{argmin}_{m, y^n} T_{m, y^n} / P_{M, Y^{(n)}|X^{(n)}}(m, y^n | X^{(n)}),$$

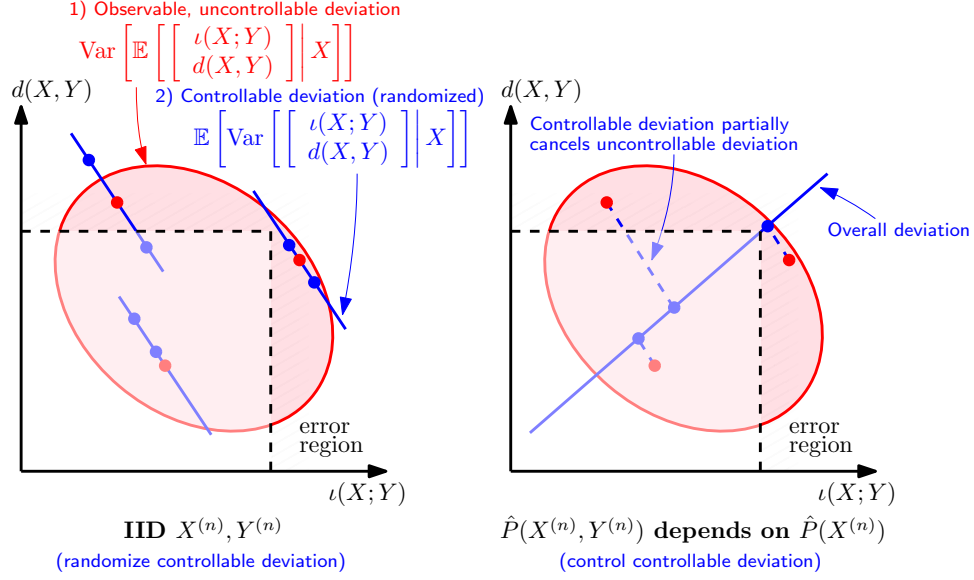


Figure 1. Left: Illustration for lossy source coding with i.i.d. $X^{(n)}, Y^{(n)}$. The three red dots are drawn from a Gaussian distribution with covariance matrix given by the first term in (10) (red ellipse is a contour of the Gaussian distribution), and the blue dots are the red dots plus a Gaussian vector with covariance matrix given by the second term in (10). Right: The optimal scheme where we control the deviation of the type of $Y^{(n)}$ according to the type of $X^{(n)}$, moving the red dots to the blue dots along the blue line.

and sends M . Note that $P_{M,Y^{(n)}|X^{(n)}}(m, y^n | X^{(n)}) = P_M(m)P_{Y^{(n)}|X^{(n)}}(y^n | X^{(n)})$. The decoder observes M and computes

$$(\hat{M}, \hat{Y}^{(n)}) = \operatorname{argmin}_{m, y^n} T_{m, y^n} / P_{M, Y^{(n)}|M}(m, y^n | M).$$

Note that $P_{M, Y^{(n)}|M}(m, y^n | M) = \mathbf{1}\{m = M\}P_{Y^{(n)}}(y^n)$.

2) Computing the asymptotic deviation of $(\mathbf{X}, \mathbf{Y})$. By Proposition (6), $\mathbf{X}$ can be coupled to be type deviation convergent with center P_X and asymptotic deviation $G_X \sim \text{NM}(P_X)$. By Proposition 12, $(\mathbf{X}, \mathbf{Y})$ is type deviation convergent with center $P_X \circ P_{Y|X}$ and asymptotic deviation

$$G_{\mathbf{X}, \mathbf{Y}} = G_X \circ P_{Y|X} + P_X \circ \zeta_{Y|X}(G_X). \quad (11)$$

3) Error bound via Poisson matching lemma and average properties. We slightly relax the error condition and consider the probability that $d(X^{(n)}, \hat{Y}^{(n)}) = \langle \hat{P}(X^{(n)}, \hat{Y}^{(n)}), d \rangle > D + \delta/n$, where $\delta > 0$ will be specified later. We have

$$\begin{aligned} P_e &= \mathbb{P}\left(\left\langle \hat{P}(X^{(n)}, \hat{Y}^{(n)}), d \right\rangle > D + \delta/n\right) \\ &\leq \mathbb{P}\left(\left\langle \hat{P}(X^{(n)}, Y^{(n)}), d \right\rangle > D + \delta/n \text{ or } Y^{(n)} \neq \hat{Y}^{(n)}\right) \\ &\stackrel{(a)}{\leq} \mathbb{P}\left(\left\langle \hat{P}(X^{(n)}, Y^{(n)}), d \right\rangle > D + \delta/n \text{ or } \right. \\ &\quad \left. \iota(M, Y^{(n)}; X^{(n)}) - \iota(M, Y^{(n)}; M) > -\frac{\log n}{2}\right) + O(n^{-1/2}) \\ &\stackrel{(b)}{\leq} \mathbb{P}\left(\langle P_{X,Y}, d \rangle + n^{-1/2} \langle G_{\mathbf{X}, \mathbf{Y}}, d \rangle + O(n^{-1}) > D + \delta/n \text{ or } \right. \\ &\quad \left. \langle P_{X,Y}, \iota_{X;Y} \rangle + n^{-1/2} \langle G_{\mathbf{X}, \mathbf{Y}}, \iota_{X;Y} \rangle - R > -O\left(\frac{\log n}{n}\right)\right) + O(n^{-1/2}) \\ &\stackrel{(c)}{\leq} \mathbb{P}(\langle G_{\mathbf{X}, \mathbf{Y}}, d \rangle > 0 \text{ or } \langle G_{\mathbf{X}, \mathbf{Y}}, \iota_{X;Y} \rangle > W) + O(n^{-1/2}), \end{aligned} \quad (12)$$

where the constants in $O(\dots)$ depend only on the distribution of $(\mathbf{X}, \mathbf{Y})$, (a) is by the Poisson matching lemma (Lemma 13), (b) is by (5) and (6), and (c) is by $\langle P_{X,Y}, d \rangle = D$, $\langle P_{X,Y}, \iota_{X;Y} \rangle = I(X; Y) = R(D)$, taking δ to be the constant in the $O(n^{-1})$ term, and taking

$$R = R(D) + W/\sqrt{n} + O((\log n)/n),$$

where W will be specified later.

4) Simplification via Gaussian vector manipulation. The problem is now reduced to bounding the probability that the Gaussian vector $G_{\mathbf{X}, \mathbf{Y}}$ violates any of the constraints $\langle G_{\mathbf{X}, \mathbf{Y}}, d \rangle \leq 0$ (the distortion constraint) and $\langle G_{\mathbf{X}, \mathbf{Y}}, \iota_{X;Y} \rangle \leq W$ (the

decodability constraint). Although we can leave the answer as an optimization problem over all choices of $\zeta_{Y|X}$, we can further simplify the answer via some simple linear algebra manipulations.

By the optimality of $P_{Y|X}$, it also minimizes $I(X; Y) + \lambda \mathbb{E}[d(X, Y)]$. If $(X, Y) \sim P_X \circ (P_{Y|X} + tV_{Y|X})$ (where $t \in \mathbb{R}$) for $V_{Y|X} \in \text{Tan}(P_{Y|X})$, by (3),

$$\begin{aligned} & \left. \frac{d}{dt} (I(X; Y) + \lambda \mathbb{E}[d(X, Y)]) \right|_{t=0} \\ &= \langle P_X \circ V_{Y|X}, \iota_{X;Y} + \lambda d \rangle. \end{aligned} \quad (13)$$

Hence, the above must be 0 for every $V_{Y|X} \in \text{Tan}(P_{Y|X})$. Letting

$$\begin{aligned} J &:= \langle G_{\mathbf{X}} \circ P_{Y|X}, \iota_{X;Y} \rangle, \\ D &:= \langle G_{\mathbf{X}} \circ P_{Y|X}, d \rangle, \\ K &:= \langle P_X \circ \zeta_{Y|X}(G_{\mathbf{X}}), d \rangle, \end{aligned}$$

we have $\langle P_X \circ \zeta_{Y|X}(G_{\mathbf{X}}), \iota_{X;Y} \rangle = -\lambda K$ by (13). By (11),

$$\begin{aligned} & \mathbb{P}(\langle G_{\mathbf{X}, \mathbf{Y}}, d \rangle > 0 \text{ or } \langle G_{\mathbf{X}, \mathbf{Y}}, \iota_{X;Y} \rangle > W) \\ &= \mathbb{P}(D + K > 0 \text{ or } J - \lambda K > W) \\ &\geq \mathbb{P}(\lambda(D + K) + J - \lambda K > W) \end{aligned} \quad (14)$$

$$= \mathbb{P}(J + \lambda D > W), \quad (15)$$

where the inequality in (14) is an equality when $D + K$ and $J - \lambda K - W$ are nonnegative constant multiples of the same random variable, for example, when $K = -D$ or $K = (J - W)/\lambda$ (since $\zeta_{Y|X}$ is arbitrary, we can choose $\zeta_{Y|X}$ to make K any Lipschitz function of (J, D)).⁹ Taking

$$\begin{aligned} V &= \text{Var}[J + \lambda D] = \text{Var}[J_{X,D}(X)], \\ W &= \sqrt{V} \mathcal{Q}^{-1} \left(\epsilon - \frac{\gamma}{\sqrt{n}} \right) = \sqrt{V} \mathcal{Q}^{-1}(\epsilon) - O(n^{-1/2}), \end{aligned}$$

where γ is chosen be the constant in the $O(n^{-1/2})$ term in (12), we have $P_e \leq \epsilon$ by (12).

5) Minor technical steps. Although the codebook is random, we can fix a particular codebook $(T_{m,y^n})_{m,y^n}$ that minimizes P_e . To complete the proof, we need to strengthen $\langle \hat{P}(X^{(n)}, \hat{Y}^{(n)}), d \rangle \leq D + \delta/n$ to $\langle \hat{P}(X^{(n)}, \hat{Y}^{(n)}), d \rangle \leq D$. We can do so by concatenating $Y^{(n)}$ with k symbols $Y_{n+1}, \dots, Y_{n+k}$, where $Y_{n+i} = \arg\min_y d(X_{n+i}, y)$. Recall that $D > D_{\min} := \min_{P_{Y|X}} \mathbb{E}[d(X, Y)] = \mathbb{E}[d(X_{n+i}, Y_{n+i})]$. By Cramér's theorem, $\mathbb{P}(k^{-1} \sum_{i=1}^k d(X_{n+i}, Y_{n+i}) > (D + D_{\min})/2)$ decays exponentially with k , and hence we can make this probability $O(n^{-1/2})$ by taking $k = O(\log n)$. As long as $k^{-1} \sum_{i=1}^k d(X_{n+i}, Y_{n+i}) \leq (D + D_{\min})/2$, the average cost is reduced by $O(k/n) = O(\log n/n)$, which is more than enough to cancel out the δ/n term. Transmitting $Y_{n+1}, \dots, Y_{n+k}$ requires $O(k) = O(\log n)$ bits, incurring only a $O((\log n)/n)$ increase in R . ■

This result can be generalized to the indirect or noisy lossy source coding [57] where only a noisy version of the source is observed at the encoder. We will show in Section VII that our approach can recover the optimal second order result [43].

VI. SECOND-ORDER WYNER-ZIV CODING

A. Second-Order Achievability for Wyner-Ziv

We now prove a new second-order achievability result for lossy source coding with side information at the decoder, also known as the Wyner-Ziv problem [7], [8], which improves upon existing results [9], [10], [11], [12], [13]. In this setting, there is a 2-discrete memoryless source $(X^{(n)}, Y^{(n)}) \sim P_{X,Y}^n$. The encoder encodes $X^{(n)}$ into a message $M \in [2^{nR}]$ with rate $R > 0$. The decoder observes M and the side information $Y^{(n)}$, and outputs $\hat{Z}^{(n)} \in \mathcal{Z}^n$. The goal is to minimize the probability of excess distortion

$$P_e := \mathbb{P}(d(X^{(n)}, \hat{Z}^{(n)}) > D),$$

⁹The only exception is when $\langle P_X \circ V_{Y|X}, d \rangle = 0$ for every $V_{Y|X} \in \text{Tan}(P_{Y|X})$, which forces $K = 0$. Note that (9) implies that $P_{Y|X}(y|x) > 0$ whenever $P_X(x) > 0$, $P_Y(y) > 0$ (we only consider finite $d(x, y)$ here). Hence, if $\langle P_X \circ V_{Y|X}, d \rangle = 0$ for every $V_{Y|X} \in \text{Tan}(P_{Y|X})$, this implies that $d(x, y)$ is only a function of x . This means that no communication is required, and the theorem is clearly true.

where $d : \mathcal{X} \times \mathcal{Z} \rightarrow \mathbb{R}$ is a distortion function, $d(X^{(n)}, \hat{Z}^{(n)}) := n^{-1} \sum_{i=1}^n d(X_i^{(n)}, \hat{Z}_i^{(n)})$, and $D \in \mathbb{R}$ is the allowed distortion level. The optimal asymptotic rate needed to have $P_e \rightarrow 0$ is given by Wyner-Ziv theorem [7], [8]

$$R(D) := \min_{P_{U|X}, z: \mathcal{U} \times \mathcal{Y} \rightarrow \mathcal{Z}, \mathbb{E}[d(X, Z)] \leq D} (I(U; X) - I(U; Y)), \quad (16)$$

where the minimum is over $P_{U|X}$ and functions $z : \mathcal{U} \times \mathcal{Y} \rightarrow \mathcal{Z}$, subject to the constraint that $\mathbb{E}[d(X, Z)] \leq D$ where $(X, Y, U) \sim P_{X,Y} \circ P_{U|X}$ and $Z = z(U, Y)$. We now prove a second-order result which improves upon existing results [9], [11], [10], [12], [35] (see Section VI-B for a comparison).

Theorem 15. *For discrete Wyner-Ziv coding, assume these two conditions are satisfied: 1) D is a value such that $\lambda := -dR(D)/dD > 0$ is finite at D ; and 2) letting $(P_{U|X}, z)$ be a minimizer in $R(D)$, $(X, Y, U) \sim P_{X,Y} \circ P_{U|X}$ and $Z = z(U, Y)$, they satisfy $\mathbb{E}[\text{Var}[\mathbb{E}[d(X, Z)|X, U]|X]] > 0$. Fix any $W > 0$. For any large enough n , there is a scheme with rate*

$$R = R(D) + \frac{W}{\sqrt{n}} + O\left(\frac{\log n}{n}\right), \quad (17)$$

achieving a probability of excess distortion

$$P_e \leq \mathbb{E}[P_e^*(W - A_{\mathbf{X}})],$$

where $A_{\mathbf{X}} \in \mathbb{R}$ is a zero mean Gaussian random variable with variance

$$\text{Var}[\mathbb{E}[\iota_{U;X}(U; X) - \iota_{U;Y}(U; Y) + \lambda d(X, Z) | X]],$$

and $P_e^* : \mathbb{R} \rightarrow \mathbb{R}$ is defined as

$$P_e^*(\alpha) := \min_{t \in \mathbb{R}} \mathbb{P}(D_{\mathbf{Y}} > t \text{ or } J_{\mathbf{Y}} > \alpha - \lambda t), \quad (18)$$

with $[J_{\mathbf{Y}}, D_{\mathbf{Y}}] \in \mathbb{R}^2$ being a zero mean Gaussian vector with covariance matrix

$$\mathbb{E} \left[\text{Var} \left[\begin{bmatrix} -\iota_{U;Y}(U; Y) \\ d(X, Z) \end{bmatrix} \middle| X, U \right] \right]. \quad (19)$$

To obtain a simpler but looser bound, applying the union bound on (18) gives the following corollary. This corollary subsumes the second-order result for lossy source coding in Theorem 14.

Corollary 16. *For discrete Wyner-Ziv coding, under the same assumptions as Theorem 15, for any fixed $0 < \epsilon < 1$ and any large enough n , there is a scheme achieving a probability of excess distortion $P_e \leq \epsilon$, with rate*

$$R = R(D) + \sqrt{\frac{V_{\text{GCC}}(D)}{n}} \mathcal{Q}^{-1}(\epsilon/2) + O\left(\frac{\log n}{n}\right), \quad (20)$$

where

$$\begin{aligned} V_{\text{GCC}}(D) &:= \text{Var}[\mathbb{E}[\iota(U; X) - \iota(U; Y) + \lambda d(X, Z) | X]] \\ &\quad + \left(\sqrt{\mathbb{E}[\text{Var}[\iota(U; Y) | X, U]]} + \lambda \sqrt{\mathbb{E}[\text{Var}[d(X, Z) | X, U]]} \right)^2. \end{aligned}$$

The $\mathcal{Q}^{-1}(\epsilon/2)$ term can be improved to $\mathcal{Q}^{-1}(\epsilon)$ if $d(X, Z)$ is a function of (X, U) , e.g., if $U = Z$.

Let $R(D, n, \epsilon)$ be the infimum of achievable rates when the blocklength is n and the probability of excess distortion is upper-bounded by ϵ , and

$$V^*(D) := \lim_{\epsilon \rightarrow 0} \limsup_{n \rightarrow \infty} \frac{n(R(D, n, \epsilon) - R(D))^2}{-2 \ln \epsilon} \quad (21)$$

be the operational dispersion (defined similarly as [58], [15]). Corollary 16 gives an upper bound

$$V^*(D) \leq V_{\text{GCC}}(D).$$

We give an intuitive explanation of Theorem 15. Loosely speaking, there are two error events: excess distortion $d(X^{(n)}, Z^{(n)}) > D$, and excess information $n^{-1}(\iota(U^{(n)}; X^{(n)}) - \iota(U^{(n)}; Y^{(n)})) > R$ (the nR -bit message is insufficient to store $U^{(n)}$). In case if

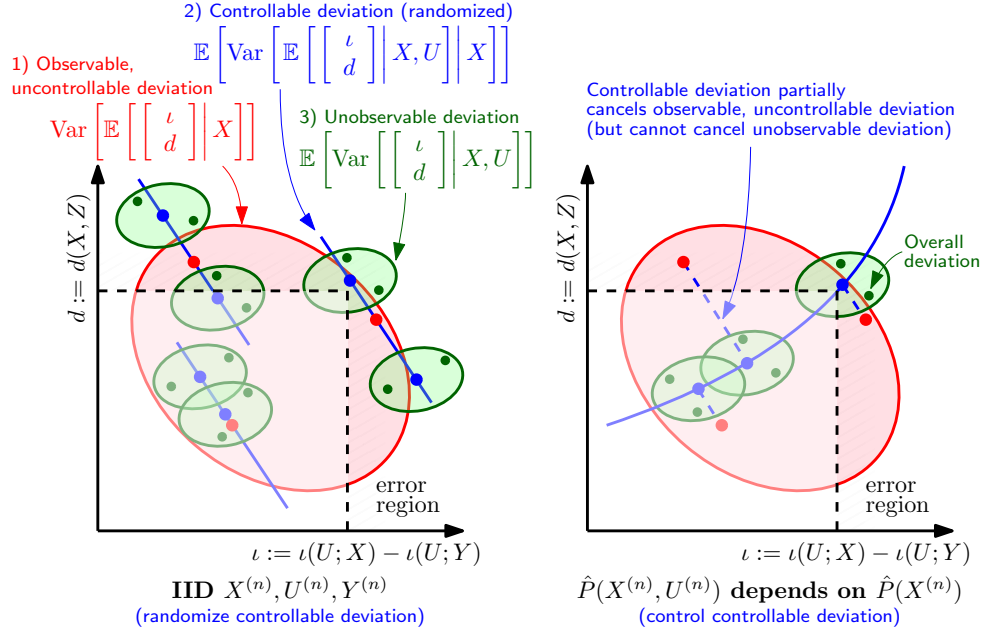


Figure 2. Left: Illustration for Wyner-Ziv coding with i.i.d. $X^{(n)}, U^{(n)}$. The red dots, blue dots and green dots are samples of $\mathbb{E}[\iota, d]^\top | X$, $\mathbb{E}[\iota, d]^\top | X, U$ and $[\iota, d]^\top$, respectively (the 3 stages in Section VI-A). Right: The optimal scheme where we control the deviation of the type of $U^{(n)}$ according to the type of $X^{(n)}$, moving the red dots to the blue dots along the blue curve.

$(X^{(n)}, U^{(n)}) \stackrel{\text{iid}}{\sim} P_X \circ P_{U|X}$, then by the central limit theorem, the vector $[n^{-1}(\iota(U^{(n)}; X^{(n)}) - \iota(U^{(n)}; Y^{(n)})), d(X^{(n)}, Z^{(n)})]^\top$ will be approximately Gaussian with covariance matrix

$$\begin{aligned} & \frac{1}{n} \text{Var} \begin{bmatrix} \iota \\ d \end{bmatrix} \\ &= \frac{1}{n} \text{Var} \left[\mathbb{E} \begin{bmatrix} \iota \\ d \end{bmatrix} \middle| X \right] + \frac{1}{n} \mathbb{E} \left[\text{Var} \left[\mathbb{E} \begin{bmatrix} \iota \\ d \end{bmatrix} \middle| X, U \right] \middle| X \right] \\ &+ \frac{1}{n} \mathbb{E} \left[\text{Var} \begin{bmatrix} \iota \\ d \end{bmatrix} \middle| X, U \right], \end{aligned} \quad (22)$$

due to the law of total covariance, where $\iota := \iota(U; X) - \iota(U; Y)$, $d := d(X, Z)$. The first, second and third term above are the contribution of the randomness in $X^{(n)}$, $U^{(n)}$ and $Y^{(n)}$, respectively. Hence, we can think of the generation $[\iota, d]^\top$ as a 3-stage process: 1) generate $\mathbb{E}[\iota, d]^\top | X$ with covariance matrix given by the first term in (22) (the observable but uncontrollable component of the deviation, contributed by $X^{(n)}$); 2) obtain $\mathbb{E}[\iota, d]^\top | X, U$ by adding a noise to $\mathbb{E}[\iota, d]^\top | X$, with covariance matrix given by the second term (the controllable deviation by $U^{(n)}$); and 3) obtain $[\iota, d]^\top$ by adding a noise with covariance matrix given by the third term (the deviation unobservable by the encoder, contributed by $Y^{(n)}$, which appears in (19)). Error occurs if one of the two components of the random vector is too large; see Figure 2 (left). The second term is rank-one since $\mathbb{E}[\iota + \lambda d | X, U]$ depends only on X by the first-order optimality of $P_{U|X}$.

To reduce the error probability, note that the encoder has full control of $U^{(n)}$, and there is no reason to randomize $U^{(n)}$ and increase the error probability. The second term in (22) can be eliminated. Moreover, we can control the deviation of the type of $U^{(n)}$ according to the type of $X^{(n)}$. Controlling U can only allow trading off $\mathbb{E}[\iota | X, U]$ and $\mathbb{E}[d | X, U]$ along a diagonal line with slope $-1/\lambda$. We move the point $\mathbb{E}[\iota, d]^\top | X$ along this diagonal line according to (18) in order to minimize the probability that the randomness of $Y^{(n)}$ (noise in the third stage with covariance (19)) would push that point into the error region. For example, we can moving the point to the blue curve in Figure 2 (right). Also, the blue curve not being a straight line implies that $[\iota, d]^\top$ has a non-Gaussian deviation. This is generally necessary to attain the minimum in (18).

We now prove Theorem 15 and Corollary 16.

Proof: The proof is divided into four steps.

1) Code construction. Consider the $P_{U|X}$, function $z : \mathcal{U} \times \mathcal{Y} \rightarrow \mathcal{Z}$ that attains $R(D)$. Consider a GCC channel $(P_{U^{(n)}|X^{(n)}})_n$ with center $P_X P_{U|X}$ and Lipschitz deviation function $\zeta_{U|X} : \text{Tan}(P_X) \rightarrow \text{Tan}(P_{U|X})$ to be specified later. Consider the joint distribution

$$(M, X^{(n)}, U^{(n)}, Y^{(n)}) \sim \text{Unif}([2^{nR}]) \times (P_X^n \circ P_{U^{(n)}|X^{(n)}} \circ P_{Y|X}^n).$$

Let $T_{m,u^n} \sim \text{Exp}(1)$, i.i.d. across $m \in [2^{nR}]$, $u^n \in \mathcal{U}^n$, which serves as a random codebook available to the encoder and the decoder. The encoder observes $X^{(n)}$, find

$$(M, U^{(n)}) = \text{argmin}_{m, u^n} T_{m, u^n} / P_{M, U^{(n)} | X^{(n)}}(m, u^n | X^{(n)}),$$

and sends M . The decoder observes $M, Y^{(n)}$ and computes

$$(\hat{M}, \hat{U}^{(n)}) = \text{argmin}_{m, u^n} T_{m, u^n} / P_{M, U^{(n)} | M, Y^{(n)}}(m, u^n | M, Y^{(n)}).$$

2) Computing the asymptotic deviation of $(\mathbf{X}, \mathbf{U}, \mathbf{Y})$. By Proposition (6), $\mathbf{X}$ can be coupled to be type deviation convergent with center P_X and asymptotic deviation $G_X \sim \text{NM}(P_X)$. By Proposition 12, $(\mathbf{X}, \mathbf{U})$ is type deviation convergent with center $P_X P_{U|X}$ and asymptotic deviation

$$G_{\mathbf{X}, \mathbf{U}} = G_X \circ P_{U|X} + P_X \circ \zeta_{U|X}(G_X). \quad (23)$$

By Proposition 7, $(\mathbf{X}, \mathbf{U}, \mathbf{Y})$ can be coupled to be type deviation convergent with center $P_X P_{U|X} P_{Y|X}$ and asymptotic deviation

$$\begin{aligned} G_{\mathbf{X}, \mathbf{U}, \mathbf{Y}} &= (G_X \circ P_{U|X} + P_X \circ \zeta_{U|X}(G_X)) \circ P_{Y|X} \\ &\quad + \sqrt{P_X \circ P_{U|X}} \circ G_{\mathbf{Y} | \mathbf{X}, \mathbf{U}}, \end{aligned} \quad (24)$$

where $G_{\mathbf{Y} | \mathbf{X}, \mathbf{U}} \sim \text{NM}(P_{Y|X, U})$ independent of G_X .

3) Error bound. We slightly relax the error condition and consider the probability that $d(X^{(n)}, \hat{Z}^{(n)}) = \langle \hat{P}(X^{(n)}, \hat{U}^{(n)}, Y^{(n)}), d \rangle > D + \delta/n$, where we let $d(x, u, y) = d(x, z(u, y))$, and $\delta > 0$ will be specified later. Using the Poisson matching lemma in a similar manner as in the proof of Theorem 14,

$$\begin{aligned} P_e &= \mathbb{P} \left(\langle \hat{P}(X^{(n)}, \hat{U}^{(n)}, Y^{(n)}), d \rangle > D + \delta/n \right) \\ &\leq \mathbb{P} (\langle G_{\mathbf{X}, \mathbf{U}, \mathbf{Y}}, d \rangle > 0 \text{ or } \langle G_{\mathbf{X}, \mathbf{U}, \mathbf{Y}}, \iota_{U;X} - \iota_{U;Y} \rangle > W) + O(n^{-1/2}), \end{aligned} \quad (25)$$

with a suitable δ (see Theorem 14), by taking

$$R = R(D) + W/\sqrt{n} + O((\log n)/n).$$

4) Simplification via Gaussian vector manipulation. The problem is now reduced to bounding the probability that the Gaussian vector $G_{\mathbf{X}, \mathbf{U}, \mathbf{Y}}$ violates any of the constraints in (25). Although we can leave the answer as an optimization problem over all choices of $\zeta_{Y|X}$, we can further simplify the answer. Let $\lambda := -dR(D)/dD$. We now consider the terms in (24). Let

$$\begin{aligned} J_X &:= \langle G_X \circ P_{U|X} \circ P_{Y|X}, \iota_{U;X} - \iota_{U;Y} \rangle, \\ D_X &:= \langle G_X \circ P_{U|X} \circ P_{Y|X}, d \rangle, \\ A_X &:= J_X + \lambda D_X, \\ J_U &:= \langle P_X \circ \zeta_{U|X}(G_X) \circ P_{Y|X}, \iota_{U;X} - \iota_{U;Y} \rangle, \\ D_U &:= \langle P_X \circ \zeta_{U|X}(G_X) \circ P_{Y|X}, d \rangle, \\ J_Y &:= \langle \sqrt{P_X \circ P_{U|X}} \circ G_{\mathbf{Y} | \mathbf{X}, \mathbf{U}}, \iota_{U;X} - \iota_{U;Y} \rangle, \\ D_Y &:= \langle \sqrt{P_X \circ P_{U|X}} \circ G_{\mathbf{Y} | \mathbf{X}, \mathbf{U}}, d \rangle. \end{aligned}$$

Note that J_Y, D_Y, A_X follow the distribution stated in the theorem by Proposition 3, and (J_Y, D_Y) is independent of A_X . By the optimality of $P_{U|X}$, using the same arguments as Theorem 14, for any $V_{U|X} \in \text{Tan}(P_{U|X})$,

$$\langle P_X \circ V_{U|X} \circ P_{Y|X}, \iota_{U;X} - \iota_{U;Y} + \lambda d \rangle = 0. \quad (26)$$

Since $\zeta_{U|X}(G_X) \in \text{Tan}(P_{U|X})$, we have $J_U = -\lambda D_U$. Hence, the probability in (25) is

$$\begin{aligned} &\mathbb{P} (\langle G_{\mathbf{X}, \mathbf{U}, \mathbf{Y}}, d \rangle > 0 \text{ or } \langle G_{\mathbf{X}, \mathbf{U}, \mathbf{Y}}, \iota_{U;X} - \iota_{U;Y} \rangle > W) \\ &= \mathbb{E} [\mathbb{P} (D_Y + D_X + D_U > 0 \text{ or } J_Y + J_X + J_U > W | A_X)] \\ &\geq \mathbb{E} [P_e^*(W - A_X)], \end{aligned} \quad (27)$$

where

$$P_e^*(\alpha) := \min_{t \in \mathbb{R}} \mathbb{P} (D_Y - t > 0 \text{ or } J_Y - \alpha + \lambda t > 0).$$

The inequality (27) is shown by considering $t = -D_X - D_U$, which makes $J_Y + J_X + J_U = J_Y + A_X + \lambda t$. We now discuss how to make (27) hold with equality. Letting $G_{U|X} \sim \text{NM}(P_{U|X})$, by Proposition 3, $\langle P_X \circ G_{U|X} \circ P_{Y|X}, d \rangle$ is

zero-mean Gaussian with variance $\mathbb{E}[\text{Var}[\mathbb{E}[d(X, Z)|X, U]|X]] > 0$ by the assumption of Theorem 15. Hence, there exists a fixed $V_{U|X} \in \text{Tan}(P_{U|X})$ with $\langle P_X \circ V_{U|X} \circ P_{Y|X}, d \rangle = 1$. We take

$$\zeta_{U|X}(G_{\mathbf{X}}) = (-\psi(W - A_{\mathbf{X}}) - D_{\mathbf{X}}) \cdot V_{U|X},$$

where $\psi : \mathbb{R} \rightarrow \mathbb{R}$ is a Lipschitz function to be specified later. We have $-D_{\mathbf{X}} - D_{\mathbf{U}} = \psi(W - A_{\mathbf{X}})$. Hence, (27) hold with equality with

$$\psi(\alpha) = \underset{t \in \mathbb{R}}{\text{argmin}} \mathbb{P}(D_{\mathbf{Y}} - t > 0 \text{ or } J_{\mathbf{Y}} - \alpha + \lambda t > 0).$$

The technical proof that ψ is Lipschitz is in Appendix D. Note that ψ is not affine, and hence $G_{\mathbf{X}, \mathbf{U}, \mathbf{Y}}$ is non-Gaussian. The remainder of the proof is similar to Theorem 14, and is omitted.

For Corollary 16, by union bound, $P_e^*(\alpha, t) \leq \mathbb{P}(D_{\mathbf{Y}} > t) + \mathbb{P}(J_{\mathbf{Y}} > \alpha - \lambda t)$. Choosing $t = \tilde{\psi}(\alpha) := \alpha \sigma_{D_{\mathbf{Y}}} / (\sigma_{J_{\mathbf{Y}}} + \lambda \sigma_{D_{\mathbf{Y}}})$ (where $\sigma_{D_{\mathbf{Y}}} := \sqrt{\text{Var}[D_{\mathbf{Y}}]}$) makes these two probabilities equal, giving $P_e^*(\alpha, t) \leq 2Q(\alpha / (\sigma_{J_{\mathbf{Y}}} + \lambda \sigma_{D_{\mathbf{Y}}}))$. The bound in the corollary follows immediately. In case if $\sigma_{D_{\mathbf{Y}}} = 0$ (e.g., if $d(X, Z)$ is a function of (X, U)), we have $t = 0$ and $\mathbb{P}(D_{\mathbf{Y}} > t) = 0$, so we have $P_e^*(\alpha, t) \leq Q(\alpha / (\sigma_{J_{\mathbf{Y}}} + \lambda \sigma_{D_{\mathbf{Y}}}))$ instead. ■

Remark 17. We conjecture that the second-order term in Theorem 15 is optimal for Wyner-Ziv coding. The informal reason is that the first-order optimality of (16) suggests that a scheme achieving the rate $R(D)$ should be close to the random coding scheme with auxiliary U . The proof of Theorem 15 is optimizing over all schemes with type deviation convergent auxiliaries where the deviation from the random coding scheme is small, that is, we have searched over all possible small perturbations of the first-order optimal scheme.

Nevertheless, proving the optimality requires a second-order converse, which is generally difficult for coding theorems involving auxiliary random variables (e.g., see [59], [31]). Considering that Theorem 15 is in a different form compared to any previous second-order result, its converse will likely require new techniques. Since the focus of this paper is the type deviation convergence framework which is purely an achievability technique, converse proofs are out of the scope of this paper. We leave the converse for future research.

B. Comparison with Existing Bounds

Theorem 15 and Corollary 16 improve upon the following existing achievability bounds (we assume that the same assumptions in Theorem 15 hold).

- Yassaee-Aref-Gohari [10] (which improves upon [9]): Achieves a rate $R = R(D) + W/\sqrt{n} + O((\log n)/n)$ with probability of excess distortion

$$P_e \leq \min_{t, \tau \in \mathbb{R}} \mathbb{P}(\bar{J}_{\mathbf{X}} > W - \lambda t - \tau \text{ or } \bar{J}_{\mathbf{Y}} > \tau \text{ or } \bar{D} > t), \quad (28)$$

where $[\bar{J}_{\mathbf{X}}, \bar{J}_{\mathbf{Y}}, \bar{D}]$ is a zero-mean Gaussian vector with the same covariance matrix as $[\iota(U; X), -\iota(U; Y), d(X, Z)]$. The bounds in [9], [10] imply the following upper bound on the operational dispersion (21): $V^*(D) \leq V_{\text{VYAG}}(D)$ where

$$V_{\text{VYAG}}(D) := \left(\sqrt{\text{Var}[\iota(U; X)]} + \sqrt{\text{Var}[\iota(U; Y)]} + \lambda \sqrt{\text{Var}[d(X, Z)]} \right)^2. \quad (29)$$

Theorem 15 improves upon (28), and Corollary 16 improves upon (29), which follow from the law of total variance. See Appendix E.

- Watanabe-Kuzuoka-Tan [11]: Achieves $R = R(D) + W/\sqrt{n} + O((\log n)/n)$ with

$$P_e \leq \min_{t, \tau \in \mathbb{R}} \mathbb{P}(\tilde{J}_{\mathbf{X}} > W - \lambda t - \tau \text{ or } \tilde{J}_{\mathbf{Y}} > \tau \text{ or } \tilde{D} > t), \quad (30)$$

where $[\tilde{J}_{\mathbf{X}}, \tilde{J}_{\mathbf{Y}}, \tilde{D}]$ is a zero-mean Gaussian vector with covariance matrix

$$\mathbb{E}[\text{Var}[\iota(\tilde{U}; X|T), -\iota(\tilde{U}; Y|T), d(X, Z)] | T],$$

where $(X, Y, T, \tilde{U}, Z) \sim P_{X,Y} \circ P_T \circ P_{\tilde{U}|Y,T} \circ P_{Z|Y,T,\tilde{U}}$ with $I(\tilde{U}; X|T) - I(\tilde{U}; Y|T) = R(D)$ and $\mathbb{E}[d(X, Z)] = D$. This implies $V^*(D) \leq V_{\text{WKT}}(D)$ where

$$V_{\text{WKT}}(D) := \left(\sqrt{\mathbb{E}[\text{Var}[\iota(\tilde{U}; X|T)|T]]} + \sqrt{\mathbb{E}[\text{Var}[\iota(\tilde{U}; Y|T)|T]]} + \lambda \sqrt{\mathbb{E}[\text{Var}[d(X, Z)|T]]} \right)^2. \quad (31)$$

This bound is tighter than (28), (29) due to the inclusion of a time-sharing random variable T . Nevertheless, such a T is unnecessary in Theorem 15 since T can be absorbed into U there. We can show that Theorem 15 and Corollary 16 improve upon (30) and (31). See Appendix E.

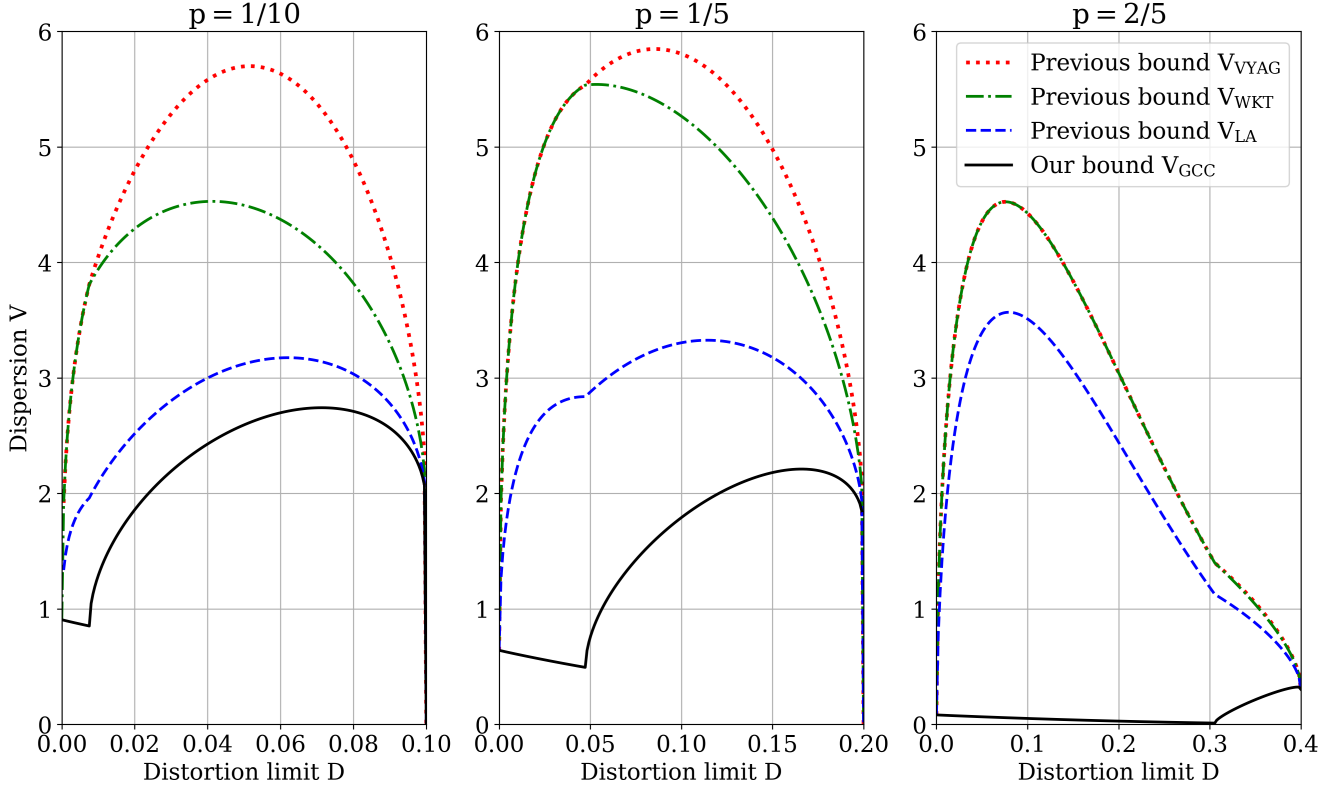


Figure 3. Our upper bound V_{GCC} on $V^*(D)$ for binary-Hamming Wyner-Ziv, and previous upper bounds V_{VYAG} , V_{WKT} , V_{LA} for $p \in \{1/10, 1/5, 2/5\}$ and $D \in [0, p]$.

- Li-Anantharam [12] (also see [35]): Achieves $R = R(D) + W/\sqrt{n} + O((\log n)/n)$ with

$$P_e \leq \min_{t \in \mathbb{R}} \mathbb{P}(\bar{J}_{\mathbf{X}} + \bar{J}_{\mathbf{Y}} > W - \lambda t \text{ or } \bar{D} > t), \quad (32)$$

with the same variables defined in (29). This implies $V^*(D) \leq V_{LA}(D)$ where

$$V_{LA}(D) := \left(\sqrt{\text{Var}[\iota(U; X) - \iota(U; Y)]} + \lambda \sqrt{\text{Var}(X, Z)} \right)^2. \quad (33)$$

This bound is tighter than (28), (29). Theorem 15 and Corollary 16 improve upon (32) and (33). See Appendix E.

In Section VI-C, we will further compare these bounds using a numerical example.

C. Binary-Hamming Wyner-Ziv Coding

For the binary-Hamming Wyner-Ziv problem where $X \sim \text{Bern}(1/2)$, $P_{Y|X}$ is a binary symmetric channel with crossover probability p , and $d(x, z) = \mathbf{1}\{x \neq z\}$ for $z \in \{0, 1\}$, the optimal rate [7], [60] is the lower convex envelope of

$$f(D) = \begin{cases} H_b(p * D) - H_b(D) & \text{if } 0 \leq D < p, \\ 0 & \text{if } D \geq p, \end{cases}$$

where H_b is the binary entropy function, and $a * b := a(1 - b) + b(1 - a)$. To obtain the dispersion bound V_{GCC} , we apply Corollary 16 on $U \in \{0, 1, 2\}$ with $P_{U|X}$ given by the conditional probability matrix $\begin{bmatrix} \gamma(1 - \beta) & \gamma\beta & 1 - \gamma \\ \gamma\beta & \gamma(1 - \beta) & 1 - \gamma \end{bmatrix}$, and $z(u, y) = u$ if $u \in \{0, 1\}$, or $z(u, y) = y$ if $u = 2$, where β, γ are chosen to minimize $I(U; X) - I(U; Y)$ subject to $\mathbb{E}[d(X, Z)] \leq D$. Figure 3 plots our upper bound V_{GCC} , and previous upper bounds V_{VYAG} , V_{WKT} , V_{LA} (see Section VI-B), for $p \in \{1/10, 1/5, 2/5\}$ and $D \in [0, p]$. We can see that our bound significantly improves upon previous bounds, especially for p close to $1/2$.

VII. INDIRECT WYNER-ZIV CODING

The Wyner-Ziv setting can be generalized to a noisy setting where only a noisy version of the source is available to the encoder [38], [39], [22]. In this setting, there is a 3-discrete memoryless source $(F^{(n)}, X^{(n)}, Y^{(n)}) \sim P_{F,X,Y}^n$, where $F^{(n)}$ is the source. The encoder encodes the observation $X^{(n)}$ into a message $M \in [2^{nR}]$ with rate $R > 0$. The decoder observes M and the side information $Y^{(n)}$, and recovers $\hat{Z}^{(n)} \in \mathcal{Z}^n$. The goal is to minimize the probability of excess distortion

$$P_e := \mathbb{P}\left(n^{-1} \sum_{i=1}^n d(F_i^{(n)}, \hat{Z}_i^{(n)}) > D\right),$$

where $d : \mathcal{F} \times \mathcal{Z} \rightarrow \mathbb{R}$ is a distortion function, and $D \in \mathbb{R}$ is the allowed distortion level. The optimal asymptotic rate needed to have $P_e \rightarrow 0$ is given by [38]

$$R(D) := \min_{P_{U|X}, z: \mathbb{E}[d(F, Z)] \leq D} (I(U; X) - I(U; Y)), \quad (34)$$

where the minimum is over $P_{U|X}$ and functions $z : \mathcal{U} \times \mathcal{Y} \rightarrow \mathcal{Z}$, subject to the constraint that $\mathbb{E}[d(F, Z)] \leq D$ where $(F, X, Y, U) \sim P_{F,X,Y} P_{U|X}$ and $Z = z(U, Y)$. We now extend Theorem 15 to this indirect setting.

Theorem 18. *For discrete indirect Wyner-Ziv coding, assume these two conditions are satisfied: 1) D is a value such that $\lambda := -dR(D)/dD > 0$ is finite at D ; and 2) letting $(P_{U|X}, z)$ be a minimizer in $R(D)$, $(F, X, Y, U) \sim P_{F,X,Y} P_{U|X}$ and $Z = z(U, Y)$, they satisfy $\mathbb{E}[\text{Var}[\mathbb{E}[d(F, Z)|X, U]|X]] > 0$. Fix any $W > 0$. For any large enough n , there is a scheme with rate $R = R(D) + W/\sqrt{n} + O((\log n)/n)$, achieving a probability of excess distortion*

$$P_e \leq \mathbb{E}[P_e^*(W - A_{\mathbf{X}})],$$

where $A_{\mathbf{X}} \in \mathbb{R}$ is a zero mean Gaussian random variable with variance $\text{Var}[\mathbb{E}[\iota(U; X) - \iota(U; Y) + \lambda d(F, Z) | X]]$, $P_e^* : \mathbb{R} \rightarrow \mathbb{R}$,

$$P_e^*(\alpha) := \min_{t \in \mathbb{R}} \mathbb{P}(D_{\mathbf{Y}} > t \text{ or } J_{\mathbf{Y}} > \alpha - \lambda t), \quad (35)$$

with $[J_{\mathbf{Y}}, D_{\mathbf{Y}}] \in \mathbb{R}^2$ being a zero mean Gaussian vector with covariance matrix

$$\mathbb{E} \left[\text{Var} \left[\begin{bmatrix} -\iota(U; Y) \\ d(F, Z) \end{bmatrix} \middle| X, U \right] \right].$$

Proof: We use the same coding scheme as in Theorem 15. We can show that $(\mathbf{X}, \mathbf{U}, \mathbf{Y}, \mathbf{F})$ can be coupled to be type deviation convergent with center $P_X P_{U|X} P_{Y,F|X}$ and asymptotic deviation

$$\begin{aligned} G_{\mathbf{X}, \mathbf{U}, \mathbf{Y}, \mathbf{F}} &= (G_{\mathbf{X}} \circ P_{U|X} + P_X \circ \zeta_{U|X}(G_{\mathbf{X}})) \circ P_{Y|X} \\ &\quad + \sqrt{P_X \circ P_{U|X}} \circ G_{\mathbf{Y}, \mathbf{F}|\mathbf{X}, \mathbf{U}}. \end{aligned} \quad (36)$$

Instead of (25), we have

$$\begin{aligned} P_e &\leq \mathbb{P}\left(\langle G_{\mathbf{X}, \mathbf{U}, \mathbf{Y}, \mathbf{F}}, d \rangle > 0 \text{ or} \right. \\ &\quad \left. \langle G_{\mathbf{X}, \mathbf{U}, \mathbf{Y}, \mathbf{F}}, \iota_{U;X} - \iota_{U;Y} \rangle > W\right) + O(n^{-1/2}), \end{aligned}$$

where $d(x, u, y, f) = d(f, z(u, y))$. The remaining steps are the same as Theorem 15, and are omitted. $\blacksquare$

We now show that Theorem 18 recovers the second order result for indirect or noisy lossy source coding [57] given in [43] by taking $Y = \emptyset$ and $U = Z$. We have $J_{\mathbf{Y}} = 0$, and hence we take $t = \alpha/\lambda$ in (35) to give

$$P_e \leq \mathbb{P}\left(D_{\mathbf{Y}} > \frac{W - A_{\mathbf{X}}}{\lambda}\right) = \mathbb{P}(A_{\mathbf{X}} + \lambda D_{\mathbf{Y}} > W),$$

where $A_{\mathbf{X}} \sim N(0, \text{Var}[\mathbb{E}[\iota(Z; X) + \lambda d(F, Z) | X]])$ independent of $D_{\mathbf{Y}} \sim N(0, \mathbb{E}[\text{Var}[d(F, Z) | X, Z]])$. Hence, we have $P_e \leq \epsilon$ for $W = \sqrt{V} \mathcal{Q}^{-1}(\epsilon)$, where

$$\begin{aligned} V &= \text{Var}[\mathbb{E}[\iota(Z; X) + \lambda d(F, Z) | X]] \\ &\quad + \lambda^2 \mathbb{E}[\text{Var}[d(F, Z) | X, Z]] \\ &\stackrel{(a)}{=} \text{Var}[\iota(Z; X) + \lambda d(F, Z)] \\ &\quad - \mathbb{E}[\text{Var}[\mathbb{E}[\iota(Z; X) + \lambda d(F, Z) | X, Z] | X]] \\ &\stackrel{(b)}{=} \text{Var}[\iota(Z; X) + \lambda d(F, Z)], \end{aligned}$$

which coincides with the dispersion in [43], where (a) is by the law of total variance, and (b) is because $\iota(Z; X) + \lambda \mathbb{E}[d(F, Z) | X, Z]$ depends only on X since it is the tilted information for the surrogate distortion function $\mathbb{E}[d(F, Z) | X, Z]$ [43].

We can extend Theorem 18 to the situation where there are multiple distortion functions $d_1, \dots, d_k : \mathcal{F} \times \mathcal{Z} \rightarrow \mathbb{R}$, and the distortion constraint is violated if $n^{-1} \sum_{j=1}^n d_i(F_j^{(n)}, \hat{Z}_j^{(n)}) > D_i$ for any i . In this case, the rate-distortion function is

$$R(\vec{D}) := \min_{P_{U|X, Z} : \mathbb{E}[\vec{d}(F, Z)] \leq \vec{D}} (I(U; X) - I(U; Y)),$$

where we write $\vec{D} = (D_i)_{i=1, \dots, k} \in \mathbb{R}^k$, $\vec{d}(f, z) = (d_i(f, z))_{i=1, \dots, k}$. The following theorem is a straightforward extension of Theorem 18. The proof is omitted.

Theorem 19. *For discrete indirect Wyner-Ziv coding with multiple distortion functions, assume that 1) $\vec{\lambda} := \nabla R(\vec{D}) \in \mathbb{R}^k$ has positive finite entries; and 2) letting $(P_{U|X}, z)$ be a minimizer in $R(\vec{D})$, $(F, X, Y, U) \sim P_{F, X, Y} P_{U|X}$ and $Z = z(U, Y)$, they satisfy that $\mathbb{E}[\text{Var}[\mathbb{E}[\vec{d}(F, Z) | X, U] | X]] \in \mathbb{R}^{k \times k}$ is a full-rank matrix. Fix any $W > 0$. For any large enough n , there is a scheme with rate $R = R(\vec{D}) + W/\sqrt{n} + O((\log n)/n)$, achieving a probability of excess distortion $P_e \leq \mathbb{E}[P_e^*(W - A_{\mathbf{X}})]$, where $A_{\mathbf{X}} \in \mathbb{R}$ is a zero mean Gaussian random variable with variance,*

$$\text{Var} \left[\mathbb{E} \left[\iota(U; X) - \iota(U; Y) + \langle \vec{\lambda}, \vec{d}(F, Z) \rangle \mid X \right] \right],$$

and $P_e^* : \mathbb{R} \rightarrow \mathbb{R}$ is defined as

$$P_e^*(\alpha) := \min_{\vec{t} \in \mathcal{D}} \mathbb{P} \left(J_{\mathbf{Y}} > \alpha - \langle \vec{\lambda}, \vec{t} \rangle \text{ or not } \vec{D}_{\mathbf{Y}} \leq \vec{t} \right),$$

with $J_{\mathbf{Y}} \in \mathbb{R}$, $\vec{D}_{\mathbf{Y}} \in \mathbb{R}^k$ such that $[J_{\mathbf{Y}}, \vec{D}_{\mathbf{Y}}] \in \mathbb{R}^{k+1}$ is a zero mean Gaussian vector with covariance matrix

$$\mathbb{E} \left[\text{Var} \left[\begin{bmatrix} -\iota(U; Y) \\ \vec{d}(F, Z) \end{bmatrix} \mid X, U \right] \right].$$

Theorem 19 improves upon the second-order result in [22] where there are two distortion functions,¹⁰ in a manner similar to how Theorem 15 improves upon [12] (see Section VI-B). Theorem 19 also recovers the dispersion of joint data and semantics lossy compression setting in [44].

VIII. LOSSY COMPRESSION WHERE SIDE INFORMATION MAY BE ABSENT

We consider a generalization of the Wyner-Ziv problem where the side information $Y^{(n)}$ may be absent, known as the Heegard-Berger problem [16], and also studied by Kaspi [17]. In this setting, there is a 2-discrete memoryless source $(X^{(n)}, Y^{(n)}) \sim P_{X, Y}^n$. The encoder encodes $X^{(n)}$ into a message $M \in [2^{nR}]$ with rate $R > 0$. There are two decoders, where Decoder 1 observes M and recovers $\hat{Z}_1^{(n)} \in \mathcal{Z}_1^n$, and Decoder 2 observes M and the side information $Y^{(n)}$, and recovers $\hat{Z}_2^{(n)} \in \mathcal{Z}_2^n$. The goal is to minimize the probability of excess distortion

$$P_e := \mathbb{P} \left(d_1(X^{(n)}, \hat{Z}_1^{(n)}) > D_1 \text{ or } d_2(X^{(n)}, \hat{Z}_2^{(n)}) > D_2 \right),$$

where $d_1 : \mathcal{X} \times \mathcal{Z}_1 \rightarrow \mathbb{R}$, $d_2 : \mathcal{X} \times \mathcal{Z}_2 \rightarrow \mathbb{R}$ are distortion functions, $d_1(x^n, z^n) = n^{-1} \sum_i d_1(x_i, z_i)$, and $D_1, D_2 \in \mathbb{R}$. The optimal asymptotic rate needed to have $P_e \rightarrow 0$ is given by [16], [17]

$$R(D_1, D_2) := \min_{P_{U_1, U_2 | X, Z_1, Z_2}} (I(U_1; X) + I(U_2; X | U_1, Y)), \quad (37)$$

where the minimum is over $P_{U_1, U_2 | X}$ and functions $z_1 : \mathcal{U}_1 \rightarrow \mathcal{Z}_1$, $z_2 : \mathcal{U}_1 \times \mathcal{U}_2 \times \mathcal{Y} \rightarrow \mathcal{Z}_2$, subject to the constraint that $\mathbb{E}[d_i(X, Z_i)] \leq D_i$ for $i = 1, 2$ where $(X, Y, U_1, U_2) \sim P_{X, Y} \circ P_{U_1, U_2 | X}$, $Z_1 = z_1(U_1)$ and $Z_2 = z_2(U_1, U_2, Y)$. We now prove a new second-order achievability result which improves upon [10]. We utilize a new technique which we call *dependent rate splitting*.

Theorem 20. *For discrete Heegard-Berger coding, assume these two conditions are satisfied: 1) $\lambda_i := -\partial R(D_1, D_2)/\partial D_i > 0$ is finite for $i = 1, 2$; 2) letting $(P_{U_1, U_2 | X}, z_1, z_2)$ be a minimizer in $R(D_1, D_2)$, $(X, Y, U_1, U_2) \sim P_{X, Y} P_{U_1, U_2 | X}$, $Z_1 = z_1(U_1)$ and $Z_2 = z_2(U_1, U_2, Y)$, they satisfy that $I(U_1; X), I(U_2; X | U_1, Y) > 0$, and*

$$\mathbb{E}[\text{Var}[\mathbb{E}[[d_1(X, Z_1), d_2(X, Z_2)]^\top \mid X, U_1, U_2] | X]] \quad (38)$$

is a 2×2 full-rank matrix. Fix any $W > 0$. For any large enough n , there is a scheme with rate

$$R = R(D) + W/\sqrt{n} + O((\log n)/n),$$

¹⁰In [22], the decoder outputs $\hat{F}^n, \hat{X}^n$, and the two distortion constraints are $d_1(F^n, \hat{F}^n) \leq D_1$, $d_2(X^n, \hat{X}^n) \leq D_2$. This is covered by the setting in Theorem 19 since we can take $F \leftarrow (F, X)$ and $Z \leftarrow (\hat{F}, \hat{X})$, so both distortion functions can be defined as functions of (F, Z) .

achieving a probability of excess distortion $P_e \leq \mathbb{E}[P_e^*(W - A_{\mathbf{X}})]$, where $A_{\mathbf{X}} \in \mathbb{R}$ is a zero mean Gaussian random variable with variance

$$\text{Var} \left[\mathbb{E} \left[\iota(U_1; X) + \iota(U_2; X|U_1, Y) + \sum_{i=1}^2 \lambda_i d_i(X, Z_i) \mid X \right] \right],$$

and $P_e^* : \mathbb{R} \rightarrow \mathbb{R}$ is defined as

$$P_e^*(\alpha) := \min_{(t_1, t_2) \in \mathbb{R}^2} \mathbb{P} \left(J_{\mathbf{Y}} > \alpha - \lambda_1 t_1 - \lambda_2 t_2 \right. \\ \left. \text{or } D_{1, \mathbf{Y}} > t_1 \text{ or } D_{2, \mathbf{Y}} > t_2 \right),$$

with $[J_{\mathbf{Y}}, D_{1, \mathbf{Y}}, D_{2, \mathbf{Y}}] \in \mathbb{R}^3$ being a zero mean Gaussian vector with covariance matrix

$$\mathbb{E} \left[\text{Var} \left[\begin{bmatrix} \iota(U_1; X) + \iota(U_2; X|U_1, Y) \\ d_1(X, Z_1) \\ d_2(X, Z_2) \end{bmatrix} \mid X, U_1, U_2 \right] \right].$$

Proof: The proof is divided into five steps.

1) Rate splitting. Conventionally, the coding scheme for the Heegard-Berger problem (and many other problems in network information theory) is constructed via rate splitting, where we allocate a rate $I(U_1; X)$ for a message M_1 for Decoder 1, and a rate $I(U_2; X|U_1, Y)$ for a message M_2 for Decoder 2, and put M_1, M_2 together to form the message M . The lengths of M_1 and M_2 are fixed in previous proofs (e.g., [10]). Here, we use a new technique which we call *dependent rate splitting*, where the lengths can depend on the type of $X^{(n)}$. Suppose M_1 has a length $\ell_1 = \lceil nR_1 \rceil$ where

$$R_1 = I(U_1; X) + \frac{W_1 + \kappa(G_{\mathbf{X}}^{(n)})}{\sqrt{n}} + O\left(\frac{\log n}{n}\right), \quad (39)$$

where $G_{\mathbf{X}}^{(n)} := \sqrt{n}(\hat{P}(X^{(n)}) - P_X)$, and $\kappa : \text{Tan}(P_X) \rightarrow \mathbb{R}$ is a Lipschitz function that will be determined later; and M_2 has a length $\ell_2 = \lceil nR_2 \rceil$ where

$$R_2 = I(U_2; X|U_1, Y) + \frac{W_2 - \kappa(G_{\mathbf{X}}^{(n)})}{\sqrt{n}} + O\left(\frac{\log n}{n}\right). \quad (40)$$

Then we have $R_1 + R_2 = R(D_1, D_2) + (W_1 + W_2)/\sqrt{n} + O((\log n)/n)$. Also, since $I(U_1; X), I(U_2; X|U_1, Y) > 0$, the probability that $R_1 < 0$ or $R_2 < 0$ is at most $O(n^{-1/2})$ if $X^{(n)}$ is type deviation convergent, and can be absorbed into ϵ .

2) Code construction. Consider a GCC channel $(P_{U_1^{(n)}, U_2^{(n)}|X^{(n)}})_n$ with center $P_X P_{U_1, U_2|X}$ and Lipschitz deviation function $\zeta_{U_1, U_2|X} : \text{Tan}(P_X) \rightarrow \text{Tan}(P_{U_1, U_2|X})$ to be specified later. Consider the joint distribution

$$(X^{(n)}, U_1^{(n)}, U_2^{(n)}, Y^{(n)}) \sim P_X^n \circ P_{U_1^{(n)}, U_2^{(n)}|X^{(n)}} \circ P_{Y|X}^n.$$

We also let $M_1^{(\ell_1)} \sim \text{Unif}([2^{\ell_1}])$, $M_2^{(\ell_2)} \sim \text{Unif}([2^{\ell_2}])$ to be independent over all $\ell_1, \ell_2 \in \mathbb{Z}_{\geq 0}$. Since there are two components of the encoding, we require two independent exponential codebooks $(T_{m_1, u_1^n}^{(\ell_1)})_{m_1 \in [2^{\ell_1}], u_1^n \in \mathcal{U}_1^n}$ and $(T_{m_2, u_2^n}^{(\ell_2)})_{m_2 \in [2^{\ell_2}], u_2^n \in \mathcal{U}_2^n}$. The encoder observes $X^{(n)}$, computes $\ell_1 = \lceil nR_1 \rceil$ and $\ell_2 = \lceil nR_2 \rceil$ according to (39) and (40), finds

$$(M_1, U_1^{(n)}) = \text{argmin}_{m_1, u_1^n} T_{m_1, u_1^n}^{(\ell_1)} / P_{M_1^{(\ell_1)}, U_1^{(n)}|X^{(n)}}(m_1, u_1^n | X^{(n)}),$$

$$(M_2, U_2^{(n)}) = \text{argmin}_{m_2, u_2^n} T_{m_2, u_2^n}^{(\ell_2)} / P_{M_2^{(\ell_2)}, U_2^{(n)}|X^{(n)}, U_1^{(n)}}(m_2, u_2^n | X^{(n)}, U_1^{(n)}),$$

and sends ℓ_1, ℓ_2 (which only takes $O(\log n)$ bits; they are required so the decoders can split M into M_1, M_2) and M_1, M_2 . Decoder 1 observes ℓ_1, ℓ_2, M_1, M_2 and computes

$$(\hat{M}_1, \hat{U}_1^{(n)}) = \text{argmin}_{m_1, u_1^n} T_{m_1, u_1^n}^{(\ell_1)} / P_{M_1, U_1^{(n)}|M_1}(m_1, u_1^n | M_1).$$

Decoder 2 observes $\ell_1, \ell_2, M_1, M_2, Y^{(n)}$, computes the same $(\hat{M}_1, \hat{U}_1^{(n)})$, and computes

$$(\hat{M}_2, \hat{U}_2^{(n)}) = \text{argmin}_{m_2, u_2^n} T_{m_2, u_2^n}^{(\ell_2)} / P_{M_2, U_2^{(n)}|M_1, U_1^{(n)}, Y^{(n)}}(m_2, u_2^n | M_2, \hat{U}_1^{(n)}, Y^{(n)}).$$

3) Computing the asymptotic deviation. Using similar steps as in Theorem 15, we have the asymptotic deviation

$$G_{\mathbf{X}, \mathbf{U}_1, \mathbf{U}_2, \mathbf{Y}} = G_{\mathbf{X}} \circ P_{U_1, U_2|X} \circ P_{Y|X} \\ + P_X \circ \zeta_{U_1, U_2|X}(G_{\mathbf{X}}) \circ P_{Y|X} \\ + \sqrt{P_{X, U_1, U_2}} \circ G_{\mathbf{Y}|\mathbf{X}, \mathbf{U}_1, \mathbf{U}_2}. \quad (41)$$

4) Error bound. Using the Poisson matching lemma in a similar manner as in the proofs of Theorems 14 and 15,¹¹ writing $G := G_{\mathbf{X}, U_1, U_2, Y}$, we have

$$P_e \leq \mathbb{P} \left(\langle G, d_1 \rangle > 0 \text{ or } \langle G, d_2 \rangle > 0 \right. \quad (42)$$

$$\left. \text{or } \langle G, \iota_{U_1; X} \rangle - \kappa(G_{\mathbf{X}}) > W_1 \right. \quad (43)$$

$$\left. \text{or } \langle G, \iota_{U_2; X, U_1} - \iota_{U_2; Y, U_1} \rangle + \kappa(G_{\mathbf{X}}) > W_2 \right) \quad (44)$$

$$+ O(n^{-1/2}). \quad (45)$$

5) Simplification via Gaussian vector manipulation. Note that

$$\langle G, \iota_{U_1; X} \rangle = \langle G_{\mathbf{X}} \circ P_{U_1|X} + P_X \circ \zeta_{U_1, U_2|X}(G_{\mathbf{X}}), \iota_{U_1; X} \rangle$$

is an Lipschitz function of $G_{\mathbf{X}}$. Hence, we can simply take $\kappa(G_{\mathbf{X}}) = \langle G, \iota_{U_1; X} \rangle$ and $W_1 = 0$ to eliminate the error event (43). We now consider the terms in (42) and (44). Write $\iota := \iota_{U_1; X} + \iota_{U_2; X, U_1} - \iota_{U_2; Y, U_1} = \iota_{U_1; X} + \iota_{U_2; X|U_1, Y}$. Then (44) becomes $\langle G, \iota \rangle > W_2$. For $i = 1, 2$, let

$$\begin{aligned} J_{\mathbf{X}} &:= \langle G_{\mathbf{X}} \circ P_{U_1, U_2|X} \circ P_{Y|X}, \iota \rangle, \\ D_{i, \mathbf{X}} &:= \langle G_{\mathbf{X}} \circ P_{U_1, U_2|X} \circ P_{Y|X}, d_i \rangle, \\ A_{\mathbf{X}} &:= J_{\mathbf{X}} + \lambda_1 D_{1, \mathbf{X}} + \lambda_2 D_{2, \mathbf{X}}, \\ J_{\mathbf{U}} &:= \langle P_X \circ \zeta_{U_1, U_2|X}(G_{\mathbf{X}}) \circ P_{Y|X}, \iota \rangle, \\ D_{i, \mathbf{U}} &:= \langle P_X \circ \zeta_{U_1, U_2|X}(G_{\mathbf{X}}) \circ P_{Y|X}, d_i \rangle, \\ J_{\mathbf{Y}} &:= \langle \sqrt{P_{X, U_1, U_2}} \circ G_{\mathbf{Y}|\mathbf{X}, U_1, U_2}, \iota \rangle, \\ D_{i, \mathbf{Y}} &:= \langle \sqrt{P_{X, U_1, U_2}} \circ G_{\mathbf{Y}|\mathbf{X}, U_1, U_2}, d_i \rangle. \end{aligned}$$

Note that $D_{1, \mathbf{Y}} = 0$ since $G_{\mathbf{Y}|\mathbf{X}, U_1, U_2} \in \text{Tan}(P_{Y|X, U_1, U_2})$. By the optimality of $P_{U_1, U_2|X}$, using the same arguments as Theorem 14, for any $V \in \text{Tan}(P_{U_1, U_2|X})$,

$$\langle P_X \circ V \circ P_{Y|X}, \iota + \lambda_1 d_1 + \lambda_2 d_2 \rangle = 0,$$

and hence $J_{\mathbf{U}} + \lambda_1 D_{1, \mathbf{U}} + \lambda_2 D_{2, \mathbf{U}} = 0$. Letting $G \sim \text{NM}(P_{U_1, U_2|X})$, $[\langle P_X \circ G \circ P_{Y|X}, d_1 \rangle, \langle P_X \circ G \circ P_{Y|X}, d_2 \rangle]^\top$ is zero-mean Gaussian with covariance matrix (38), which is full rank. Hence, there exist fixed $V_1, V_2 \in \text{Tan}(P_{U_1, U_2|X})$ with $\langle P_X \circ V_i \circ P_{Y|X}, d_j \rangle = \mathbf{1}\{i = j\}$ for $i, j = 1, 2$. We take

$$\begin{aligned} \zeta_{U_1, U_2|X}(G_{\mathbf{X}}) &= (-\psi_1(W_2 - A_{\mathbf{X}}) - D_{1, \mathbf{X}}) \cdot V_1 \\ &\quad + (-\psi_2(W_2 - A_{\mathbf{X}}) - D_{2, \mathbf{X}}) \cdot V_2, \end{aligned}$$

where $\psi : \mathbb{R} \rightarrow \mathbb{R}^2$ is a Lipschitz function. The desired result follows from taking

$$\begin{aligned} \psi(\alpha) &= \underset{(t_1, t_2) \in \mathbb{R}^2}{\text{argmin}} \mathbb{P} \left(J_{\mathbf{Y}} - \alpha + \lambda_1 t_1 + \lambda_2 t_2 > 0 \right. \\ &\quad \left. \text{or } D_{1, \mathbf{Y}} - t_1 > 0 \text{ or } D_{2, \mathbf{Y}} - t_2 > 0 \right). \end{aligned}$$

Refer to Appendix D for the technical proof that ψ is Lipschitz.¹² The remaining steps are similar to Theorem 15, and are omitted. ■

IX. SECOND-ORDER CHANNEL CODING WITH COST CONSTRAINT

Apart from source coding results, type deviation convergence is applicable to channel coding as well using the exact same workflow. We demonstrate this technique by recovering the second-order result on channel coding for a discrete memoryless channel with cost constraint [2], [3], [41]. In this setting, the encoder encodes a message $M \sim \text{Unif}[[2^{nR}]]$ with rate $R > 0$ into $X^{(n)} \in \mathcal{X}^n$, and sends it through a memoryless channel $P_{Y|X}^n$. The decoder observes the output $Y^{(n)} \in \mathcal{Y}^n$ and recovers $\hat{M}$. We also require that $n^{-1} \sum_{i=1}^n d(X_i^{(n)}) \leq D$, where $d : \mathcal{X} \rightarrow \mathbb{R}$ is a cost function. The goal is to minimize the probability of error $P_e := \mathbb{P}(M \neq \hat{M})$. We now recover the second-order result in [2], [41].

¹¹Although now ℓ_1, ℓ_2 depends on $X^{(n)}$, the same error analysis works since the Poisson matching lemma (8) is a pointwise bound that applies to every tuple of values of $x^n u_1^n, u_2^n, y^n, \ell_1, \ell_2$.

¹²Although Appendix D only proves the case for 2D Gaussian vector, the generalization to 3D is straightforward and is omitted.

Theorem 21. For discrete channel coding with cost constraint, for any $0 < \epsilon < 1$, if $D > \min_x d(x)$, then for every n , there is a scheme achieving a probability of error $P_e \leq \epsilon$, with rate

$$R = C - \sqrt{\frac{V}{n}} \mathcal{Q}^{-1}(\epsilon) - O\left(\frac{\log n}{n}\right),$$

as long as this rate is positive, where $C := \max_{P_X: \mathbb{E}[d(X)] \leq D} I(X; Y)$ is the capacity, $V := \mathbb{E}[\text{Var}[l_{X;Y}(X; Y)|X]]$ (with (X, Y) induced by the P_X attaining the capacity) is the channel dispersion,¹³ and the constant in $O((\log n)/n)$ depends on $P_{Y|X}$, d , D and ϵ .

Proof: The proof is divided into three steps.

1) Code construction. Consider a GCC channel $(P_{X^{(n)}})_n$ from $\emptyset$ (constant random variable) to X with center P_X and function $\zeta = 0$. Consider the joint distribution

$$(M, X^{(n)}, Y^{(n)}) \sim \text{Unif}([2^{nR}]) \times (P_{X^{(n)}} \circ P_{Y|X}^n).$$

Let $T_{m,x^n} \sim \text{Exp}(1)$, i.i.d. across $m \in [2^{nR}]$, $x^n \in \mathcal{X}^n$, which serves as a random codebook available to the encoder and the decoder. The encoder observes M , find

$$(M, X^{(n)}) = \text{argmin}_{m,x^n} T_{m,x^n} / P_{M,X^{(n)}|M}(m, x^n | M), \quad (46)$$

(note that $P_{M,X^{(n)}|M}(m, x^n | M) = P_{X^{(n)}}(x^n) \cdot \mathbf{1}\{m = M\}$) and sends $X^{(n)}$ if $\langle \hat{P}(X^{(n)}), d \rangle = n^{-1} \sum_{i=1}^n d(X_i^{(n)}) \leq D + \delta/n$, where $\delta > 0$ will be specified later; or else sends an arbitrary $X^{(n)}$ satisfying this constraint. Note that although we have “redefined” M in (46), the two M ’s on the left-hand side and the right-hand side must match. The channel outputs $Y^{(n)} | X^{(n)} \sim P_{Y|X}^n$. The decoder observes $Y^{(n)}$ and computes

$$(\hat{M}, \hat{X}^{(n)}) = \text{argmin}_{m,x^n} T_{m,x^n} / P_{M,X^{(n)}|Y^{(n)}}(m, x^n | Y^{(n)}).$$

Note that $P_{M,X^{(n)}|Y^{(n)}}(m, x^n | Y^{(n)}) = P_{X^{(n)}|Y^{(n)}}(x^n) P_M(m)$.

2) Computing the asymptotic deviation of $(\mathbf{X}, \mathbf{Y})$. We now utilize the type deviation convergence of the sequences. By Proposition 12, $\mathbf{X}$ is type deviation convergent with center P_X and asymptotic deviation 0, and hence

$$\begin{aligned} \mathbb{P}\left(\langle \hat{P}(X^{(n)}), d \rangle > D + \delta/n\right) &\leq \mathbb{P}\left(\langle G_{\mathbf{X}}^{(n)}, d \rangle > \delta/\sqrt{n}\right) \\ &= O(1/\sqrt{n}) \end{aligned} \quad (47)$$

if we take δ to be the constant in Definition 4. By Proposition 7, we can assume that $(\mathbf{X}, \mathbf{Y})$ is type deviation convergent with center $P_X P_{Y|X}$ and asymptotic deviation $G_{\mathbf{X}, \mathbf{Y}} = \sqrt{P_X} \circ G_{\mathbf{Y}|\mathbf{X}}$, where $G_{\mathbf{Y}|\mathbf{X}} \sim \text{NM}(P_{Y|X})$.

3) Error bound. By Proposition 8,

$$\iota(X^{(n)}; Y^{(n)}) \geq nC + \sqrt{n} \left\langle \sqrt{P_X} \circ G_{\mathbf{Y}|\mathbf{X}}, \iota_{X;Y} \right\rangle - c \log n. \quad (48)$$

with probability $1 - O(n^{-1/2})$ for some $c > 0$. By Proposition 3, $\langle \sqrt{P_X} \circ G_{\mathbf{Y}|\mathbf{X}}, \iota_{X;Y} \rangle$ is Gaussian with mean 0 and variance $V = \mathbb{E}[\text{Var}[\iota(X; Y)|X]]$. We have

$$\begin{aligned} P_e &\leq \mathbb{P}\left(\langle \hat{P}(X^{(n)}), d \rangle > D + \delta/n \text{ or } M \neq \hat{M}\right) \\ &\stackrel{(a)}{\leq} \mathbb{P}\left(M \neq \hat{M}\right) + O(n^{-1/2}) \\ &\stackrel{(b)}{\leq} \mathbb{P}\left(\iota(M, X^{(n)}; M) - \iota(M, X^{(n)}; Y^{(n)}) > -\frac{\log n}{2}\right) + O(n^{-1/2}) \\ &\leq \mathbb{P}\left(nR - \iota(X^{(n)}; Y^{(n)}) > -\frac{1}{2} \log n\right) + O(n^{-1/2}) \\ &\stackrel{(c)}{\leq} \mathbb{P}\left(n(R - C) - \sqrt{n} \left\langle \sqrt{P_X} \circ G_{\mathbf{Y}|\mathbf{X}}, \iota_{X;Y} \right\rangle > -\left(c + \frac{1}{2}\right) \log n\right) + O(n^{-1/2}) \\ &\stackrel{(d)}{=} \mathbb{P}\left(\left\langle \sqrt{P_X} \circ G_{\mathbf{Y}|\mathbf{X}}, \iota_{X;Y} \right\rangle < \sqrt{V} \mathcal{Q}^{-1}\left(\epsilon - \frac{\gamma}{\sqrt{n}}\right)\right) + O(n^{-1/2}) \\ &= \epsilon, \end{aligned}$$

¹³In case there are multiple P_X ’s attaining the capacity, choose the one with the smallest V if $\epsilon \leq 1/2$, or the largest V if $\epsilon > 1/2$.

where (a) is by (47), (b) is by the Poisson matching lemma on $M, (M, X^{(n)}), Y^{(n)}$, (c) is by (48), and (d) is by taking

$$R = C - \sqrt{\frac{V}{n}} \mathcal{Q}^{-1} \left(\epsilon - \frac{\gamma}{\sqrt{n}} \right) - \left(c + \frac{1}{2} \right) \frac{\log n}{n},$$

where γ is chosen be the constant in the $O(n^{-1/2})$ term (note that $\mathcal{Q}^{-1}(\epsilon - \gamma n^{-1/2}) = \mathcal{Q}^{-1}(\epsilon) + O(n^{-1/2})$). To complete the proof, we need to strengthen $\langle \hat{P}(X^{(n)}), d \rangle \leq D + \delta/n$ to $\langle \hat{P}(X^{(n)}), d \rangle \leq D$. We can do so by concatenating $X^{(n)}$ with a constant number of symbols x that minimizes $d(x)$. Since $D > \min_x d(x)$, this can reduce the average cost by $O(1/n)$. Also, this will only incur a $O(1/n)$ penalty on R . ■

X. SECOND-ORDER GELFAND-PINSKER WITH COST CONSTRAINT

A. Second-Order Achievability for Gelfand-Pinsker with Cost

We consider channel coding with noncausal state information at the encoder, also known as the Gelfand-Pinsker problem [24], which was also studied by Heegard and El Gamal [25]. In this setting, the encoder observes a message $M \sim \text{Unif}[[2^{nR}]]$ with rate $R > 0$ and the i.i.d. channel state sequence $S^{(n)} \sim P_S^n$, and sends $X^{(n)} \in \mathcal{X}^n$ through a memoryless channel $P_{Y|S,X}^n$. The decoder observes the output $Y^{(n)} \in \mathcal{Y}^n$ and recovers $\hat{M}$. We also require that $n^{-1} \sum_{i=1}^n d(S_i^{(n)}, X_i^{(n)}) \leq D$, where $d : \mathcal{S} \times \mathcal{X} \rightarrow \mathbb{R}$ is a cost function. The goal is to minimize the probability of error $P_e := \mathbb{P}(M \neq \hat{M})$. The asymptotic capacity is given by [24], [25]

$$C(D) := \max_{P_{U|S}, x : \mathbb{E}[d(S, X)] \leq D} (I(U; Y) - I(U; S)), \quad (49)$$

where the minimum is over $P_{U|S}$ and functions $x : \mathcal{S} \times \mathcal{U} \rightarrow \mathcal{X}$, subject to the constraint that $\mathbb{E}[d(S, X)] \leq D$ where $(S, U, X, Y) \sim P_S \circ P_{U|S} \circ P_{X|S,U} \circ P_{Y|S,X}$ with $P_{X|S,U}(x(s, u)|s, u) = 1$. We now show a second-order result that improves upon [11] and [12], [35] (after a straightforward generalization to include cost).

Theorem 22. *For discrete Gelfand-Pinsker coding with cost constraint, assume these two conditions are satisfied: 1) D is a value such that $\lambda := dC(D)/dD \geq 0$ is finite at D ; and 2) letting $(P_{U|S}, x)$ be a minimizer in $C(D)$, $(S, U, X, Y) \sim P_S \circ P_{U|S} \circ P_{X|S,U} \circ P_{Y|S,X}$, they satisfy $\mathbb{E}[\text{Var}[d(S, X)|S]] > 0$. Then for every n , there is a scheme achieving a probability of error $P_e \leq \epsilon$, with rate*

$$R = C(D) - \sqrt{\frac{V}{n}} \mathcal{Q}^{-1}(\epsilon) - O\left(\frac{\log n}{n}\right),$$

as long as this rate is positive, where

$$V := \text{Var}[\mathbb{E}[\iota(U; S) - \iota(U; Y) + \lambda d(S, X) | S]] \\ + \mathbb{E}[\text{Var}[\iota(U; Y) | S, U]],$$

and the constant in $O((\log n)/n)$ depends on $P_{Y|S,X}$, $P_{U|S}$, d , D , ϵ and the function x .

Our result recovers the second-order results in [27], [12] when the cost constraint is inactive, where D is large enough and $\lambda = 0$, giving $V = \text{Var}[\iota(U; S) - \iota(U; Y)]$. We note that [27] does not contain a result for general cost functions (it only covers the power constraint in dirty paper coding [26]). While the analysis in [12] can be extended to incorporate a cost constraint, it requires two error events and leads to a worse second-order result.

We now prove Theorem 22.

Proof: The proof is divided into four steps.

1) Code construction. Consider a GCC channel $(P_{U^{(n)}|S^{(n)}})_n$ with center $P_{S,U}$ and a Lipschitz deviation function $\zeta_{U|S} : \text{Tan}(P_S) \rightarrow \text{Tan}(P_{U|S})$ to be specified later. Consider the joint distribution

$$(M, S^{(n)}, U^{(n)}, Y^{(n)}) \\ \sim \text{Unif}([2^{nR}]) \times (P_{S^{(n)}} \circ P_{U^{(n)}|S^{(n)}} \circ P_{Y^{(n)}|S^{(n)}, U^{(n)}}),$$

where $P_{Y^{(n)}|S^{(n)}, U^{(n)}}$ is the conditional distribution where $X = x(S, U)$ and Y follows $P_{Y|S,X}$. Let $T_{m,u^n} \sim \text{Exp}(1)$, i.i.d. across $m \in [[2^{nR}]]$, $x^n \in \mathcal{X}^n$, which serves as a random codebook available to the encoder and the decoder. The encoder observes M , find

$$(M, U^{(n)}) = \text{argmin}_{m, u^n} T_{m, u^n} / P_{M, U^{(n)}|M}(m, u^n | M), \quad (50)$$

and sends $X^{(n)}$ with $X_i^{(n)} = x(S_i^{(n)}, U_i^{(n)})$ if $\langle \hat{P}(S^{(n)}, X^{(n)}), d \rangle = n^{-1} \sum_{i=1}^n d(S_i^{(n)}, X_i^{(n)}) \leq D + \delta/n$, where $\delta > 0$ will be specified later; or else sends an arbitrary $X^{(n)}$ satisfying this constraint. The decoder observes $Y^{(n)}$ and computes

$$(\hat{M}, \hat{U}^{(n)}) = \text{argmin}_{m, u^n} T_{m, u^n} / P_{M, U^{(n)}|Y^{(n)}}(m, u^n | Y^{(n)}).$$

2) Computing the asymptotic deviation. By Propositions 6, 12 and 7, $(\mathbf{S}, \mathbf{U}, \mathbf{Y})$ can be coupled to be type deviation convergent with center $P_S \circ P_{U|S} \circ P_{Y|S,U}$ and asymptotic deviation

$$G_{\mathbf{S},\mathbf{U},\mathbf{Y}} = (G_{\mathbf{S}} \circ P_{U|S} + P_S \circ \zeta_{U|S}(G_{\mathbf{S}})) \circ P_{Y|S,U} + \sqrt{P_S \circ P_{U|S}} \circ G_{\mathbf{Y}|\mathbf{S},\mathbf{U}}, \quad (51)$$

where $G_{\mathbf{S}} \sim \text{NM}(P_S)$, $G_{\mathbf{Y}|\mathbf{S},\mathbf{U}} \sim \text{NM}(P_{Y|S,U})$.

3) Error bound. Let $d(s, u) := d(s, x(s, u))$. Using the Poisson matching lemma in a similar manner as in the proofs of Theorems 14, 15, 21,

$$P_e \leq \mathbb{P}(\langle G_{\mathbf{S},\mathbf{U},\mathbf{Y}}, d \rangle > 0 \text{ or } \langle G_{\mathbf{S},\mathbf{U},\mathbf{Y}}, \iota_{U;S} - \iota_{U;Y} \rangle > W) + O(n^{-1/2}), \quad (52)$$

with a suitable δ (see Theorem 21), by taking

$$R = C(D) - \frac{W}{\sqrt{n}} - O\left(\frac{\log n}{n}\right),$$

with W to be specified later.

4) Simplification via Gaussian vector manipulation. The problem is now reduced to bounding the probability that the Gaussian vector $G_{\mathbf{S},\mathbf{U},\mathbf{Y}}$ violating any of the constraints in (52). Similar to Theorem 15, define

$$\begin{aligned} J_{\mathbf{S}} &:= \langle G_{\mathbf{S}} \circ P_{U|S} \circ P_{Y|S,U}, \iota_{U;S} - \iota_{U;Y} \rangle, \\ D_{\mathbf{S}} &:= \langle G_{\mathbf{S}} \circ P_{U|S} \circ P_{Y|S,U}, d \rangle, \\ A_{\mathbf{S}} &:= J_{\mathbf{S}} + \lambda D_{\mathbf{S}}, \\ J_{\mathbf{U}} &:= \langle P_S \circ \zeta_{U|S}(G_{\mathbf{S}}) \circ P_{Y|S,U}, \iota_{U;S} - \iota_{U;Y} \rangle, \\ D_{\mathbf{U}} &:= \langle P_S \circ \zeta_{U|S}(G_{\mathbf{S}}) \circ P_{Y|S,U}, d \rangle, \\ J_{\mathbf{Y}} &:= \langle \sqrt{P_S \circ P_{U|S}} \circ G_{\mathbf{Y}|\mathbf{S},\mathbf{U}}, \iota_{U;S} - \iota_{U;Y} \rangle. \end{aligned}$$

By the optimality of $P_{U|S}$, using the same arguments as Theorem 14, for any $V_{U|S} \in \text{Tan}(P_{U|S})$,

$$\langle P_S \circ V_{U|S} \circ P_{Y|S,U}, \iota_{U;S} - \iota_{U;Y} + \lambda d \rangle = 0. \quad (53)$$

Since $\zeta_{U|S}(G_{\mathbf{S}}) \in \text{Tan}(P_{U|S})$, we have $J_{\mathbf{U}} = -\lambda D_{\mathbf{U}}$. We first eliminate the “ $\langle G_{\mathbf{S},\mathbf{U},\mathbf{Y}}, d \rangle > 0$ ” error event in (52). Letting $G_{U|S} \sim \text{NM}(P_{U|S})$, $\langle P_S \circ G_{U|S} \circ P_{Y|S,U}, d \rangle$ is zero-mean Gaussian with variance $\mathbb{E}[\text{Var}[d(S, X)|S]] > 0$, and hence there exists a fixed $V_{U|S} \in \text{Tan}(P_{U|S})$ with $\langle P_S \circ V_{U|S} \circ P_{Y|S,U}, d \rangle = 1$. Take

$$\zeta_{U|S}(G_{\mathbf{S}}) = -D_{\mathbf{S}} \cdot V_{U|S}.$$

We have $D_{\mathbf{U}} = -D_{\mathbf{S}}$ and $\langle G_{\mathbf{S},\mathbf{U},\mathbf{Y}}, d \rangle = D_{\mathbf{S}} + D_{\mathbf{U}} = 0$. The probability in (52) becomes

$$\mathbb{P}(J_{\mathbf{S}} + J_{\mathbf{U}} + J_{\mathbf{Y}} > W) = \mathbb{P}(J_{\mathbf{S}} + \lambda D_{\mathbf{S}} + J_{\mathbf{Y}} > W),$$

which is at most ϵ when $W = \mathcal{Q}^{-1}(\epsilon)\sqrt{\text{Var}[J_{\mathbf{S}} + \lambda D_{\mathbf{S}} + J_{\mathbf{Y}}]}$, which gives the desired bound. Since λ is finite, we have $D > \mathbb{E}[\min_x d(S, x)]$, and hence we can concatenate $X^{(n)}$ with $O(\log n)$ symbols to strengthen $\langle \hat{P}(S^{(n)}, X^{(n)}), d \rangle \leq D + \delta/n$ to $\langle \hat{P}(S^{(n)}, X^{(n)}), d \rangle \leq D$ as in the last step in the proof of Theorem 14. ■

B. Comparisons with Existing Bounds

Theorem 22 improves upon the following existing achievability bounds (we assume that the same assumptions in Theorem 22 hold).

- Watanabe-Kuzuoka-Tan [11].¹⁴ Achieves $R = C(D) - W/\sqrt{n} - O((\log n)/n)$ with

$$P_e \leq \min_{t, \tau \in \mathbb{R}} \mathbb{P}(\tilde{J}_{\mathbf{S}} > W - \lambda t - \tau \text{ or } \tilde{J}_{\mathbf{Y}} > \tau \text{ or } \tilde{D} > t), \quad (54)$$

where $[\tilde{J}_{\mathbf{S}}, \tilde{J}_{\mathbf{Y}}, \tilde{D}]$ is a zero-mean Gaussian vector with covariance matrix

$$\mathbb{E}[\text{Var}[\iota(\tilde{U}; S|T), -\iota(\tilde{U}; Y|T), d(S, X)]^T | T],$$

¹⁴We slightly generalize [11] to allow the cost to depend on the state.

where $(S, \tilde{U}, X, Y, T) \sim P_S \circ P_T \circ P_{\tilde{U}|S,T} \circ P_{X|S,\tilde{U},T} \circ P_{Y|S,X}$ with $I(\tilde{U}; Y|T) - I(\tilde{U}; S|T) = C(D)$ and $\mathbb{E}[d(S, X)] = D$. Note that the time-sharing random variable T is needed in [11], but is unnecessary in Theorem 15 since T can be absorbed into U . We can show that Theorem 22 improves upon (54) using a similar argument as Appendix E.

- Li-Anantharam [12] (after a straightforward generalization to include cost; also see [35]): Achieves $R = C(D) - W/\sqrt{n} - O((\log n)/n)$ with

$$P_e \leq \min_{t \in \mathbb{R}} \mathbb{P}(\bar{J}_X + \bar{J}_Y > W - \lambda t \text{ or } \bar{D} > t), \quad (55)$$

where $[\bar{J}_X, \bar{J}_Y, \bar{D}]$ is a zero-mean Gaussian vector with covariance matrix

$$\text{Var}[\iota(U; S), -\iota(U; Y), d(S, X)]^\top,$$

with the same variables in Theorem 22. We can show that Theorem 22 improves upon (55) using a similar argument as Appendix E.

XI. BROADCAST CHANNELS

The type deviation convergence technique is applicable to a wide range of problems in network information theory. We now consider the broadcast channel with common message, where the encoder encodes the messages $M_i \sim \text{Unif}[\llbracket 2^{nR_i} \rrbracket]$ for $i = 0, 1, 2$ into $X^{(n)} \in \mathcal{X}^n$, and sends it through a memoryless broadcast channel $P_{Y_1, Y_2|X}^n$. Decoder 1 observes the output $Y_1^{(n)} \in \mathcal{Y}_1^n$ and recovers $\hat{M}_{01}, \hat{M}_1$. Decoder 2 observes the output $Y_2^{(n)} \in \mathcal{Y}_2^n$ and recovers $\hat{M}_{02}, \hat{M}_2$. The error probability is

$$P_e := 1 - \mathbb{P}(M_0 = \hat{M}_{01} = \hat{M}_{02}, M_1 = \hat{M}_1, M_2 = \hat{M}_2).$$

We now apply type deviation convergence on [12, Theorem 5] to derive the following achievable dispersion, which is a second-order refinement upon the generalization of Marton's inner bound [61] in [62], [63], [64]. This recovers the currently best known bound in [45],¹⁵ which improves upon previous results [9], [65] (with i.i.d. random codebook), [10] and [12]. For the special case where $R_2 = 0$ (i.e., the asymmetric broadcast channel), this result reduces to [15] applied on constant-composition codes.

Theorem 23. *For a discrete broadcast channel $P_{Y_1, Y_2|X}$, for any P_{U_0, U_1, U_2} , function $x : \mathcal{U}_0 \times \mathcal{U}_1 \times \mathcal{U}_2 \rightarrow \mathcal{X}$, $0 < \epsilon < 1$ and for any n , there is a scheme achieving a probability of error $P_e \leq \epsilon$, with rates R_0, R_1, R_2 if there exists $0 \leq S_1 \leq I(U_1; U_2|U_0)$ (let $S_2 := I(U_1; U_2|U_0) - S_1$) and $0 \leq \tilde{R}_i \leq R_i$ for $i = 1, 2$ such that*

$$\begin{bmatrix} \tilde{R}_1 + S_1 \\ \tilde{R}_2 + S_2 \\ R_0 + R_1 + R_2 - \tilde{R}_2 + S_1 \\ R_0 + R_1 + R_2 - \tilde{R}_1 + S_2 \end{bmatrix} \in \mathbb{E}[J] - \frac{1}{\sqrt{n}} \mathcal{Q}^{-1}(\mathbb{E}[\text{Var}[J|U_0, U_1, U_2]], \epsilon) - \frac{c \log n}{n},$$

$$J := \begin{bmatrix} \iota(U_1; Y_1|U_0) \\ \iota(U_2; Y_2|U_0) \\ \iota(U_0, U_1; Y_1) \\ \iota(U_0, U_2; Y_2) \end{bmatrix},$$

where we assume $(U_0, U_1, U_2) \sim P_{U_0, U_1, U_2}$, $X = x(U_0, U_1, U_2)$, $(Y_1, Y_2)|X \sim P_{Y_1, Y_2|X}$, $\mathcal{Q}^{-1}(\Sigma, \epsilon) := \{v \in \mathbb{R}^k : \mathbb{P}(V \leq v) \geq 1 - \epsilon\}$ for $\Sigma \in \mathbb{R}^{k \times k}$ where $V \sim \mathcal{N}(0, \Sigma)$ and “ $\leq$ ” denotes entrywise comparison (as in [10]), where c is a constant that depends only on $P_{Y_1, Y_2|X}$, P_{U_0, U_1, U_2} , ϵ and the function x .

Proof: Generate $U_0^{(n)}, U_1^{(n)}, U_2^{(n)}$ via a GCC channel $(P_{U_0^{(n)}, U_1^{(n)}, U_2^{(n)}})_n$ from $\emptyset$ (constant random variable) to U_0, U_1, U_2 with center P_{U_0, U_1, U_2} and function $\zeta = 0$. Using Propositions 7 and 12, we have the asymptotic deviation

$$\begin{aligned} & G_{\mathbf{U}_0, \mathbf{U}_1, \mathbf{U}_2, \mathbf{Y}_1, \mathbf{Y}_2} \\ &= \sqrt{P_{U_0, U_1, U_2}} \circ G_{\mathbf{Y}_1, \mathbf{Y}_2|U_0, U_1, U_2}. \end{aligned}$$

The result follows directly from applying the achievability result in [12, Theorem 5], and computing the relevant information density terms using Proposition 8. ■

¹⁵[45] only contains the second-order refinement of the 2-auxiliary Marton's inner bound. The 3-auxiliary bound was conveyed by M. H. Yassaee in a private communication.

XII. ACKNOWLEDGEMENTS

This work was partially supported by two grants from the Research Grants Council of the Hong Kong Special Administrative Region, China [Project No.s: CUHK 24205621 (ECS), CUHK 14209823 (GRF)].

APPENDIX

A. Proof of Proposition 8

For $p \in \hat{P}_n(\mathcal{X})$, its type class is given by the preimage $\hat{P}_n^{-1}(p) := \{x^n : \hat{P}(x^n) = p\}$. Consider $x^n \in \mathcal{X}^n$ with $\hat{p} := \hat{P}(x^n)$. Since $X^{(n)}$ is exchangeable,

$$\begin{aligned} \mathbb{P}(X^{(n)} = x^n) &= |\hat{P}_n^{-1}(\hat{p})|^{-1} \mathbb{P}(\hat{P}(X^{(n)}) = \hat{p}) \\ &= |\hat{P}_n^{-1}(\hat{p})|^{-1} \mathbb{P}(G_{\mathbf{X}}^{(n)} = \sqrt{n}(\hat{p} - P_X)) \\ &= |\hat{P}_n^{-1}(\hat{p})|^{-1} P_{G_{\mathbf{X}}^{(n)}}(\sqrt{n}(\hat{p} - P_X)). \end{aligned}$$

Recall that for $\hat{p} \in \hat{P}(\mathcal{X}^n)$, the logarithm of the size of its type class is [66]

$$\log |\hat{P}_n^{-1}(\hat{p})| = nH(\hat{p}) - O(|\mathcal{X}| \log n).$$

Since $\mathbb{P}(\|G_{\mathbf{X}}^{(n)} - G_{\mathbf{X}}\| > cn^{-1/2}) < cn^{-1/2}$, with probability $1 - cn^{-1/2}$, $\sqrt{n} \langle G_{\mathbf{X}}, \iota_X \rangle$ and $\sqrt{n} \langle G_{\mathbf{X}}^{(n)}, \iota_X \rangle$ differs by at most a constant. We have

$$\begin{aligned} &\iota_{X^{(n)}}(X^{(n)}) - \left(nH(X) + \sqrt{n} \langle G_{\mathbf{X}}^{(n)}, \iota_X \rangle \right) \\ &= \log |\hat{P}_n^{-1}(\hat{P}(X^{(n)}))| - \log P_{G_{\mathbf{X}}^{(n)}}(G_{\mathbf{X}}^{(n)}) - nH(X) - \sqrt{n} \langle G_{\mathbf{X}}^{(n)}, \iota_X \rangle \\ &= \log \left| \hat{P}_n^{-1} \left(P_X + \frac{G_{\mathbf{X}}^{(n)}}{\sqrt{n}} \right) \right| - nH(X) - \sqrt{n} \langle G_{\mathbf{X}}^{(n)}, \iota_X \rangle - \log P_{G_{\mathbf{X}}^{(n)}}(G_{\mathbf{X}}^{(n)}) \\ &= nH \left(P_X + \frac{G_{\mathbf{X}}^{(n)}}{\sqrt{n}} \right) - nH(X) - \sqrt{n} \langle G_{\mathbf{X}}^{(n)}, \iota_X \rangle - \log P_{G_{\mathbf{X}}^{(n)}}(G_{\mathbf{X}}^{(n)}) - O(|\mathcal{X}| \log n) \\ &= n \left(H(X) + \left\langle \frac{G_{\mathbf{X}}^{(n)}}{\sqrt{n}}, \iota_X \right\rangle - O \left(\left\| \frac{G_{\mathbf{X}}^{(n)}}{\sqrt{n}} \right\|^2 \right) \right) \\ &\quad - nH(X) - \sqrt{n} \langle G_{\mathbf{X}}^{(n)}, \iota_X \rangle - \log P_{G_{\mathbf{X}}^{(n)}}(G_{\mathbf{X}}^{(n)}) - O(|\mathcal{X}| \log n) \\ &= -O \left(\left\| G_{\mathbf{X}}^{(n)} \right\|^2 \right) - \log P_{G_{\mathbf{X}}^{(n)}}(G_{\mathbf{X}}^{(n)}) - O(|\mathcal{X}| \log n). \end{aligned}$$

Since there are no more than $(n+1)^{|\mathcal{X}|-1}$ possibilities for $\hat{P}(X^{(n)})$ and $G_{\mathbf{X}}^{(n)}$ [66],

$$\begin{aligned} \mathbb{P} \left(-\log P_{G_{\mathbf{X}}^{(n)}}(G_{\mathbf{X}}^{(n)}) > |\mathcal{X}| \log n \right) &= \mathbb{P} \left(P_{G_{\mathbf{X}}^{(n)}}(G_{\mathbf{X}}^{(n)}) < n^{-|\mathcal{X}|} \right) \\ &\leq n^{-|\mathcal{X}|} (n+1)^{|\mathcal{X}|-1} \\ &\leq 2^{|\mathcal{X}|} n^{-1}. \end{aligned}$$

Consider a constant $c_2 > 0$ which will be specified later. By Definition 4, as long as n is large enough such that $cn^{-1/2} \leq \sqrt{\ln n}$,

$$\begin{aligned} &\mathbb{P} \left(\left\| G_{\mathbf{X}}^{(n)} \right\|^2 > (c_2 + 1)^2 \ln n \right) \\ &= \mathbb{P} \left(\left\| G_{\mathbf{X}}^{(n)} \right\| > (c_2 + 1) \sqrt{\ln n} \right) \\ &\stackrel{(a)}{\leq} \mathbb{P} \left(\|G_{\mathbf{X}}\| > (c_2 + 1) \sqrt{\ln n} - cn^{-1/2} \right) + cn^{-1/2} \\ &\leq \mathbb{P} \left(\|G_{\mathbf{X}}\| > c_2 \sqrt{\ln n} \right) + cn^{-1/2} \\ &\stackrel{(b)}{=} O \left(e^{-(\ln n)/2} \right) + cn^{-1/2} \\ &= O(n^{-1/2}), \end{aligned}$$

where (a) is because $d_{\Pi}(G_{\mathbf{X}}^{(n)}, G_{\mathbf{X}}) \leq cn^{-1/2}$, and (b) holds for an appropriate choice of c_2 since $G_{\mathbf{X}}$ is subgaussian. The result follows.

B. Proof of Proposition 7

Generate $G_{\mathbf{Y}|\mathbf{X}}$, independent of $(\mathbf{X}, G_{\mathbf{X}})$. By [67], for $x^n \in \mathcal{X}^n$

$$\begin{aligned} d_{\text{TV}} \left(\sqrt{\hat{P}(x^n)} \circ G_{\mathbf{Y}|\mathbf{X}}, \sqrt{P_X} \circ G_{\mathbf{Y}|\mathbf{X}} \right) \\ \leq \min\{\zeta \|\hat{P}(x^n) - P_X\|, 1\}, \end{aligned}$$

for some constant $\zeta > 1$, where d_{TV} is the total variation distance. By the coupling lemma for d_{TV} , we can define $\tilde{G}_{\mathbf{Y}|\mathbf{X}}$ which has the same distribution as $G_{\mathbf{Y}|\mathbf{X}}$ and is also independent of $(\mathbf{X}, G_{\mathbf{X}})$, such that

$$\begin{aligned} \mathbb{P} \left(\sqrt{\hat{P}(X^n)} \circ \tilde{G}_{\mathbf{Y}|\mathbf{X}} \neq \sqrt{P_X} \circ G_{\mathbf{Y}|\mathbf{X}} \mid \mathbf{X}, G_{\mathbf{X}} \right) \\ \leq \min\{\zeta \|\hat{P}(X^n) - P_X\|, 1\} \end{aligned} \quad (56)$$

almost surely. Since $\mathbf{X}$ is type deviation convergent, $\|G_{\mathbf{X}}^{(n)} - G_{\mathbf{X}}\| \leq cn^{-1/2}$ with probability at least $1 - cn^{-1/2}$, and hence

$$\begin{aligned} & \mathbb{E}[\min\{\zeta \|\hat{P}(X^n) - P_X\|, 1\}] \\ & \leq \zeta \mathbb{E}[\min\{n^{-1/2} \|G_{\mathbf{X}}^{(n)}\|, 1\}] \\ & \leq \zeta \mathbb{E}[\min\{n^{-1/2} (\|G_{\mathbf{X}}\| + cn^{-1/2}), 1\}] + \zeta cn^{-1/2} \\ & \leq \zeta n^{-1/2} \mathbb{E}[\|G_{\mathbf{X}}\|] + \zeta cn^{-1} + \zeta cn^{-1/2} \\ & = O(n^{-1/2}). \end{aligned}$$

Therefore,

$$\mathbb{P} \left(\sqrt{\hat{P}(X^n)} \circ \tilde{G}_{\mathbf{Y}|\mathbf{X}} \neq \sqrt{P_X} \circ G_{\mathbf{Y}|\mathbf{X}} \right) = O(n^{-1/2}). \quad (57)$$

Generate $Y^{(n)}|X^{(n)} \sim P_{Y|X}^n$ for each n . Recall that $G_{\mathbf{X}}^{(n)} = \sqrt{n}(\hat{P}(X^{(n)}) - P_X)$. Let $\mathbf{n} \in \mathbb{Z}^{\mathcal{X}}$, $n_x := |\{i : X_i^{(n)} = x\}|$ (note that $\mathbf{n} = n\hat{P}(X^{(n)}) = \sqrt{n}G_{\mathbf{X}}^{(n)} + nP_X$). Since $\mathbf{X}$ is type deviation convergent, we have $\|G_{\mathbf{X}}^{(n)}\| \leq c \log n$ and $n_x \geq nP_X(x)/2$ for all x , with probability at least $1 - cn^{-1/2}$ (for some constant c). Consider $\tilde{Y}_x^{(n)} := (Y_i^{(n)})_{i: X_i^{(n)}=x}$, i.e., the vector formed by $Y_i^{(n)}$ for $i \in [n]$ where $X_i^{(n)} = x$. Let $G_{\mathbf{Y}|\mathbf{X}}^{(n)} \in \mathbb{R}^{\mathcal{X} \times \mathcal{Y}}$, $G_{\mathbf{Y}|\mathbf{X}}^{(n)}(\cdot|x) := \sqrt{n_x}(\hat{P}(\tilde{Y}_x^{(n)}) - P_{Y|X=x})$. Applying the central limit theorem in [54], we know that conditional on $X^{(n)} = x^{(n)}$, $G_{\mathbf{Y}|\mathbf{X}}^{(n)}(\cdot|x)$ follows $\text{NM}(P_{Y|X=x})$ approximately, in the sense that there exists a constant c_x (does not depend on n) such that

$$d_{\Pi}(G_{\mathbf{Y}|\mathbf{X}}^{(n)}(\cdot|x), \tilde{G}_{\mathbf{Y}|\mathbf{X}}(\cdot|x)) < \frac{c_x}{\sqrt{n_x}}, \quad (58)$$

where we recall that $\tilde{G}_{\mathbf{Y}|\mathbf{X}}(\cdot|x) \sim \text{NM}(P_{Y|X=x})$, independent of $(\mathbf{X}, G_{\mathbf{X}})$. Recall that we have $n_x \geq nP_X(x)/2$, and hence $c_x/\sqrt{n_x} \leq c_x/\sqrt{nP_X(x)/2}$, with probability at least $1 - cn^{-1/2}$. By the Strassen-Dudley theorem [50, Theorem 6.9], there exists a coupling of $G_{\mathbf{Y}|\mathbf{X}}^{(n)}$ and $\tilde{G}_{\mathbf{Y}|\mathbf{X}}$ such that

$$\mathbb{P} \left(\|G_{\mathbf{Y}|\mathbf{X}}^{(n)} - \tilde{G}_{\mathbf{Y}|\mathbf{X}}\| > \frac{\tilde{c}}{\sqrt{n}} \right) < \frac{\tilde{c}}{\sqrt{n}}, \quad (59)$$

where $\tilde{c} := \sum_x (c_x/\sqrt{P_X(x)/2} + c)$. Therefore, $(Y^{(n)})_n$ can be coupled with $\mathbf{X}$ and $\tilde{G}_{\mathbf{Y}|\mathbf{X}}$ such that the above bound holds for all n . Also, recall that $\tilde{G}_{\mathbf{Y}|\mathbf{X}}$ is independent of $(\mathbf{X}, G_{\mathbf{X}})$ (and hence $(\mathbf{X}, G_{\mathbf{X}}, G_{\mathbf{X}}^{(n)})$). We have

$$\begin{aligned} G_{\mathbf{X}, \mathbf{Y}}^{(n)} &= \sqrt{n}(\hat{P}(X^{(n)}, Y^{(n)}) - P_{X,Y}) \\ &= \left(\left(\frac{n_x}{\sqrt{n}} - \sqrt{n}P_X(x) \right) P_{Y|X=x} + \sqrt{\frac{n_x}{n}} G_{\mathbf{Y}|\mathbf{X}}^{(n)}(\cdot|x) \right)_{x \in \mathcal{X}} \\ &= \sqrt{n} \left(\hat{P}(X^{(n)}) - P_X \right) \circ P_{Y|X} + \sqrt{\hat{P}(X^{(n)})} \circ G_{\mathbf{Y}|\mathbf{X}}^{(n)} \\ &= \sqrt{n} \left(\hat{P}(X^{(n)}) - P_X \right) \circ P_{Y|X} + \sqrt{\hat{P}(X^{(n)})} \circ \tilde{G}_{\mathbf{Y}|\mathbf{X}} + O(n^{-1/2}) \\ &= G_{\mathbf{X}}^{(n)} \circ P_{Y|X} + \sqrt{\hat{P}(X^{(n)})} \circ \tilde{G}_{\mathbf{Y}|\mathbf{X}} + O(n^{-1/2}) \\ &= G_{\mathbf{X}}^{(n)} \circ P_{Y|X} + \sqrt{P_X} \circ G_{\mathbf{Y}|\mathbf{X}} + O(n^{-1/2}) \\ &= G_{\mathbf{X}} \circ P_{Y|X} + \sqrt{P_X} \circ G_{\mathbf{Y}|\mathbf{X}} + O(n^{-1/2}), \end{aligned}$$

with probability $1 - O(n^{-1/2})$, due to (57) and (59), and $G_{\mathbf{X}}^{(n)} - G_{\mathbf{X}} = O(n^{-1/2})$ with probability $1 - O(n^{-1/2})$ since $\mathbf{X}$ is type deviation convergent.¹⁶ This concludes the proof.

C. Proof of Proposition 11

For $x^n \in \mathcal{X}^n$, let $G_{x^n}^{(n)} := \sqrt{n}(\hat{P}(x^n) - P_X)$,

$$\tilde{P}_{U|X}^{x^n}(u|x) := P_{U|X}(u|x) + \frac{P_X(x)}{\sqrt{n}\hat{P}(x^n)(x)}\zeta(G_{x^n}^{(n)})(u|x).$$

(If $P_X(x) = 0$, take $\tilde{P}_{U|X}^{x^n}(u|x) = P_{U|X}(u|x)$.) We will check that $\tilde{P}_{U|X}^{x^n}$ is a valid conditional distribution with nonnegative entries for large enough n and small enough $\|\hat{P}(x^n) - P_X\|$. Since ζ takes values over $\text{Tan}(P_{U|X})$, $\tilde{P}_{U|X}^{x^n}(u|x) = 0$ whenever $P_{U|X}(u|x) = 0$. Also, $\tilde{P}_{U|X}^{x^n}(u|x) = P_{U|X}(u|x)$ when $P_X(x) = 0$. It is left to check the case $P_X(x)P_{U|X}(u|x) > 0$, where we have

$$\begin{aligned} & \tilde{P}_{U|X}^{x^n}(u|x) \\ &= P_{U|X}(u|x) + \frac{P_X(x)\zeta(G_{x^n}^{(n)})(u|x)}{\sqrt{n}\hat{P}(x^n)(x)} \\ &\stackrel{(a)}{\geq} P_{U|X}(u|x) - \frac{P_X(x)(\|\zeta(0)\| + \tilde{c}\sqrt{n}\|\hat{P}(x^n) - P_X\|)}{\sqrt{n}(P_X(x) - \|\hat{P}(x^n) - P_X\|)} \\ &\stackrel{(b)}{\geq} P_{U|X}(u|x) - \frac{2(\|\zeta(0)\| + \tilde{c}\sqrt{n}/c)}{\sqrt{n}} \\ &\stackrel{(c)}{\geq} 0, \end{aligned}$$

where (a) holds for some $\tilde{c} > 0$ since ζ is Lipschitz, (b) is by restricting $\|\hat{P}(x^n) - P_X\| \leq 1/c$ where $1/c \leq (1/2) \min_{x: P_X(x) > 0} P_X(x)$, and (c) is by taking $n \geq c$ and $c \geq 16(\tilde{c} + \|\zeta(0)\|^2) / \min_{u,x: P_{U|X}(u|x) > 0} P_{U|X}(u|x)^2$. Hence, we can find $c > 0$ such that as long as $n \geq c$ and $\|\hat{P}(x^n) - P_X\| \leq 1/c$, $\tilde{P}_{U|X}^{x^n}$ is a valid conditional distribution with nonnegative entries. We have

$$\begin{aligned} & \sqrt{n}(\hat{P}(x^n) \circ \tilde{P}_{U|X}^{x^n} - P_{X,U}) \\ &= \sqrt{n}\hat{P}(x^n) \circ P_{U|X} + P_X \circ \zeta(G_{x^n}^{(n)}) - \sqrt{n}P_{X,U} \\ &= G_{x^n}^{(n)} \circ P_{U|X} + P_X \circ \zeta(G_{x^n}^{(n)}). \end{aligned}$$

To construct the conditional distribution $P_{U^{(n)}|X^{(n)}}$, given $X^{(n)} = x^n$ with $\|\hat{P}(x^n) - P_X\| \leq 1/c$, we can let $U^{(n)}$ be uniform over sequences u^n with conditional type $\tilde{P}_{U|X}^{x^n}$ (rounded off) given x^n . The round off error is $\|\hat{P}(x^n, u^n) - \hat{P}(x^n) \circ \tilde{P}_{U|X}^{x^n}\| = O(1/n)$. Hence,

$$\begin{aligned} & \|\sqrt{n}(\hat{P}(x^n, u^n) - P_{X,U}) - G_{x^n}^{(n)} \circ P_{U|X} - P_X \circ \zeta(G_{x^n}^{(n)})\| \\ &\leq \|\sqrt{n}(\hat{P}(x^n) \circ \tilde{P}_{U|X}^{x^n} - P_{X,U}) - G_{x^n}^{(n)} \circ P_{U|X} - P_X \circ \zeta(G_{x^n}^{(n)})\| + O(n^{-1/2}) \\ &= O(n^{-1/2}). \end{aligned}$$

This completes the proof.

D. Proof of $\psi(a)$ being Lipschitz in Theorem 15

Proposition 24. Let $F(x, y) = \mathbb{P}(X \leq x, Y \leq y)$ be the cumulative distribution function of a Gaussian vector $(X, Y) \in \mathbb{R}^2$. Then there exists a Lipschitz function $\psi : \mathbb{R} \rightarrow \mathbb{R}$ such that $F(\psi(a), a - \psi(a)) = \max_{t \in \mathbb{R}} F(t, a - t)$ for $a \in \mathbb{R}$.

Proof: Without loss of generality, assume (X, Y) is zero mean. Let the covariance matrix be $\Sigma = \begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{bmatrix}$. It is straightforward to check the case where Σ is not full rank, i.e., (X, Y) is supported over a line. Therefore, we assume Σ

¹⁶We say $A_n = B_n + O(n^{-1/2})$ with probability $1 - O(n^{-1/2})$ if there exists a constant c such that $\mathbb{P}(\|A_n - B_n\| \leq cn^{-1/2}) \geq 1 - cn^{-1/2}$ for all n .

is full rank. Let f be the probability density function of (X, Y) , and write $f_x = \partial f / \partial x$, $f_y = \partial f / \partial y$. Let ϕ and Φ be the probability density function and cumulative distribution function of $N(0, 1)$. Let

$$\begin{aligned}\dot{F}_x(x, y) &:= \int_{-\infty}^y f(x, t) dt, & \dot{F}_y(x, y) &:= \int_{-\infty}^x f(s, y) ds, \\ \ddot{F}_x(x, y) &:= \int_{-\infty}^y f_x(x, t) dt, & \ddot{F}_y(x, y) &:= \int_{-\infty}^x f_y(s, y) ds.\end{aligned}$$

Step 1: Showing that $\psi'(a)$ is bounded for $a \gg 0$. By direct evaluation,

$$\psi'(a) = \frac{\ddot{F}_y(x, y) - f(x, y)}{\ddot{F}_x(x, y) + \ddot{F}_y(x, y) - 2f(x, y)},$$

where $(x, y) = (\psi(a), a - \psi(a))$. Hence, $0 \leq \psi'(a) \leq 1$ if $\ddot{F}_x(x, y), \ddot{F}_y(x, y) \leq 0$. We have

$$\dot{F}_x(x, y) = \frac{1}{\sigma_X} \phi\left(\frac{x}{\sigma_X}\right) \Phi\left(\frac{y - \sigma_{XY}\sigma_X^{-2}x}{\sqrt{\sigma_Y^2 - \sigma_{XY}^2\sigma_X^{-2}}}\right).$$

If $\sigma_{XY} \geq 0$, then $\dot{F}_x(x, y)$ is decreasing in x for $x \geq 0$ since both the ϕ and Φ terms are decreasing, and hence $\ddot{F}_x(x, y) \leq 0$ for $x \geq 0$. If $\sigma_{XY} < 0$, then since $\dot{F}_x(x, 0)/(\sigma_X^{-1}\phi(\sigma_X^{-1}x)) \rightarrow -\infty$, we can find $x_1 \geq 0$ such that $\ddot{F}_x(x, 0) < 0$ for $x \geq x_1$, and hence, $\dot{F}_x(x, 0)$ is decreasing in x for $x \geq x_1$, implying that $\dot{F}_x(x, y)$ is decreasing in x for $x \geq x_1$ for any fixed $y \geq 0$ (since $f(x, y)$ is decreasing in x for $x \geq 0$ for fixed $y \geq 0$). Combining both cases, we know that there exists $x_1 \geq 0$ such that $\ddot{F}_x(x, y) \leq 0$ for $x \geq x_1, y \geq 0$. Applying the same arguments on y , there exists $x_1, y_1 \geq 0$ such that $\ddot{F}_x(x, y), \ddot{F}_y(x, y) \leq 0$ for $x \geq x_1, y \geq y_1$, and hence $0 \leq \psi'(a) \leq 1$ whenever $\psi(a) \geq x_1, a - \psi(a) \geq y_1$. Letting $a_1 > 2 \max\{x_1, y_1\}$ be such that $F(a_1/2, a_1/2) > \max\{\Phi(\sigma_X^{-1}x_1), \Phi(\sigma_Y^{-1}y_1)\}$, we have $\psi(a) \geq x_1, a - \psi(a) \geq y_1$ for $a \geq a_1$ by the optimality of ψ , and hence $0 \leq \psi'(a) \leq 1$ for $a \geq a_1$.

Step 2: Showing that $\psi'(a)$ is bounded for $a \ll 0$. Performing a change of coordinate $a = x + y, b = x - y, c = b - \gamma a$ (where γ is specified later), we have

$$\begin{aligned}F(x, y) &= \int_0^\infty \frac{1}{\sigma_A} \phi\left(\frac{a-s}{\sigma_A}\right) \left(\Phi\left(\frac{c+(\gamma+1)s}{\sigma_{B|A}}\right) - \Phi\left(\frac{c+(\gamma-1)s}{\sigma_{B|A}}\right) \right) ds \\ &= \frac{\sigma_A}{-2a} \phi\left(\frac{a}{\sigma_A}\right) K(a, c),\end{aligned}$$

where $\begin{bmatrix} \sigma_A^2 & \sigma_{AB} \\ \sigma_{AB} & \sigma_B^2 \end{bmatrix}$ is the covariance matrix of $(X+Y, X-Y)$, $\gamma := \sigma_{AB}\sigma_A^{-2}$, $\sigma_{B|A}^2 := \sigma_B^2 - \sigma_{AB}^2\sigma_A^{-2}$, and

$$\begin{aligned}K(a, c) &:= \int_0^\infty \frac{-2a}{\sigma_A^2} \exp\left(-\frac{s(s-2a)}{\sigma_A^2}\right) \left(\Phi\left(\frac{c+(\gamma+1)s}{\sigma_{B|A}}\right) - \Phi\left(\frac{c+(\gamma-1)s}{\sigma_{B|A}}\right) \right) ds \\ &= \frac{2}{\sigma_{B|A}} \phi\left(\frac{c}{\sigma_{B|A}}\right) + O\left(\frac{\log(-a)}{-a}\right)\end{aligned}$$

as $a \rightarrow -\infty$ since $-2a\sigma_A^{-2}e^{-s(s-2a)\sigma_A^{-2}}$ approaches the density of the exponential distribution with rate $-2a\sigma_A^{-2}$. Hence,

$$\lim_{a \rightarrow -\infty} \operatorname{argmax}_c K(a, c) = 0.$$

Therefore, for a sufficiently negative, we can focus on $|c| \leq \epsilon$ for some small $\epsilon > 0$ such that $\phi''(t) < 0$ for all $|t| \leq \sigma_{B|A}^{-1}\epsilon/2$. For $|c| \leq \epsilon$,

$$\begin{aligned}& -\sigma_{B|A}^2 \frac{\partial^2 K(a, c)}{\partial c^2} \\ &= \int_0^\infty \frac{-2a}{\sigma_A^2} \exp\left(-\frac{s(s-2a)}{\sigma_A^2}\right) \left(\phi'\left(\frac{c+(\gamma-1)s}{\sigma_{B|A}}\right) - \phi'\left(\frac{c+(\gamma+1)s}{\sigma_{B|A}}\right) \right) ds \\ &\geq \int_0^\infty \frac{-2a}{\sigma_A^2} \exp\left(-\frac{s(s-2a)}{\sigma_A^2}\right) \min_{t \in \sigma_{B|A}^{-1}(\gamma s + (s+\epsilon)[-1, 1])} (-\phi''(t)) ds \\ &\rightarrow \min_{t \in \sigma_{B|A}^{-1}\epsilon[-1, 1]} (-\phi''(t)) > 0\end{aligned}$$

as $a \rightarrow -\infty$ since $\min_{t \in \sigma_{B|A}^{-1}(\gamma s + (s+\epsilon)[-1,1])}(-\phi''(t))$ is bounded and continuous in s . Also, for $a < 0$,

$$\begin{aligned} & \left| \sigma_{B|A} \frac{\partial^2 K(a, c)}{\partial a \partial c} \right| \\ &= \left| \int_0^\infty \frac{2s-2a}{\sigma_A^2 \sigma_A^2} \exp\left(-\frac{s(s-2a)}{\sigma_A^2}\right) \left(\phi\left(\frac{c + (\gamma+1)s}{\sigma_{B|A}}\right) - \phi\left(\frac{c + (\gamma-1)s}{\sigma_{B|A}}\right) \right) ds \right| \\ &\leq \int_0^\infty \frac{2s-2a}{\sigma_A^2 \sigma_A^2} \exp\left(-\frac{s(s-2a)}{\sigma_A^2}\right) \frac{1}{\sqrt{2\pi}} ds \\ &\leq \frac{1}{-a\sqrt{2\pi}}. \end{aligned}$$

Hence, there exists a_0 such that

$$\begin{aligned} \psi'(a) &= \frac{1+\gamma}{2} + \frac{1}{2} \frac{d}{da} \operatorname{argmax}_c K(a, c) \\ &= \frac{1+\gamma}{2} - \frac{1}{2} \frac{\partial^2 K(a, c) / (\partial a \partial c)}{\partial^2 K(a, c) / \partial c^2} \end{aligned}$$

is bounded for $a \leq a_0$.

Step 3: Showing that $\psi'(a)$ is bounded for $a_0 \leq a \leq a_1$. For $a \in [a_0, a_1]$, we have $F(a/2, a/2) \geq F(a_0/2, a_0/2)$, and hence by the optimality of ψ ,

$$\psi(a) \geq \sigma_X \Phi^{-1}(F(a_0/2, a_0/2)),$$

$$a - \psi(a) \geq \sigma_Y \Phi^{-1}(F(a_0/2, a_0/2)).$$

Therefore, to study the behavior of $\psi(a)$, it suffices to study $F(x, y)$ for (x, y) satisfying $x \geq \sigma_X \Phi^{-1}(F(a_0/2, a_0/2))$, $y \geq \sigma_Y \Phi^{-1}(F(a_0/2, a_0/2))$ and $x + y \leq a_1$, which is a closed bounded region. Let $H(x, y) := \nabla^2(-\ln F(x, y)) \in \mathbb{R}^{2 \times 2}$ be the Hessian matrix of $-\ln F(x, y)$. We will prove that $H(x, y)$ is positive definite. By the Prékopa-Leindler inequality, for any log-concave function $g(x, y)$, the function $G(x, y) = \int_{-\infty}^x \int_{-\infty}^y g(s, t) dt ds$ is log-concave as well. Since f is log-concave, $H(x, y)$ must be positive semidefinite. By direct computation,

$$\begin{aligned} H_{11}(x, y) &= (\dot{F}_x(x, y)^2 - F(x, y)\ddot{F}_x(x, y))/F(x, y)^2, \\ H_{22}(x, y) &= (\dot{F}_y(x, y)^2 - F(x, y)\ddot{F}_y(x, y))/F(x, y)^2, \\ H_{12}(x, y) &= (\dot{F}_x(x, y)\dot{F}_y(x, y) - F(x, y)f(x, y))/F(x, y)^2. \end{aligned}$$

We can see that $H(x, y)$ depends only on $F(x, y)$ and the values of f around the neighborhood of $\{(x, t) : t \leq y\} \cup \{(s, y) : s \leq x\}$. For a fixed (x, y) , if we can construct an alternative log-concave probability density function g such that its cumulative distribution function G satisfies that $g(x, y) = f(x, y)$ (at this particular fixed (x, y)), $G(x, y) = F(x, y)$, $\dot{G}_x(x, y) = \dot{F}_x(x, y)$, $\dot{G}_y(x, y) = \dot{F}_y(x, y)$, but $\ddot{G}_x(x, y) > \ddot{F}_x(x, y)$ and $\ddot{G}_y(x, y) > \ddot{F}_y(x, y)$, then $H(x, y)$ would be the sum of the Hessian of $-\ln G(x, y)$ and a positive definite diagonal matrix, and hence is strictly positive definite. To do this, we can take the small perturbation $g(s, t) = f(s, t)e^{b(s-x, t-y+2)+b(t-y, s-x+2)}$, where $b(s, t)$ is a suitable bump function supported over $[-1, 1]^2$ chosen to satisfy the requirements.¹⁷ Hence, $H(x, y)$ is strictly positive definite. Let

$$D(x, y) := \frac{H_{22}(x, y) - H_{12}(x, y)}{H_{11}(x, y) + H_{22}(x, y) - 2H_{12}(x, y)}.$$

We have $\psi'(a) = D(\psi(a), a - \psi(a))$. Since $H(x, y)$ is positive definite, the denominator of $D(x, y)$ is positive, and hence $D(x, y)$ is a continuous function, and is bounded within a closed bounded set of (x, y) . Hence, $\psi'(a)$ is bounded for $a_0 \leq a \leq a_1$. Therefore, combining all three steps, we know $\psi'(a)$ is always bounded, and ψ is Lipschitz. ■

E. Proof of the Comparisons in Section VI-B

We first prove that Theorem 15 improves upon (32), and hence (28). Let $\bar{J} := \bar{J}_X + \bar{J}_Y$. Then $[\bar{J}, \bar{D}]$ has a covariance matrix $\operatorname{Var}[\iota(U; X) - \iota(U; Y), d(X, Z)]$. Let $[J_X, D_X]$, $[\hat{J}_U, \hat{D}_U]$, $[J_Y, D_Y]$ be independent a zero mean Gaussian vectors with covariance matrices

$$\operatorname{Var} \left[\mathbb{E} \left[\begin{bmatrix} \iota(U; X) - \iota(U; Y) \\ d(X, Z) \end{bmatrix} \middle| X \right] \right],$$

¹⁷More precisely, we require $\nabla^2 b \preceq \nabla^2(-\ln f)$, where $\preceq$ is the semidefinite order (note that $\nabla^2(-\ln f)$ is constant), $b(s, t) = 0$ when $s = 0$ or $t = 0$, $\partial b(s, t)/\partial s \geq 0$ at $s = 0$ with strict inequality for some t , $\partial b(s, t)/\partial t \geq 0$ at $t = 0$ with strict inequality for some s , and $G(x, y) = F(x, y)$. An explicit construction would be $b(s, t) = \delta(\gamma s^2 + s) \exp(1/((s^2 - 1)(t^2 - 1)))$ for $(s, t) \in [-1, 1]^2$, where $\delta > 0$, $\gamma \in \mathbb{R}$ are suitable small constants.

$$\mathbb{E} \left[\text{Var} \left[\mathbb{E} \left[\begin{bmatrix} \iota(U; X) - \iota(U; Y) \\ d(X, Z) \end{bmatrix} \middle| X, U \right] \middle| X \right] \right],$$

and

$$\mathbb{E} \left[\text{Var} \left[\begin{bmatrix} -\iota(U; Y) \\ d(X, Z) \end{bmatrix} \middle| X, U \right] \right],$$

respectively. Let $A_{\mathbf{X}} = J_{\mathbf{X}} + \lambda D_{\mathbf{X}}$. By the law of total variance, the covariance matrix of $[J_{\mathbf{X}}, D_{\mathbf{X}}] + [\hat{J}_{\mathbf{U}}, \hat{D}_{\mathbf{U}}] + [J_{\mathbf{Y}}, D_{\mathbf{Y}}]$ is the same as that of $[\bar{J}, \bar{D}]$, and hence we can assume $\bar{J} = J_{\mathbf{X}} + \hat{J}_{\mathbf{U}} + J_{\mathbf{Y}}$ and $\bar{D} = D_{\mathbf{X}} + \hat{D}_{\mathbf{U}} + D_{\mathbf{Y}}$. Also, by the first-order optimality of $P_{U|X}$, we have $\hat{J}_{\mathbf{U}} = -\lambda \hat{D}_{\mathbf{U}}$ (see the proof of Theorem 15). To compare (32) with Theorem 15, for any $t \in \mathbb{R}$,

$$\begin{aligned} & \mathbb{P}(\bar{J} > W - \lambda t \text{ or } \bar{D} > t) \\ &= \mathbb{P}(J_{\mathbf{X}} + \hat{J}_{\mathbf{U}} + J_{\mathbf{Y}} > W - \lambda t \text{ or } D_{\mathbf{X}} + \hat{D}_{\mathbf{U}} + D_{\mathbf{Y}} > t) \\ &= \mathbb{P}(J_{\mathbf{Y}} > W - A_{\mathbf{X}} - \lambda(t - D_{\mathbf{X}} - \hat{D}_{\mathbf{U}}) \text{ or } D_{\mathbf{Y}} > t - D_{\mathbf{X}} - \hat{D}_{\mathbf{U}}) \\ &\geq \mathbb{E}[P_e^*(W - A_{\mathbf{X}})], \end{aligned} \tag{60}$$

where $P_e^*(\alpha) = \min_{t'} \mathbb{P}(J_{\mathbf{Y}} > \alpha - \lambda t' \text{ or } D_{\mathbf{Y}} > t')$ as in Theorem 15, by substituting $t' = t - D_{\mathbf{X}} - \hat{D}_{\mathbf{U}}$. Hence, (32) is greater than or equal to the bound in Theorem 15. Also, Corollary 16 improves upon (33) since they are the low excess distortion probability limits of Theorem 15 and (32), respectively.

Next, we prove that Theorem 15 improves upon (30). First, we can assume that Z is a function of $(\tilde{U}, Y, T)$ since the randomness in Z can be absorbed into T by the functional representation lemma [68]. Since T is independent of (X, Y) , we have $\iota(\tilde{U}; X|T) = \iota(\tilde{U}, T; X)$ and $\iota(\tilde{U}; Y|T) = \iota(\tilde{U}, T; Y)$. Take $U = (\tilde{U}, T)$. Let $\tilde{J} := \tilde{J}_{\mathbf{X}} + \tilde{J}_{\mathbf{Y}}$. Let $K := [\iota(U; X) - \iota(U; Y), d(X, Z)] \in \mathbb{R}^2$ and $\bar{K} := \mathbb{E}[K|X, U]$. Let $[\tilde{J}_{\mathbf{U}}, \tilde{D}_{\mathbf{U}}]$ be a zero mean Gaussian vector with covariance matrix

$$\begin{aligned} & \mathbb{E}[\text{Var}[\bar{K}|T]] - \text{Var}[\mathbb{E}[\bar{K}|X]]. \\ &= \text{Var}[\bar{K}] - \text{Var}[\mathbb{E}[\bar{K}|T]] - \text{Var}[\mathbb{E}[\bar{K}|X]] \\ &= \text{Var}[\bar{K} - \mathbb{E}[\bar{K}|T] - \mathbb{E}[\bar{K}|X]] \end{aligned}$$

since X, T are independent. Define $[J_{\mathbf{X}}, D_{\mathbf{X}}]$, $[J_{\mathbf{Y}}, D_{\mathbf{Y}}]$ as in the previous part. By the law of total variance, the covariance matrix of $[J_{\mathbf{X}}, D_{\mathbf{X}}] + [\tilde{J}_{\mathbf{U}}, \tilde{D}_{\mathbf{U}}] + [J_{\mathbf{Y}}, D_{\mathbf{Y}}]$ is the same as that of $[\tilde{J}, \tilde{D}]$, and hence we can assume $\tilde{J} = J_{\mathbf{X}} + \tilde{J}_{\mathbf{U}} + J_{\mathbf{Y}}$ and $\tilde{D} = D_{\mathbf{X}} + \tilde{D}_{\mathbf{U}} + D_{\mathbf{Y}}$. We have

$$\begin{aligned} & \text{Var}[J_{\mathbf{X}}, D_{\mathbf{X}}] + \text{Var}[\tilde{J}_{\mathbf{U}}, \tilde{D}_{\mathbf{U}}] + \text{Var}[J_{\mathbf{Y}}, D_{\mathbf{Y}}] \\ &= \text{Var}[\tilde{J}, \tilde{D}] \\ &\preceq \text{Var}[\bar{J}, \bar{D}] \\ &= \text{Var}[J_{\mathbf{X}}, D_{\mathbf{X}}] + \text{Var}[\hat{J}_{\mathbf{U}}, \hat{D}_{\mathbf{U}}] + \text{Var}[J_{\mathbf{Y}}, D_{\mathbf{Y}}], \end{aligned}$$

and hence $\text{Var}[\tilde{J}_{\mathbf{U}}, \tilde{D}_{\mathbf{U}}] \preceq \text{Var}[\hat{J}_{\mathbf{U}}, \hat{D}_{\mathbf{U}}]$, where “ $\preceq$ ” denotes the positive semidefinite order. Since $\hat{J}_{\mathbf{U}} = -\lambda \hat{D}_{\mathbf{U}}$, we also have $\tilde{J}_{\mathbf{U}} = -\lambda \tilde{D}_{\mathbf{U}}$. Using the same steps as (60), for any $t, \tau \in \mathbb{R}$,

$$\begin{aligned} & \mathbb{P}(\tilde{J}_{\mathbf{X}} > W - \lambda t - \tau \text{ or } \tilde{J}_{\mathbf{Y}} > \tau \text{ or } \tilde{D} > t) \\ &\geq \mathbb{P}(\bar{J} > W - \lambda t \text{ or } \bar{D} > t) \\ &\geq \mathbb{E}[P_e^*(W - A_{\mathbf{X}})]. \end{aligned}$$

Hence, (30) is greater than or equal to the bound in Theorem 15, and Corollary 16 improves upon (31) since they are the low excess distortion probability limits of Theorem 15 and (30), respectively.

REFERENCES

- [1] V. Strassen, “Asymptotic estimates in Shannon’s information theory,” in *Proc. 3rd Trans. Prague Conf. Inf. Theory*, 1962, pp. 689–723.
- [2] M. Hayashi, “Information spectrum approach to second-order coding rate in channel coding,” *IEEE Transactions on Information Theory*, vol. 55, no. 11, pp. 4947–4966, 2009.
- [3] Y. Polyanskiy, H. V. Poor, and S. Verdú, “Channel coding rate in the finite blocklength regime,” *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2307–2359, 2010.
- [4] I. Kontoyiannis, “Pointwise redundancy in lossy data compression and universal lossy data compression,” *IEEE Transactions on Information Theory*, vol. 46, no. 1, pp. 136–152, 2000.
- [5] V. Kostina and S. Verdú, “Fixed-length lossy compression in the finite blocklength regime,” *IEEE Transactions on Information Theory*, vol. 58, no. 6, pp. 3309–3338, 2012.
- [6] A. Ingber and Y. Kochman, “The dispersion of lossy source coding,” in *2011 Data Compression Conference*, March 2011, pp. 53–62.
- [7] A. D. Wyner and J. Ziv, “The rate-distortion function for source coding with side information at the decoder,” *IEEE Transactions on Information Theory*, vol. 22, no. 1, pp. 1–10, January 1976.
- [8] A. D. Wyner, “The rate-distortion function for source coding with side information at the decoder-ii. general sources,” *Information and control*, vol. 38, no. 1, pp. 60–80, 1978.

- [9] S. Verdú, “Non-asymptotic achievability bounds in multiuser information theory,” in *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*, Oct 2012, pp. 1–8.
- [10] M. H. Yassaee, M. R. Aref, and A. Gohari, “A technique for deriving one-shot achievability results in network information theory,” in *2013 IEEE ISIT*, July 2013, pp. 1287–1291.
- [11] S. Watanabe, S. Kuzuoka, and V. Y. F. Tan, “Nonasymptotic and second-order achievability bounds for coding with side-information,” *IEEE Transactions on Information Theory*, vol. 61, no. 4, pp. 1574–1605, April 2015.
- [12] C. T. Li and V. Anantharam, “A unified framework for one-shot achievability via the Poisson matching lemma,” *IEEE Transactions on Information Theory*, vol. 67, no. 5, pp. 2624–2651, 2021.
- [13] Y. Liu and C. T. Li, “One-shot coding over general noisy networks,” in *2024 IEEE International Symposium on Information Theory (ISIT)*, 2024, pp. 3124–3129.
- [14] D. Slepian and J. K. Wolf, “Noiseless coding of correlated information sources,” *IEEE Trans. Inf. Theory*, vol. 19, no. 4, pp. 471–480, Jul. 1973.
- [15] V. Y. Tan and O. Kosut, “On the dispersions of three network information theory problems,” *IEEE Transactions on Information Theory*, vol. 60, no. 2, pp. 881–903, 2013.
- [16] C. Heegard and T. Berger, “Rate distortion when side information may be absent,” *IEEE Transactions on Information Theory*, vol. 31, no. 6, pp. 727–734, 1985.
- [17] A. H. Kaspi, “Rate-distortion function when side-information may be present at the decoder,” *IEEE Transactions on Information Theory*, vol. 40, no. 6, pp. 2031–2034, 1994.
- [18] S. C. Draper and G. W. Wornell, “Side information aware coding strategies for sensor networks,” *IEEE Journal on Selected Areas in Communications*, vol. 22, no. 6, pp. 966–976, 2004.
- [19] Y. Kochman and G. W. Wornell, “On the excess distortion exponent of the quadratic-Gaussian Wyner-Ziv problem,” in *2010 IEEE International Symposium on Information Theory*. IEEE, 2010, pp. 36–40.
- [20] B. Phan, A. Khisti, and C. Louizos, “Importance matching lemma for lossy compression with side information,” in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2024, pp. 1387–1395.
- [21] M. Gkagkos and C. D. Charalambous, “Structural properties of the Wyner-Ziv rate distortion function: Applications for multivariate gaussian sources,” *Entropy*, vol. 26, no. 4, p. 306, 2024.
- [22] J. Wei, P. Mary, and E. Dupraz, “Non-asymptotic achievable rate-distortion region for indirect Wyner-Ziv source coding,” *arXiv preprint arXiv:2507.17432*, 2025.
- [23] S. Watanabe, “Tight exponential strong converses for lossy source coding with side-information and distributed function computation,” *arXiv preprint arXiv:2504.16380*, 2025.
- [24] S. I. Gel’fand and M. S. Pinsker, “Coding for channel with random parameters,” *Probl. Contr. and Inf. Theory*, vol. 9, no. 1, pp. 19–31, 1980.
- [25] C. Heegard and A. El Gamal, “On the capacity of computer memory with defects,” *IEEE Transactions on Information Theory*, vol. 29, no. 5, pp. 731–739, 1983.
- [26] M. Costa, “Writing on dirty paper (corresp.),” *IEEE Transactions on Information Theory*, vol. 29, no. 3, pp. 439–441, 1983.
- [27] J. Scarlett, “On the dispersions of the Gel’fand-Pinsker channel and dirty paper coding,” *IEEE Transactions on Information Theory*, vol. 61, no. 9, pp. 4569–4586, 2015.
- [28] I. Csiszár, “The method of types,” *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2505–2523, 1998.
- [29] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2011.
- [30] V. Y. Tan, “Asymptotic estimates in information theory with non-vanishing error probabilities,” *Foundations and Trends® in Communications and Information Theory*, vol. 11, no. 1-2, pp. 1–184, 2014.
- [31] L. Zhou, M. Motani *et al.*, “Finite blocklength lossy source coding for discrete memoryless sources,” *Foundations and Trends® in Communications and Information Theory*, vol. 20, no. 3, pp. 157–389, 2023.
- [32] Y.-W. Huang and P. Moulin, “Finite blocklength coding for multiple access channels,” in *2012 IEEE International Symposium on Information Theory Proceedings*. IEEE, 2012, pp. 831–835.
- [33] J. Scarlett, A. Martinez, and A. G. i Fàbregas, “Second-order rate region of constant-composition codes for the multiple-access channel,” *IEEE Transactions on Information Theory*, vol. 61, no. 1, pp. 157–172, 2014.
- [34] C. T. Li and A. El Gamal, “Strong functional representation lemma and applications to coding theorems,” *IEEE Transactions on Information Theory*, vol. 64, no. 11, pp. 6967–6978, Nov 2018.
- [35] Y. Liu and C. T. Li, “One-shot coding over general noisy networks,” *IEEE Transactions on Information Theory*.
- [36] —, “One-shot information hiding,” in *2024 IEEE Information Theory Workshop (ITW)*, 2024, pp. 169–174.
- [37] Y. Liu, S. H. Advary, and C. T. Li, “Nonasymptotic oblivious relaying and variable-length noisy lossy source coding,” in *2025 IEEE International Symposium on Information Theory (ISIT)*, 2025.
- [38] H. Yamamoto and K. Itoh, “Source coding theory for multiterminal communication systems with a remote source,” *IEICE TRANSACTIONS (1976-1990)*, vol. 63, no. 10, pp. 700–706, 1980.
- [39] D. Rebollo-Monedero and B. Girod, “Generalization of the rate-distortion function for Wyner-Ziv coding of noisy sources in the quadratic-Gaussian case,” in *Data Compression Conference*. IEEE, 2005, pp. 23–32.
- [40] A. Feinstein, “A new basic theorem of information theory,” *IRE Trans. Inf. Theory*, no. 4, pp. 2–22, 1954.
- [41] V. Kostina and S. Verdú, “Channels with cost constraints: Strong converse and dispersion,” *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2415–2429, 2015.
- [42] V. Kostina and S. Verdú, “Fixed-length lossy compression in the finite blocklength regime,” *IEEE Transactions on Information Theory*, vol. 58, no. 6, pp. 3309–3338, June 2012.
- [43] —, “Nonasymptotic noisy lossy source coding,” *IEEE Transactions on Information Theory*, vol. 62, no. 11, pp. 6111–6123, 2016.
- [44] H. Yang, Y. Shi, S. Shao, and X. Yuan, “Joint data and semantics lossy compression: Nonasymptotic converse bounds and second-order asymptotics,” 2024. [Online]. Available: <https://arxiv.org/abs/2402.02501>
- [45] M. H. Yassaee, M. R. Aref, and A. Gohari, “Non-asymptotic output statistics of random binning and its applications,” in *2013 IEEE ISIT*, July 2013, pp. 1849–1853.
- [46] —, “Achievability proof via output statistics of random binning,” *IEEE Transactions on Information Theory*, vol. 60, no. 11, pp. 6760–6786, 2014.
- [47] Y. Oohama, “Exponential strong converse for source coding with side information at the decoder,” *Entropy*, vol. 20, no. 5, p. 352, 2018.
- [48] R. M. Dudley, “Distances of probability measures and random variables,” *Ann. Math. Statist.*, vol. 39, no. 6, pp. 1563–1572, 1968.
- [49] —, *Real analysis and probability*. Chapman and Hall/CRC, 2018.
- [50] P. Billingsley, *Convergence of probability measures*. John Wiley & Sons, 2013.
- [51] J. Liu, “Dispersion bound for the wyner-ahlsvede-körner network via reverse hypercontractivity on types,” in *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2018, pp. 1854–1858.
- [52] R. Chen and T. B. Fomby, “Forecasting with stable seasonal pattern models with an application to Hawaiian tourism data,” *Journal of Business & Economic Statistics*, vol. 17, no. 4, pp. 497–504, 1999.
- [53] T. S. Han, “The reliability functions of the general source with fixed-length coding,” *IEEE Transactions on Information Theory*, vol. 46, no. 6, pp. 2117–2132, 2006.

- [54] A. Jurinskii, "A smoothing inequality for estimates of the Levy-Prokhorov distance," *Theory of Probability & Its Applications*, vol. 20, no. 1, pp. 1–10, 1975.
- [55] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," *IRE Nat. Conv. Rec.*, vol. 4, no. 142-163, p. 1, 1959.
- [56] I. Csiszár, "On an extremum problem of information theory," *Studia Scientiarum Mathematicarum Hungarica*, vol. 9, no. 1, pp. 57–71, 1974.
- [57] R. Dobrushin and B. Tsybakov, "Information transmission with additional noise," *IRE Transactions on Information Theory*, vol. 8, no. 5, pp. 293–304, 1962.
- [58] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Dispersion of Gaussian channels," in *2009 IEEE International Symposium on Information Theory*. IEEE, 2009, pp. 2204–2208.
- [59] S. Watanabe, "Second-order region for Gray-Wyner network," *IEEE Transactions on Information Theory*, vol. 63, no. 2, pp. 1006–1018, 2016.
- [60] R. J. Barron, B. Chen, and G. W. Wornell, "The duality between information embedding and source coding with side information and some applications," *IEEE Transactions on Information Theory*, vol. 49, no. 5, pp. 1159–1180, 2003.
- [61] K. Marton, "A coding theorem for the discrete memoryless broadcast channel," *IEEE Transactions on Information Theory*, vol. 25, no. 3, pp. 306–311, May 1979.
- [62] S. I. Gel'fand and M. S. Pinsker, "Capacity of a broadcast channel with one deterministic component," *Problemy Peredachi Informatsii*, vol. 16, no. 1, pp. 24–34, 1980.
- [63] Y. Liang and G. Kramer, "Rate regions for relay broadcast channels," *IEEE Transactions on Information Theory*, vol. 53, no. 10, pp. 3517–3535, Oct 2007.
- [64] Y. Liang, G. Kramer, and H. V. Poor, "On the equivalence of two achievable regions for the broadcast channel," *IEEE Transactions on Information Theory*, vol. 57, no. 1, pp. 95–100, Jan 2011.
- [65] J. Liu, P. Cuff, and S. Verdú, "One-shot mutual covering lemma and Marton's inner bound with a common message," in *2015 IEEE ISIT*, June 2015, pp. 1457–1461.
- [66] T. M. Cover and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. USA: Wiley-Interscience, 2006.
- [67] L. Devroye, A. Mehrabian, and T. Reddad, "The total variation distance between high-dimensional Gaussians with the same mean," *arXiv preprint arXiv:1810.08693*, 2018.
- [68] A. El Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge University Press, 2011.