Dispersion Bound for the Wyner-Ahlswede-Körner Network via a Semigroup Method on Types

Jingbo Liu, *Member, IEEE*

***Abstract*— We revisit the Wyner-Ahlswede-Körner network, focusing especially on the converse part of the dispersion analysis, which is known to be challenging. Using the functional-entropic duality and the reverse hypercontractivity of the transposition semigroup, we lower bound the error probability for each joint type. Then by averaging the error probability over types, we lower bound the c-dispersion (which characterizes the second-order behavior of the weighted sum of the rates of the two compressors when a nonvanishing error probability is small) as the variance of the gradient of $\inf_{P_{U|X}}\{cH(Y|U)+I(U;X)\}$ with respect to Q_{XY}, the per-letter side information and source distribution. In comparison, using standard achievability arguments based on the method of types, we upper-bound the c-dispersion as the variance of $c\imath_{Y|U}(Y|U)+\imath_{U;X}(U;X)$, which improves the existing upper bounds but has a gap to the aforementioned lower bound. Our converse analysis should be immediately extendable to other distributed source-type problems, such as distributed source coding, common randomness generation, and hypothesis testing with communication constraints. We further present improved bounds for the general image-size problem via our semigroup technique.**

***Index Terms*— Shannon theory, concentration of measure, hypercontractivity, Markov semigroups, functional inequalities, converse bounds, distributed systems.**

I. Introduction

In THE Wyner-Ahlswede-Körner (WAK) problem [1], [2], [49] (Figure 1), a source Y^n and a side information X^n are compressed separately as integers $W_2 = W_2(Y^n)$ and $W_1 = W_1(X^n)$, respectively, and a decoder reconstructs $\hat{Y}^n$ based on W_1 and W_2. Consider the discrete memoryless setting where the per-letter source distribution is Q_{XY}, for any $c > 0$, define

$$\begin{aligned}\phi_c(Q_{XY}) &:= \inf_{P_{U|X}} \{I(U;X) + cH(Y|U)\} \qquad (1)\\ &= \inf_{P_{U|X}} \{I(U;X) - cI(U;Y) + cH(Y)\}\end{aligned}$$

Manuscript received September 12, 2018; revised June 8, 2020; accepted November 24, 2020. Date of publication December 3, 2020; date of current version January 21, 2021. This work was supported in part by the NSF under Grant CCF-1350595, Grant CCF-1016625, Grant CCF-0939370, and Grant DMS-1148711; in part by the ARO under Grant W911NF-15-1-0479, Grant W911NF-14-1-0094, and Grant AFOSR FA9550-15-1-0180; and in part by the Center for Science of Information. This article was presented in part at the 2018 IEEE International Symposium on Information Theory (ISIT).

The author is with the Department of Statistics, University of Illinois at Urbana–Champaign, Champaign, IL 61820 USA (e-mail: jingbol@illinois.edu).

Communicated by A. Gohari, Associate Editor At Large.

Digital Object Identifier 10.1109/TIT.2020.3041791

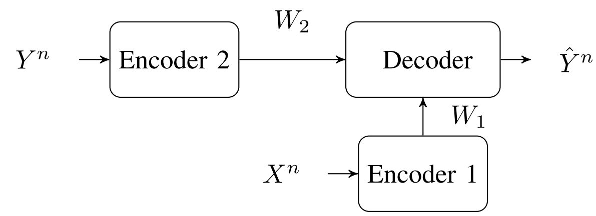

Fig. 1. Source coding with compressed side information.

where $(U, X, Y) \sim P_{U|X}Q_{XY}$, and the cardinality of the auxiliary can be bounded by $|\mathcal{U}| \leqslant |\mathcal{X}| + 2$. The following *strong converse* result was proved in [1] using the *blowing-up lemma* (BUL): if the error probability $\mathbb{P}[\hat{Y}^n \neq Y^n]$ is below some $\epsilon \in (0, 1)$, then

$$\ln|\mathcal{W}_1| + c\ln|\mathcal{W}_2| \geqslant n\phi_c(Q_{XY}) - O\left(\sqrt{n}\ln^{3/2} n\right). \qquad (2)$$

The first-order term in (2) is tight since it is the precise single-letter characterization [1], [49]. Note that for any $c \leqslant 1$, we have $\phi_c(Q_{XY}) = cH(Y)$ by the data processing inequality. Moreover, $\ln|\mathcal{W}_1| + c\ln|\mathcal{W}_2| \geqslant c\ln|\mathcal{W}_1 \times \mathcal{W}_2| \geqslant cnH(Y) - O(\sqrt{n})$, which follows trivially from a method of type analysis [10] of the single source compression problem. Therefore it suffices for us to focus on the case of $c \geqslant 1$.

The study of the *second-order* rates in information theory, initiated by Strassen [41] and recently popularized by Hayashi [14], [15] and Polyanskiy-Poor-Verdú [38], [39], can be viewed as refining the strong converse results. For example, in (2), the term $n\phi_c(Q_{XY})$ is called the first-order approximation, whereas the second-order analysis concerns the asymptotic behavior of the remainder. Common approaches for the second-order analysis include the information spectrum methods (initiated by Polyanskiy-Poor-Verdú [39] and Hayashi [15]) and the method of types (initiated by Ingber and Kochman [16]). In most solved cases from the network information theory (see e.g., the summary in [42]), the second-order term behaves as $\sqrt{nV}\mathrm{Q}^{-1}(\epsilon)$, where V is called the *dispersion* which depends on the channel/source distributions, and Q^{-1} denotes the inverse function of the Gaussian tail probability.

While recent research has succeeded in characterizing the *second-order* rates for various single-user and selected multi-user problems (see e.g., [20], [39], [42]), it remained a formidable challenge to precisely characterize second-order term in (2) for the WAK problem. Indeed, [42, Section 9.2.2, 9.2.3] listed the second-order rate in WAK as a major open problem, since previous converse techniques (e.g. straightforward uses of method of types, information spectrum methods, or meta-converses) appear insufficient for cases where the auxiliary random variable satisfies a Markov chain. In fact, the WAK problem represents a typical challenge in a class of distributed source coding problems (or more generally, distributed source-type problems including common randomness generation [3] or hypothesis testing [4]) involving side information (a.k.a. a helper). Recently, Watanabe [46] examined the converse bound obtained by taking limits in the Gray-Wyner network, yielding a strong converse for WAK but not appearing to improve the second-order term. In [44], Tyagi and Watanabe proposed an approach of dealing with such Markov chain constraints, by replacing it with a bound on the conditional mutual information and then taking the limits. While their approach yields strong converses in interesting applications such as Wyner-Ziv and wiretap channels, it is not clear whether such an approach yields sharper second-order estimates in (2).

To our knowledge, the first proof of an $O(\sqrt{n})$ second-order converse was by a novel semigroup method due to the author and van Handel and Verdú [30].[1] The $O(\sqrt{n})$ rate is sharp for $\epsilon > 1/2$ since an $\Omega(\sqrt{n})$ bound follows by applying the central limit theorem to the standard random coding argument. The converse technique in [30] is based on functional-entropic duality and the reverse hypercontractivity of semigroups, which is widely applicable to multiuser information theory problems. It appears that all previous strong converses via BUL or the image-size characterization in [1], [10] can be upgraded to an $O(\sqrt{n})$ second-order converse (see the further discussions in [22, Chapter 4,5]).

While there have been many applications of (the forward) hypercontractivity, mostly in the converse proofs (see e.g., [17], [26], [31], [32], [50]), the applications of the reverse hypercontractivity have been very rare; see the discussion [33] (see however, an application in [37, Section 4]). Besides the applications in operational problems, some pure mathematical aspects of hypercontractivity and its reverse also received some recent attentions in the information theory community [6]–[8], [18], [24], [25], [34].

The connection between our analysis of the WAK problem and (reverse) hypercontractivity is two-folded: First, the entropic quantity (1) has an equivalent functional formulation which is related to hypercontractivity of the conditional expectation operator $Q_{Y|X}$; see e.g. [5, Theorem 5(b)] and [6, Theorem 1(e)]. This connection has been well-studied in the literature. The second connection, which is the novelty of [30] and the present paper, is the reverse hypercontractivity of certain Markov semigroup operators that replaces the blowing-up lemma machinery. To our knowledge, [30] is among the rare applications of the reverse hypercontractivity of Markov semigroups beyond the obviously related settings such as Markov chain mixing times.

[1] Almost concurrently, Zhou-Tan-Yu-Motani [51] and Oohama [35] amended the technique in an earlier version of Oohama [35] and announced an $O(\sqrt{n})$ converse for WAK. Their proof is based on a novel single-letterization method for the information spectrum together with sophisticated method of types analysis. See also [36].

Let us also remark that the reverse hypercontractivity is known to imply sub-gaussian concentration (i.e. BUL) in rather general settings. Here, the Markov semigroup corresponds to a Dirichlet form which is the inner product of a certain notion of gradients. indeed, reverse hypercontractivity of such a reversible semigroup is equivalent to the modified log-Sobolev inequality [33], which in turn implies concentration by a general convexity argument [45, Lemma 3.16]. Therefore, the improvement on the converse bounds by the reverse hypercontractivity over BUL may be attributed to the fact that we have identified the right (stronger) tool to attack the given problem. More recently, a similar method based on the reverse hypercontractivity of semigroups has successfully improved the (first-order) outer-bound for the relay channel [27], [28].

The idea of [30] is roughly described as follows: first we note that an entropic quantity related to ϕ_c has an equivalent functional formulation (52) which contains quantities such as $\int \ln f dP$. If one directly takes f to be the indicator function of a decoding set, then generally $\int \ln f \mathrm{d}P = -\infty$ which is useless. However, using a machinery called reverse hypercontractivity, we design some "dominating operator" Λ such that $\int \ln(\Lambda f)\mathrm{d}P \geqslant \ln \int f \mathrm{d}P$, and $\int f \mathrm{d}P$ is the probability of correct decoding which we desired. For all source and channel networks where a strong converse was proved in [10], we can now bound the second-order term as $C\sqrt{n \ln \frac{1}{1-\epsilon}}$.

One deficiency of the machinery of [30] is the inability to estimate the sign of the second-order term correctly when $\epsilon < 1/2$: for example, in the WAK problem, one would expect (e.g. by analyzing achievability schemes) that $\ln|\mathcal{W}_1| + c\ln|\mathcal{W}_2| - n\phi_c(Q_{XY})$ is positive when $\epsilon < 1/2$ and negative when $\epsilon > 1/2$, but the method of [30] always lower bound it by $-C\sqrt{n}$, for all $\epsilon \in (0,1)$, where $C > 0$ is some constant depending on Q_{XY}, c and ϵ. In other words, the sign is wrong for $\epsilon < 1/2$. In this paper, we integrate the semigroup technique with the method of types, which is capable of showing, among others, that for sufficiently small (but independent of n) ϵ, we have $\ln|\mathcal{W}_1| + c\ln|\mathcal{W}_2| - n\phi_c(Q_{XY}) \geqslant C\sqrt{n}$ where $C > 0$ is some constant depending on Q_{XY}, c and ϵ. More precisely, let us define the *c-dispersion* as

$$V_c := \lim_{\epsilon\downarrow 0} \limsup_{n\to\infty} \frac{[\ln M_c(n,\epsilon) - n\phi_c(Q_{XY})]^2}{2n\ln\frac{1}{\epsilon}} \tag{3}$$

where $M_c(n,\epsilon)$ denotes the infimum of $|\mathcal{W}_1|\cdot|\mathcal{W}_2|^c$ over codes for which $\mathbb{P}[\mathcal{E}_n] \leqslant \epsilon$. Then, under a mild differentiability assumption on $\phi_c(\cdot)$, we show that

$$V_c \geqslant \mathrm{Var}\left(\nabla\phi_c|_{Q_{XY}}(X,Y)\right) \tag{4}$$

$$= \mathrm{Var}(\mathbb{E}[c\imath_{Y|U}(Y|U) + \imath_{U;X}(U;X)|XY]) \tag{5}$$

where $(U,X,Y) \sim Q_{UXY} := Q_{U|X}Q_{XY}$, $Q_{U|X}$ is *any* infimizer for (1), and we used the notations

$$\imath_{Y|U}(y|u) := \frac{1}{Q_{Y|U}(y|u)}, \quad \forall(y,u); \tag{6}$$

$$\imath_{U;X}(u;x) := \frac{Q_{X|U}(x|u)}{Q_X(x)}, \quad \forall(u,x). \tag{7}$$

We can take $|\mathcal{U}| \leqslant |\mathcal{X}| + 2$ [1]. We remark that the second-order bound $C\sqrt{n \ln \frac{1}{1-\epsilon}}$ (for some constant $C > 0$ depending on c and Q_{XY}) in [30], as well as the strong converse exponent in [36], does not give a nontrivial bound on the dispersion, whereas the present bound (4) is analogous to the dispersion formula in most other previously solved problems from the network information theory. Let us also remark, however, that (4) only captures the asymptotics as $\epsilon \to 0$; we do not show a second-order converse bound of the form $\sqrt{nV}\mathrm{Q}^{-1}(\epsilon)$ for *any* ϵ.

Instead of the semigroup associated with the Glauber dynamics (a.k.a. Gibbs sampler) used in [30], the present paper uses the transposition semigroup and its reverse hypercontractivity estimate in [13], which is a tailor-made for type-class analysis. We remark that the reverse hypercontractivity estimate for the transposition semigroup is order-wise the same as the Glauber dynamics; see [13].

Apart from reverse hypercontractivity and functional-entropic duality, another interesting ingredient in our proof is an argument in analyzing certain information quantity for the equiprobable distribution on a type class (Lemma 2). We need this because, unlike Glauber dynamics, the standard tensorization argument for the reverse hypercontractivity (see e.g., [33]) does not apply to the transposition semigroup, and an interesting induction argument is used instead. More precisely, we perform an algebraic expansion employing the symmetry of the type class, but which is different from the standard tensorization argument for the i.i.d. distribution. This gives rise to a certain martingale, whose variance equals the gap to the same quantity evaluated for the i.i.d. distribution.

On the achievability side, [48] previously showed that $V_c \leqslant \left(\sqrt{\mathrm{Var}(c\imath_{Y|U}(Y|U))} + \sqrt{\mathrm{Var}(\imath_{U;X}(U;X))}\right)^2$. In this paper we use the method of types to show an improved bound

$$V_c \leqslant \mathrm{Var}\left(c\imath_{Y|U}(Y|U) + \imath_{U;X}(U;X)\right). \tag{8}$$

The achievability proof uses standard techniques. We remark that generally, (5) can indeed be smaller than the right side of (8), due to the reduction of the variance by conditioning on the sources. In contrast, note that in the single source lossy compression problem, conditioning on the source does *not* decrease the variance (see Proposition 1.1 ahead).

The paper is organized as follows: the main results about the converse and achievability bounds on the dispersion are presented in Section II. We review and compare with the classical (pre)image-size approach [1], [10] in Section III, and present our new estimates via the semigroup method on types. Some basic properties of the related information-theoretic quantities are discussed in Section IV, which serves as a preparation for the converse and achievability proofs of the main results in Section V and VI. The second one is the new one which we intended to refer to in the title.

Notation. Given an alphabet $\mathcal{Y}$, define $\mathcal{H}_+(\mathcal{Y})$ as the set of all nonnegative functions on $\mathcal{Y}$, and $\mathcal{H}_{[0,1]}(\mathcal{Y})$ the set of functions from $\mathcal{Y}$ to $[0,1]$. For $f \in \mathcal{H}_+(\mathcal{Y})$, define $P(f) := \mathbb{E}_P[f]$, and define $P_{Y|X}(f) := \mathbb{E}_{P_{Y|X=\cdot}}[f(Y)]$ as a function on $\mathcal{H}_+(\mathcal{X})$. Given an n-type P_{XY}, let $\mathcal{T}_n(P_X)$ be the set of all x^n with type P_X, and $\mathcal{T}_{x^n}(P_{Y|X})$ the set of all y^n such that (x^n, y^n) is type P_{XY}. The total variation distance is denoted by $|P - Q|$. We use $P_X \to P_{Y|X} \to P_Y$ to define an output distribution P_X for a given input and a random transformation. Define the Gaussian tail probability $\mathrm{Q}(t) := \int_t^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \mathrm{d}x$, for $t \in \mathbb{R}$. The bases for all exponentials and entropic quantities are natural. Unless otherwise stated, the constants used in bounding (e.g. E,F,G,λ) may depend on c and Q_{XY}. Given a probability measure Q on $\mathcal{X}$, and a functional ϕ on the probability simplex $\Delta_{\mathcal{X}}$, the gradient $\nabla\, \phi|_Q$ can be regarded as a function on $\mathcal{X}$, and $\langle \nabla\, \phi|_Q\,, Q\rangle := \int \nabla\, \phi|_Q\, \mathrm{d}Q$.

II. Main Results

We will analyze and present our dispersion bounds using the weighted sum-rate formulation; see e.g. [42, Section 6.4]. That is, for any $c \geqslant 0$ and $D \in \mathbb{R}$, we estimate the error probability of the best code for which $\ln|\mathcal{W}_1| + c\ln|\mathcal{W}_2| = n\phi_c(Q_{XY}) + D\sqrt{n}$. Furthermore we only need to focus on the $c \geqslant 1$ case, since otherwise, as argued at the beginning of the introduction, the problem degenerates into single-source coding.

A. Converse

All converse analysis in this paper assumes finite alphabets $\mathcal{X}$ and $\mathcal{Y}$, and that $\phi_c(\cdot)$ (defined in (1)) has bounded second derivatives in a neighborhood of Q_{XY}. The main ingredient of the converse proof is the following bound in the case where the source sequences are equiprobably distributed on a given type class.

Lemma 1: Given Q_{XY} and $c \geqslant 1$, there exists $\lambda \in (0,1)$ and $E > 0$ such that the following holds: For any $n \geqslant 1$ and n-type P_{XY} such that $|P_{XY} - Q_{XY}| \leqslant \lambda$, let (X^n, Y^n) be equiprobable on the type class $\mathcal{T}_n(P_{XY})$. If there exists a WAK coding scheme for (X^n, Y^n) with error probability $\epsilon \in (0,1)$, then

$$\begin{aligned} &\ln|\mathcal{W}_1| + c\ln|\mathcal{W}_2| \\ &\geqslant n\phi_c(P_{XY}) - 2c\sqrt{\frac{n}{\min_x P_X(x)} \ln\frac{1}{1-\epsilon}} \\ &\quad - E\ln n + c\ln(1-\epsilon). \end{aligned} \tag{9}$$

By averaging the error probability bounded in Lemma 1 over the types, we obtain the following converse for stationary memoryless sources:

Theorem 1: Fix $c \geqslant 1$, $D \in \mathbb{R}$, and Q_{XY}. Let $(X^n, Y^n) \sim Q_{XY}^{\otimes n}$. If a WAK coding scheme satisfies

$$\ln|\mathcal{W}_1| + c\ln|\mathcal{W}_2| \leqslant n\phi_c(Q_{XY}) + D\sqrt{n} \tag{10}$$

for all n then we lower bound the error probability

$$\liminf_{n\to\infty} \mathbb{P}[\mathcal{E}_n] \geqslant \sup_{\delta\in(0,1)} \delta\mathrm{Q}\left(\frac{D + c\sqrt{\frac{8}{\min_x Q_X(x)} \ln\frac{1}{1-\delta}}}{\sqrt{\mathrm{Var}(\nabla\phi_c|_{Q_{XY}}(X,Y))}}\right). \tag{11}$$

In particular, the c-dispersion (see (3)) satisfies

$$\begin{aligned} V_c &\geqslant \mathrm{Var}(\nabla\phi_c|_{Q_{XY}}(X,Y)) && (12)\\ &= \mathrm{Var}(\mathbb{E}[c\imath_{Y|U}(Y|U) + \imath_{U;X}(U;X)|XY]). && (13) \end{aligned}$$

Proofs of Lemma 1 and (11) are given in Section V.

B. Achievability

As alluded, we also use the method of types (in lieu of the information spectrum approach of [48]) to obtain the following improved upper bound on c-dispersion.

Theorem 2: Fix Q_{XY} on finite alphabets, $c \geqslant 1$, and $D \in \mathbb{R}$. Then there exists a WAK scheme scheme such that

$$\ln|\mathcal{W}_1| + c\ln|\mathcal{W}_2| \leqslant n\phi_c(Q_{XY}) + D\sqrt{n}, \quad \forall n; \tag{14}$$

$$\limsup_{n\to\infty} \mathbb{P}[\mathcal{E}_n] \leqslant \mathrm{Q}\left(\frac{D}{\sqrt{\mathrm{Var}\left(c\imath_{Y|U}(Y|U)+\imath_{U;X}(U;X)\right)}}\right). \tag{15}$$

In particular, the c-dispersion (see (3)) satisfies

$$V_c \leqslant \mathrm{Var}\left(c\imath_{Y|U}(Y|U) + \imath_{U;X}(U;X)\right). \tag{16}$$

Proof of Theorem 2 is given in Section VI. We remark that generally there is a gap between the bounds on c-dispersion in Theorem 1 and Theorem 2.

III. Image-Size and Preimage-Size Bounds

In this section, we review the "preimage-size" method in the original work of Ahlswede *et al.* [1] and the related "image-size" method popularized by the classical book of Csiszár and Körner [10], and discuss how the techniques in the present paper will bear on them. These methods, introduced more than 40 years ago, have remained the only way towards the strong converse of a number of multiuser information theory problems in [10] until recently (see e.g. the discussion in [42]). The present paper does not use the precise forms of these images, and the current section is not at all needed for understanding the proofs of the converse results in Section II-A. Yet, for the interest of readers who are conversant with the classical literature, we will present here some new estimates for the (pre)image-size for a type class using our semigroup technique.

Consider a random transformation $Q_{Y^n|X^n}$ between finite sets $\mathcal{X}^n$ and $\mathcal{Y}^n$, where $n \geqslant 1$. Let $\epsilon \in (0,1)$, and $\mathcal{B} \subseteq \mathcal{Y}^n$. In [1] the authors introduced

$$\Psi_\epsilon(\mathcal{B}) := \{x^n \in \mathcal{X}^n : Q_{Y^n|X^n}[\mathcal{B}|x^n] \geqslant 1-\epsilon\}. \tag{17}$$

For ease of comparison with the image-size later, we shall call $\Psi_\epsilon(\mathcal{B})$ the $(1-\epsilon)$-preimage of $\mathcal{B}$. Now let us fix a $Q_{X^nY^n}$ compatible with $Q_{Y^n|X^n}$, and an arbitrary reference measure ν_{Y^n} (which is set as the counting measure in the application to the WAK problem). For any $a \in \mathbb{R}$ and $\epsilon \in (0,1)$, [1] defined

$$\hat{S}_n(c,\epsilon) := \frac{1}{n}\ln\min \nu_{Y^n}[\mathcal{B}] \tag{18}$$

where the minimum is over $\mathcal{B}$ such that

$$\frac{1}{n}\ln Q_{X^n}(\Psi_\epsilon(\mathcal{B}) \cap \mathcal{C}_n) \geqslant a \tag{19}$$

and $\mathcal{C}_n$ denotes the strongly typical set, which is the set of x^n such that the deviation of the empirical distribution from Q_X is in $\omega(\frac{1}{\sqrt{n}}) \cap o(1)$ (see [1, Definition 4]). The main technical result of [1] is that, in the i.i.d. case where $Q_{X^nY^n} = Q_{XY}^{\otimes n}$ and $\nu_{Y^n} = \nu_Y^{\otimes n}$, we have

$$\lim_{n\to\infty} \hat{S}_n(c,\epsilon) = \min\{-D(P_{Y|U}\|\nu_Y|P_U)\} \tag{20}$$

where the minimum is over P_{UXY} such that $P_{XY} = Q_{XY}$ and $-I(U;X) \geqslant a$. The constraint of $P_X = Q_X$, which is important to the information-theoretic applications, comes from the restriction to typical sets. If the intersection with the strongly typical set was removed in (19), then a similar result would hold in (20), with $P_{XY} = Q_{XY}$ relaxed to $P_{Y|X} = Q_{Y|X}$ in the constraints of the minimization over P_{UXY} [1, Lemma 1A].

Since a closed convex set can be expressed as the intersection of closed half spaces, the above results from [1] can be equivalently stated in terms of extremal sum rates:

$$\lim_{n\to\infty}\sup_{\mathcal{B}\subseteq\mathcal{Y}^n}\left\{\frac{1}{n}\ln Q_{X^n}[\Psi_\epsilon(\mathcal{B})\cap\mathcal{C}_n] - \frac{c}{n}\ln\nu_{Y^n}[\mathcal{B}]\right\}$$
$$= \sup_{P_{U|X}}\{-I(U;X) + cD(P_{Y|U}\|\nu_Y|P_U)\} \tag{21}$$

and

$$\lim_{n\to\infty}\sup_{\mathcal{B}\subseteq\mathcal{Y}^n}\left\{\frac{1}{n}\ln Q_{X^n}[\Psi_\epsilon(\mathcal{B})] - \frac{c}{n}\ln\nu_{Y^n}[\mathcal{B}]\right\}$$
$$= \sup_{P_{XY}:\, P_{Y|X}=Q_{Y|X}}\{-D(P_X\|Q_X) + cD(P_Y\|\nu_Y)\} \tag{22}$$

for all $c \geqslant 0$ (equivalently, for all $c \geqslant 1$, since this range is enough to characterize the convex set, which can be seen by the strong data processing inequality). Note that there is no need of conditioning on U on the right side of (22), since U in [1, Lemma 1A] only serves to convexify the region, whereas the sum-rate characterization corresponds to intersection of half-spaces which is automatically convex.

Let us now turn to the related concept of image-size. Consider $Q_{Y^n|X^n}$ and ν_{Y^n} as above. The following is taken from [10, Definition 6.2]:

Definition 1: For any $\eta \in (0,1)$, an η-image of $\mathcal{A} \subseteq \mathcal{X}^n$ is any set $\mathcal{B} \subseteq \mathcal{Y}^n$ for which $Q_{Y^n|X^n}[\mathcal{B}|x^n] \geqslant \eta$, for all $x^n \in \mathcal{A}$. Define the η-image size of $\mathcal{A}$

$$\mathrm{g}(\mathcal{A},\nu_{Y^n},\eta) := \min_{\mathcal{B}:\ \eta\text{-image of }\mathcal{A}} \nu_{Y^n}[\mathcal{B}]. \tag{23}$$

In many applications such as the WAK problem, we are interested in maximizing the measure of a set while minimizing the corresponding image size. Since the preimage of the image of a set always contains the original set, it is easy to see that the image-size problem is equivalent to the preimage-size problem in such settings.

Let us remark that the BUL technique of [1] shows that the gap between the two supremums on the two sides of (21), i.e., the second-order term is $O(\sqrt{n}\ln^{3/2} n)$, whereas in [29] we used the semigroup technique to show that[2] the second-order term is $O(\sqrt{n})$ which is order-wise optimal. Recall that the method of [1] proceeds by first defining a measure P_{X^n} which is the conditioning of Q_{X^n} on the r-blowup of a certain preimage, and then connect the probability of sets and

[2]More precisely, we did not explicitly state an image-size bound in [29], but such a result is clearly contained in our proof of the WAK converse therein.

relative entropy terms via the data processing inequality. It is clear from their proof that, if an $O(\sqrt{n})$ second-order bound were possible via the BUL technique, then r must satisfy the following conditions:

1) For any $\mathcal{A} \subseteq \mathcal{X}^n$ such that $Q_{X^n}[\mathcal{A}]$ is bounded away from 0 and 1, the r-blowup of $\mathcal{A}$, denoted as $\mathcal{A}^r$, satisfies $Q_{X^n}[\mathcal{A}^r] = 1 - O(n^{-1/2})$.
2) For such $\mathcal{A}$ we also have the cardinality bound $|\mathcal{A}^r| \leqslant e^{O(\sqrt{n})}|\mathcal{A}|$.

In the stationary memoryless setting it is easy to see that there is no r satisfying the two conditions simultaneously, and a careful analysis (see e.g., [29]) reveals that $O(\sqrt{n}\ln^{3/2} n)$ is the best second-order bound via BUL. In fact, the sub-optimality is not only because of BUL, but also inherent in the data-processing argument: an analysis in [29] shows that it is impossible to design a new set transformation $\mathcal{A} \mapsto \tilde{\mathcal{A}}$ so that the data-processing argument gives a $o(\sqrt{n \ln n})$ second-order bound.

We now turn to (pre)image-size analysis for type classes. Below we shall define $\Psi_\epsilon(\cdot)$ and $\mathrm{g}(\cdot)$ similarly as (17) and (23) but with the i.i.d. distribution $Q_{X^nY^n}$ replaced by $P_{X^nY^n}$ which is the equiprobable distribution on some n-type P_{XY}. The first-order terms are the same as (21) and (22). The BUL approach does not yield an $O(\sqrt{n})$ second-order term, since we can verify that when the i.i.d. distribution Q_{X^n} is replaced by the equiprobable distribution P_{X^n} on $\mathcal{T}_n(P_X)$, it is still impossible to choose r to fulfill conditions 1) and 2) above. In contrast, we can indeed prove an $O(\sqrt{n})$ second-order bound using our semigroup techniques:

Theorem 3: Given Q_{XY} and $c \in [0, \infty)$, there exists $\lambda \in (0, 1)$ and $E > 0$ such that the following holds: Let $n \geqslant 1$ and P_{XY} be an n-type such that $|P_{XY} - Q_{XY}| \leqslant \lambda$. Let $P_{X^nY^n}$ be the equiprobable distribution on $\mathcal{T}_n(P_{XY})$. For any $\epsilon \in (0, 1)$, and $\mathcal{B} \subseteq \mathcal{Y}^n$, we have

$$\begin{aligned} &\ln P_{X^n}[\Psi_\epsilon(\mathcal{B})] \\ &\leqslant c \ln P_{Y^n}[\mathcal{B}] + n \sup_{P_{U|X}} \{cI(U; Y) - I(U; X)\} \\ &\quad + E \ln n + \frac{c}{1-\epsilon} + 2c\sqrt{\frac{n}{\min_x P_X(x)} \ln \frac{1}{1-\epsilon}}. \end{aligned} \tag{24}$$

Equivalently, for any $\mathcal{A} \subseteq \mathcal{X}^n$,

$$\begin{aligned} &\ln P_{X^n}[\mathcal{A}] \\ &\leqslant c \ln \mathrm{g}(\mathcal{A}, P_{Y^n}, 1-\epsilon) + n \sup_{P_{U|X}} \{cI(U; Y) - I(U; X)\} \\ &\quad + E \ln n + \frac{c}{1-\epsilon} + 2c\sqrt{\frac{n}{\min_x P_X(x)} \ln \frac{1}{1-\epsilon}}. \end{aligned} \tag{25}$$

Remark 1: The first order terms in (24) and (25) are optimal, as in the i.i.d. case (21) and (22), since otherwise we would obtain a converse result that contradicts the achievability. The first-order tightness can also be argued directly as in the proof of [1, Theorem 1,2]: Let $P_{UXY} = P_{U|X}P_{XY}$ where $P_{U|X}$ is a maximizer for (24). Assume without loss of generality that $\mathcal{U} = \{1, \ldots, |\mathcal{U}|\}$. For each u define the integer $n_u := nP_U(u)$; if $P_U(u)$ is not an integer we can perturb $P_{U|X}$ by $O(1/n)$ so that it is, and the proof will still work by a continuity argument. Let (α_n) be a sequence in $\omega(\frac{1}{\sqrt{n}}) \cap o(1)$. Define $x^{(u)} := (x_{n_1+\cdots+n_{u-1}+1}, \ldots, x_{n_1+\cdots+n_u}) \in \mathcal{X}^{n_u}$, and define $y^{(u)}$ similarly. Put $\mathcal{B} = \{y^n : |\widehat{P}_{y^{(u)}} - P_{Y|U=u}| \leqslant \alpha_n, \forall u\}$. We have the following estimates from standard large deviation analysis (e.g. writing the probabilities in terms of combinatorial numbers and approximating them via the Stirling formula):

$$P_{Y^n}[\mathcal{B}] = e^{-nI(U;Y)+o(n)} \tag{26}$$

and

$$P_{X^n}[x^n : x^{(u)} \in \mathcal{T}_{nP_U(u)}(P_{X|U=u}), \forall u] = e^{-nI(U;X)+o(n)}. \tag{27}$$

In the above we assumed that $P_{X|U=u}$ is an $nP_U(u)$-type (if not, we can still perturb $P_{U|X}$ and P_X by $O(1/n)$ to ensure it is, and continue the proof with a continuity argument). For large n we have $\{x^n : x^{(u)} \in \mathcal{T}_{nP_U(u)}(P_{X|U=u}), \forall u\} \subseteq \Psi_\epsilon(\mathcal{B})$, so $\Psi_\epsilon(\mathcal{B})$ is lower bounded by the right $e^{-nI(U;X)+o(n)}$. This shows the first-order tightness.

Some steps of the proof overlap with that of Lemma 1, which we shall avoid repeating.

Proof of Theorem 3: Using the same arguments from the line (46) to (49) ahead, we see that, for any $c \geqslant 0$ (note that these steps do not require $c \geqslant 1$), $\mathcal{B} \subseteq \mathcal{Y}^n$, and $t > 0$,

$$\begin{aligned} &\int P^{c(1+\frac{1}{t})}_{Y^n|X^n}[\mathcal{B}|x^n] \mathrm{d}P_{X^n}(x^n) \\ &\leqslant e^d \exp_e\left(\frac{nt}{\min_x P_X(x)}\right) P^c_{Y^n}[\mathcal{B}] \end{aligned} \tag{28}$$

where

$$d := \sup_{S_{X^n}} \{cD(S_{Y^n} \| P_{Y^n}) - D(S_{X^n} \| P_{X^n})\} \tag{29}$$

in which the supremum is over S_{X^n} supported on $\mathcal{T}_n(P_X)$, and S_{Y^n} is induced by S_{X^n} and the random transformation $P_{Y^n|X^n}$. Using Markov's inequality, for any $\epsilon \in (0, 1)$, we can lower bound the left side of (28) by $P_{X^n}[\Psi_\epsilon(\mathcal{B})](1-\epsilon)^{c(1+\frac{1}{t})}$. Thus we have obtained

$$\begin{aligned} &P_{X^n}[\Psi_\epsilon(\mathcal{B})] \\ &\leqslant P^c_{Y^n}[\mathcal{B}] \exp_e\left(d + c \inf_{t>0}\left\{\frac{nt}{\min_x P_X(x)} + \left(1 + \frac{1}{t}\right)\ln\frac{1}{1-\epsilon}\right\}\right) \end{aligned} \tag{30}$$

$$= P^c_{Y^n}[\mathcal{B}] \exp_e\left(d + c \ln \frac{1}{1-\epsilon} + 2c\sqrt{\frac{n}{\min_x P_X(x)} \ln \frac{1}{1-\epsilon}}\right). \tag{31}$$

In Lemma 2 ahead we show that there exists E depending only on Q_{XY} and c such that

$$d \leqslant c \ln |\mathcal{T}_n(P_Y)| - n\phi_c(P_{XY}) + E \ln n \tag{32}$$

$$\leqslant nH(P_Y) - n\phi_c(P_{XY}) + E \ln n \tag{33}$$

$$= \sup_{P_{U|X}} \{cI(U; Y) - I(U; X)\} + E \ln n \tag{34}$$

Therefore we have established (24). Then (25) is obtained by setting $\mathcal{B}$ to be the $(1-\epsilon)$-image of $\mathcal{A}$, and noting that $\mathcal{A} \subseteq \Psi_\epsilon(\mathcal{B})$. $\square$

Once a preimage size bound such as (24) is obtained, one can connect it to the error probability in the WAK problem with a reverse Markov inequality argument, as did in [1] as well as our previous work [29]; see Remark 3 ahead. There is an extra parameter created in the reverse Markov inequality step, corresponding to the parameter ϵ in $\Psi_\epsilon(\cdot)$, which needs to be chosen carefully when one cares about the prefactor in the second-order rate. As demonstrated in the proof of Lemma 1 in the present paper, however, these extra steps of applying the reverse Markov inequality and introducing the preimage are not necessary, and the whole proof can proceed more seamlessly using the functional inequalities.

In some other applications, however, it still seems convenient to use the notation of image-size. For example, the set $\mathcal{A}$ in Definition 1 may be naturally associated with a codebook; see the proof of the strong converse of the asymmetric broadcast channel with a common message in [10]. For stationary memoryless settings, it is shown in [22, Chapter 5] that the semigroup and functional inequality techniques still apply in those more involved image-size applications, yielding the optimal $\sqrt{n}$ rate, and simplifying the proofs (e.g. avoiding the "maximal code lemma" previously used in [10, Chapter 16]). It should be possible to extend the approach in [22, Chapter 5] using semigroup on types to further estimate the second-order prefactors in the corresponding coding theorems.

IV. Basic Properties of the Single-Letter Expression

To better interpret our results and prepare for the proofs, it is instructive to understand some of the basic properties of the single-letter rate expressions. To fix ideas, let us first recall the situation in the simpler and well-studied problem of lossy compression of a single source (see e.g. [21]). In that problem, we are given a single source with per-letter distribution Q_X, and a per-letter distortion $\mathrm{d}\colon \mathcal{U}\times\mathcal{X} \to \mathbb{R}$ on the reconstruction alphabet and the source alphabet.

Proposition 1: If $P_{U|X}$ is an optimizer for $\varphi_\lambda(Q_X) := \inf_{P_{U|X}}\{I(U;X) + \lambda\mathbb{E}[\mathrm{d}(U;X)]\}$, then the stationarity condition implies that the $\imath_{U;X}(u;x) + \lambda\,\mathrm{d}(u,x)$ is

1) independent of u, $P_{U|X}Q_X$-a.s., and in particular, $\mathrm{Var}(\mathbb{E}[\imath_{U;X}(U;X) + \lambda\,\mathrm{d}(U,X)|X]) = \mathrm{Var}(\imath_{U;X}(U;X) + \lambda\,\mathrm{d}(U,X))$, where $(U,X) \sim P_{U|X}Q_X$;
2) equal to $\nabla\varphi_\lambda|_{Q_X}(x)$, regardless of the choice of the optimal $P_{U|X}$. It is known (e.g. [19]) that the dispersion equals $\mathrm{Var}\left(\nabla\varphi_\lambda|_{Q_X}(X)\right)$.

Now in WAK, our lower and upper bounds on the dispersion in Theorem 1 and Theorem 2, although different, are both analogous to the solution in single-user lossy source coding in certain senses. More precisely, we observe Proposition 2 and Proposition 3 below, which are parallel to the two properties listed above for single-user lossy compression.

Proposition 2: For any Q_{XY} on finite alphabets and $c \geqslant 1$, let $P_{U|X}$ be optimal in the definition of $\phi_c(\cdot)$ in (1) and suppose that $\mathcal{U}$ is finite[3]. Then $\mathbb{E}[c\imath_{Y|U}(Y|U) + \imath_{U;X}(U;X)|UX]$ is independent of U almost surely.

Proof: Let us introduce the notations

$$f(P'_{U|X}) = cH(Y'|U') + I(U';X'), \tag{35}$$

$$g(u,x,y) = c\imath_{Y|U}(y|u) + \imath_{U;X}(u;x), \tag{36}$$

where $P'_{U|X}$ is any random transformation from $\mathcal{X}$ to $\mathcal{U}$, and $(U',X',Y') \sim P'_{U|X}Q_{XY}$. The first order term in the Taylor expansion of $f(P'_{U|X}) - f(P_{U|X})$ equals $\sum_{u,x,y} g(u,x,y)(P'_{U|X} - P_{U|X})(u|x)Q_{XY}(x,y)$, which must vanish by the optimality of $P_{U|X}$. In particular, fix any x such that $Q_X(x) \neq 0$, and consider $P'_{U|X}$ such that $(P'_{U|X} - P_{U|X})(u|x') = 0$ for all u, unless $x' = x$. By the first order condition for such $P'_{U|X}$, we have

$$\sum_u \mathbb{E}[g(U,X,Y)|U=u, X=x](P'_{U|X} - P_{U|X})(u|x) = 0. \tag{37}$$

This shows that for that particular x, $(\mathbb{E}[g(U,X,Y)|U=u,X=x])_{u\colon P_{U|X}(u|x)>0}$, viewed as a vector of dimension $|\{u\colon P_{U|X}(u|x) > 0\}|$, is orthogonal to the subspace of vectors whose coordinates sum to zero, so itself must be a vector with constant coordinates. This means that $\mathbb{E}[g(U,X,Y)|U=u,X=x]$ is independent of u, $P_{U|X}Q_X$-almost surely. $\square$

We remark that $c\imath_{Y|U}(y|u) + \imath_{U;X}(u;x)$ is generally not independent of u. Below is an explicit example.

Binary symmetric sources: Suppose that X and Y are both equiprobable on $\{-1,1\}$ and $\mathbb{E}[XY] = \rho$. Consider any $c \in [\rho^{-2},\infty)$. Remark that $c < \rho^{-2}$ is the degenerate case since ρ^2 is the strong data processing constant. Let U be equiprobable on $\{-1,1\}$ and such that $U - X - Y$ and $\mathbb{E}[UX] = \eta$, where η is defined as the solution to

$$c = \frac{\ln\frac{1+\eta}{1-\eta}}{\rho\ln\frac{1+\eta\rho}{1-\eta\rho}}. \tag{38}$$

Note that the η satisfying (38) maximizes $cH(Y|U)+I(U;X)$ for the given c and ρ. Using Mrs. Gerbers lemma (see e.g. [11]) one can show that such $P_{U|X}$ is an infimizer for (1). We can compute that

$$c\imath_{Y|U}(1|1) + \imath_{U;X}(1;1) = c\ln\frac{2}{1+\rho\eta} + \ln(1+\eta); \tag{39}$$

$$c\imath_{Y|U}(1|-1) + \imath_{U;X}(-1;1) = c\ln\frac{2}{1-\rho\eta} + \ln(1-\eta). \tag{40}$$

Proposition 3: For any Q_{XY} on finite alphabets and $c \geqslant 1$, suppose that $\phi_c(\cdot)$ is differentiable at Q_{XY}. Then for any optimal $P_{U|X}$,

$$\mathbb{E}[c\imath_{Y|U}(Y|U) + \imath_{U;X}(U;X)|X,Y] = \nabla\,\phi_c\big|_Q(X,Y). \tag{41}$$

In particular, the left side does not depend on the choice of the optimal $P_{U|X}$.

[3]This is merely a simplifying assumption and is without loss of generality. Indeed, Carathéodory's theorem implies that one can take $|\mathcal{U}| \leqslant |\mathcal{X}| + 2$ [1].

Proof: Recall that $\phi_c(Q_{XY})$ is defined as the infimum of $cH(Y|U) + I(U;X)$ over $P_{U|X}$. As a general fact, the derivative of an infimum equals the partial derivative of the objective function evaluated at an optimizer, under suitable differentiability assumption (see e.g. the calculation in [23, Lemma 13]. Now for fixed $P_{U|X}$, the partial derivative of $cH(Y|U)+I(U;X)$ with respect to Q_{XY} equals the left side of (41). □

V. Proof of the Converse

A. Overview

1) Fixed Composition Argument: Imagine that a genie tells both encoders and the decoder the joint type of the source, and they all design coding strategies for each type. A converse for this oracle setup gives a converse to the original problem in the stationary memoryless setting. For the purpose of second-order rate analysis, the intuition behind many previously solved problems (including the challenging ones such as the Gray-Wyner network [47]) may be described as follows. One roughly sees a dichotomy: for some "good types", the error probability essentially equals 0, and for the rest "bad types", the error probability is essentially 1. Thus the total error probability is tightly approximated by the probability of those "bad types".

2) Challenge of Fixed Composition Argument for WAK: In network information theory problems where the auxiliary in the single-letter expression satisfies a Markov chain, a straightforward fixed composition argument as described above does not seem to give even a strong converse (unsurprisingly, since otherwise the authors of [1], [10] who are familiar with the method of types would not have needed the blowing-up lemma for strong converses in these works). Moreover, when the blowing-up lemma is applied to a type class, we can verify that the second-order term is still $O(\sqrt{n}\ln^{3/2} n)$, no better than the i.i.d. case. In fact, the above 0-1 dichotomy may not be true when the auxiliary satisfies a Markov chain.

3) New Machineries: In the present paper, we perform fixed composition analysis in the nonvanishing error regime, and the second-order rate is improved to $O(\sqrt{n})$. In lieu of the blowing-up lemma, we use the dual representation of $\phi_c(\cdot)$ as well as a semigroup technique – both ingredients integrate naturally and are responsible for the improved rates. We remark that the $O(\sqrt{n})$ rate is the same order as the i.i.d. case solved in [30]; If it were $o(\sqrt{n})$ instead, we could have bounded the second-order term prefactor as $\sqrt{\mathrm{Var}\left(\nabla\phi_c|_{Q_{XY}}(X,Y)\right)}\mathrm{Q}^{-1}(\epsilon)$ for *each* $\epsilon \in (0,1)$. In reality, the bound on the second-order prefactor is not so clean, and involves nuisance constants depending on Q_X (see (11)). However, the nuisance constants disappear in computing *dispersion* where we take $\epsilon \to 0$. A technical part of the converse is to show that there exists a certain "dominating operator" Λ satisfying desired norm estimates. This is done in Section VII, where we use the estimate of the modified log-Sobolev constant in [13].

B. Proof of Lemma 1

Suppose that $f\colon \mathcal{X}^n \to \mathcal{W}_1$, $g\colon \mathcal{Y}^n \to \mathcal{W}_2$ are the encoders, and $V\colon \mathcal{W}_1 \times \mathcal{W}_2 \mapsto \mathcal{Y}^n$ denotes the decoder. For each $w \in \mathcal{W}_1$, define the "correctly decodable set":

$$\mathcal{B}_w := \{y^n \colon V(w, g(y^n)) = y^n\}. \tag{42}$$

Let $P_{X^nY^n}$ be the equiprobable distribution on $\mathcal{T}_n(P_{XY})$, and let $P_{Y^n|X^n}$ be the induced random transformation. By the assumption,

$$\int P_{Y^n|X^n}[\mathcal{B}_{f(x^n)}|x^n]\mathrm{d}P_{X^n}(x^n) \geqslant 1-\epsilon. \tag{43}$$

Next, we lower bound the error probability using the functional inequality and reverse hypercontractivity approach. We introduce a "dominating" linear operator $\Lambda_{n,t}\colon \mathcal{H}_+(\mathcal{Y}) \to \mathcal{H}_+(\mathcal{Y})$, apply it to the indicator function of a decodable set, and plug the resulting function into the functional inequality. To streamline the presentation, we postpone the definition of $\Lambda_{n,t}$ to (171). The key properties of $\Lambda_{n,t}$, the proofs of which is deferred to Section VII, are the following: for $f \in \mathcal{H}_{[0,1]}(\mathcal{Y}^n)$ and $t = 1/\sqrt{n}$,

- **Lower bound** (173) $P_{Y^n|X^n}(\ln\Lambda_{n,t}f) \geqslant O(\sqrt{n})\ln P_{Y^n|X^n}(f)$,
- **Upper bound** (178) $P_{Y^n}(\Lambda_{n,t}f) \leqslant \exp(O(\sqrt{n}))P_{Y^n}(f)$.

Now, for any $t > 0$,

$$\begin{aligned}
&(1-\epsilon)^{c\left(1+\frac{1}{t}\right)} \\
&\leqslant \int P^{c\left(1+\frac{1}{t}\right)}_{Y^n|X^n}[\mathcal{B}_{f(x^n)}|x^n]\mathrm{d}P_{X^n}(x^n) && (44)\\
&= \sum_{w\in\mathcal{W}_1}\int_{x^n\colon f(x^n)=w} P^{c\left(1+\frac{1}{t}\right)}_{Y^n|X^n}[\mathcal{B}_w|x^n]\mathrm{d}P_{X^n}(x^n) && (45)\\
&\leqslant |\mathcal{W}_1|\int P^{c\left(1+\frac{1}{t}\right)}_{Y^n|X^n}[\mathcal{B}_{w*}|x^n]\mathrm{d}P_{X^n}(x^n) && (46)\\
&\leqslant |\mathcal{W}_1|\int \exp_e\left(cP_{Y^n|X^n=x^n}(\ln\Lambda_{n,t}1_{\mathcal{B}_{w*}})\right)\mathrm{d}P_{X^n} && (47)\\
&\leqslant e^d|\mathcal{W}_1|P^c_{Y^n}(\Lambda_{n,t}1_{\mathcal{B}_{w*}}) && (48)\\
&\leqslant e^d|\mathcal{W}_1|\exp_e\left(\frac{cnt}{\min_x P_X(x)}\right)P^c_{Y^n}[\mathcal{B}_{w*}] && (49)\\
&\leqslant e^d|\mathcal{W}_1|\exp_e\left(\frac{cnt}{\min_x P_X(x)}\right)|\mathcal{W}_2|^c\cdot|\mathcal{T}_n(P_Y)|^{-c}. && (50)
\end{aligned}$$

Here,

- (44) used Jensen's inequality.
- For (46), we can clearly choose some $w^* \in \mathcal{W}_1$ such that this line holds.
- (47) used the precise form of the lower bound stated above. This is the reverse hypercontractivity step.
- In (48), we defined[4]

  $$d := \sup_{S_{X^n}}\{cD(S_{Y^n}\|P_{Y^n}) - D(S_{X^n}\|P_{X^n})\} \tag{51}$$

  where the supremum is over S_{X^n} supported on $\mathcal{T}_n(P_X)$, and $S_{X^n} \to P_{Y^n|X^n} \to S_{Y^n}$. (48) follows by taking

[4] Although not used in the proof, we remark that the largest $c > 1$ for which $d = 0$ equals the reciprocal of the strong data processing constant.

$f = \Lambda_{n,t} 1_{\mathcal{B}_{w*}}$ in the following basic functional-entropic duality result (see e.g. [24, Theorem 1]):

$$d = \sup_{f \in \mathcal{H}_+(\mathcal{Y}^n)} \left\{ \ln P_{X^n}(e^{cP_{Y^n|X^n}(\ln f)}) - c \ln P_{Y^n}(f) \right\}. \tag{52}$$

For completeness, we recall the proof of the $\geqslant$ part in (52): given any $f \in \mathcal{H}_+(\mathcal{Y}^n)$, define S_{X^n} by $\frac{\mathrm{d}S_{X^n}}{\mathrm{d}P_{X^n}} = \frac{e^{cP_{Y^n|X^n}(\ln f)}}{P_{X^n}\left(e^{cP_{Y^n|X^n}(\ln f)}\right)}$. We have

$$\begin{aligned}
&\ln P_{X^n}(e^{cP_{Y^n|X^n}(\ln f)}) - c \ln P_{Y^n}(f) \\
&= S_{X^n}(cP_{Y^n|X^n}(\ln f)) - D(S_{X^n} \| P_{X^n}) - c \ln P_{Y^n}(f) && (53)\\
&= cS_{Y^n}(\ln f) - D(S_{X^n} \| P_{X^n}) - c \ln P_{Y^n}(f) && (54)\\
&\leqslant cD(S_{Y^n} \| P_{Y^n}) - D(S_{X^n} \| P_{X^n}) && (55)\\
&\leqslant d && (56)
\end{aligned}$$

where (55) follows from the Gibbs' variation formula.

- (49) used the precise form of the upper bound stated above.
- (50) used $|\mathcal{B}_{w*}| \leqslant |\mathcal{W}_2|$.

We thus obtain

$$\begin{aligned}
&\ln |\mathcal{W}_1| + c \ln |\mathcal{W}_2| \\
&\geqslant -d + c \ln |\mathcal{T}_n(P_Y)| \\
&\quad - \inf_{t>0} \left\{ \frac{cnt}{\min_x P_X(x)} + c\left(1 + \frac{1}{t}\right) \ln \frac{1}{1-\epsilon} \right\} && (57)\\
&\geqslant -d + c \ln |\mathcal{T}_n(P_Y)| + c \ln(1-\epsilon) \\
&\quad - 2c \sqrt{\frac{n}{\min_x P_X(x)} \ln \frac{1}{1-\epsilon}}. && (58)
\end{aligned}$$

Finally, Lemma 2 in Section V-D ahead shows that $-d + c \ln |\mathcal{T}_n(P_Y)| \geqslant n\phi_c(P_{XY}) - E \ln n$, and the proof of Lemma 1 is completed.

Remark 2: From the proof we see that the result continues to hold if the $\mathcal{Y}$-encoder is allowed to access the message of the $\mathcal{X}$-encoder: $g\colon \mathcal{Y} \times \mathcal{W}_1 \to \mathcal{W}_2$.

Remark 3: We used Jensen inequality to get (46) from (43). In contrast, [1] used a reverse Markov inequality, essentially deducing from (43) that

$$P_{X^n}[x^n\colon P_{Y^n|X^n=x^n}[\mathcal{B}_{w*}] \geqslant 1-\epsilon'] \geqslant \frac{\epsilon' - \epsilon}{\epsilon' |\mathcal{W}_1|} \tag{59}$$

for any $\epsilon' \in (\epsilon, 1)$, which gives rise to a new parameter ϵ' to be optimized. It is possible to follow (59) with the functional approach as we did in [30]. However, proceeding with (46) is more natural and better manifests the simplicity and flexibility of the functional approach [30].

C. Proof Theorem 1

Let P_{XY} be an arbitrary n-type such that $|P_{XY} - Q_{XY}| \leqslant \lambda$ as in Lemma 1. Then if the error probability conditioned on type P_{XY} is less than $\delta \in (0,1)$, we have

$$\begin{aligned}
n\phi_c(Q_{XY}) + D\sqrt{n} \geqslant n\phi_c(P_{XY}) &- 2c\sqrt{\tfrac{n}{\min_x P_X(x)} \ln \tfrac{1}{1-\delta}} \\
&- E \ln n - c \ln 2. && (60)
\end{aligned}$$

Remark that the last two terms will be immaterial for the asymptotic analysis. Note that by the Taylor expansion, there exists $F > 0$ (depending on Q_{XY} and c) such that for any P_{XY} in the λ-neighborhood of Q_{XY},

$$\begin{aligned}
\phi_c(P_{XY}) \geqslant \phi_c(Q_{XY}) &+ \langle \nabla\, \phi_c|_Q \,, P_{XY} - Q_{XY} \rangle \\
&- F|P_{XY} - Q_{XY}|^2. && (61)
\end{aligned}$$

Combining the two bounds above, the error conditioned on type P_{XY} exceeds δ if

$$\begin{aligned}
&n\langle \nabla\, \phi_c|_Q \,, P_{XY} - Q_{XY} \rangle \\
&> D\sqrt{n} + 2c\sqrt{\frac{n}{\min_x P_X(x)} \ln \frac{1}{1-\delta}} \\
&\quad + E \ln n + nF|P_{XY} - Q_{XY}|^2 + c \ln 2. && (62)
\end{aligned}$$

Now particularize P_{XY} to be the empirical distribution of $(X^n, Y^n) \sim Q_{XY}^{\otimes n}$. Then with probability $1 - O(e^{-n^{1/3}})$ we have $|P_{XY} - Q_{XY}| < n^{-1/3}$ (by Hoeffding's inequality) and $\frac{1}{\min_x P_X(x)} < \frac{2}{\min_x Q_X(x)}$, and (62) holds if (for some $G > 0$)

$$\begin{aligned}
&\sum_{i=1}^n \phi_c|_Q\,(X_i, Y_i) - \mathbb{E}[\,\phi_c|_Q\,(X, Y)] \\
&> D\sqrt{n} + 2c\sqrt{\frac{2n}{\min_x Q_X(x)} \ln \frac{1}{1-\delta}} + Gn^{1/3}. && (63)
\end{aligned}$$

Thus by CLT, we conclude that the probability of type with error exceeding δ is at least

$$\mathrm{Q}\left(\frac{D + c\sqrt{\frac{8}{\min_x Q_X(x)} \ln \frac{1}{1-\delta}}}{\sqrt{\mathrm{Var}(\nabla\, \phi_c|_Q\,(X,Y))}} \right) - o(1). \tag{64}$$

Averaging over types, we obtain

$$\liminf_{n\to\infty} \mathbb{P}[\mathcal{E}_n] \geqslant \delta \mathrm{Q}\left(\frac{D + c\sqrt{\frac{8}{\min_x Q_X(x)} \ln \frac{1}{1-\delta}}}{\sqrt{\mathrm{Var}(\,\nabla \phi_c|_{Q_{XY}}\,(X,Y))}} \right). \tag{65}$$

and (11) follows by taking the infimum over δ.

To bound the c-dispersion, we will take $\delta = 1/2$ in the above. For any $\epsilon \in (0,1)$, set $D_\epsilon \in \mathbb{R}$ to be such that

$$\frac{1}{2}\mathrm{Q}\left(\frac{D_\epsilon + c\sqrt{\frac{8}{\min_x Q_X(x)} \ln 2}}{\sqrt{\mathrm{Var}(\nabla\, \phi_c|_Q\,(X,Y))}} \right) = \epsilon + \epsilon^2. \tag{66}$$

Then

$$\begin{aligned}
V_c &\geqslant \liminf_{\epsilon \downarrow 0} \limsup_{n\to\infty} \frac{[\ln M_c(n,\epsilon) - n\phi_c(Q_{XY})]^2}{2n \ln(1/\epsilon)} && (67)\\
&\geqslant \liminf_{\epsilon \downarrow 0} \limsup_{n\to\infty} \frac{D_\epsilon^2}{2\ln 2 - 2\ln \mathrm{Q}\left(\frac{D_\epsilon + c\sqrt{\frac{8}{\min_x Q_X(x)} \ln 2}}{\sqrt{\mathrm{Var}(\nabla\, \phi_c|_Q\,(X,Y))}} \right)} && (68)
\end{aligned}$$

where the last line follows since given $\epsilon \in (0,1)$, for sufficiently large n any code with $\ln|\mathcal{W}_1| + c\ln|\mathcal{W}_2| \leqslant n\phi_c(Q_{XY}) + D_\epsilon\sqrt{n}$ has error probability larger than ϵ, which implies that $\ln M_c(n,\epsilon) - n\phi_c(Q_{XY}) \geqslant \sqrt{n}D_\epsilon$. Since clearly $D_\epsilon \to +\infty$ as $\epsilon \downarrow 0$, we can further lower bound (68) as

$$\liminf_{D\to\infty}\limsup_{n\to\infty} \frac{D^2}{-2\ln \mathrm{Q}\left(\frac{D+c\sqrt{\frac{8}{\min_x Q_X(x)}\ln 2}}{\sqrt{\mathrm{Var}(\nabla\,\phi_c|_Q(X,Y))}}\right)}$$

$$\geqslant \lim_{D\to\infty} \frac{D^2}{\left(\frac{D+c\sqrt{\frac{8}{\min_x Q_X(x)}\ln\frac{1}{1-\delta}}}{\sqrt{\mathrm{Var}(\,\nabla\phi_c|_{Q_{XY}}(X,Y))}}\right)^2} \tag{69}$$

$$= \mathrm{Var}(\,\nabla\phi_c|_{Q_{XY}}(X,Y)) \tag{70}$$

where (69) uses the fact that $\lim_{r\to\infty}\frac{\ln \mathrm{Q}(r)}{r^2} = -\frac{1}{2}$. The equivalent expression for the gradient was shown in (41). This establishes the dispersion bound in Theorem 1.

D. Single-Letterization on Types

Given an n-type P_{XY}, let $P_{X^nY^n}$ be the equiprobable distribution on $\mathcal{T}_n(P_{XY})$, and let $P_{Y^n|X^n}$ be the induced random transformation defined for distributions supported on $\mathcal{T}_n(P_X)$. Let

$$\psi_{c,n}(P_{XY}) := \inf_{S_{X^n}}\{cH(S_{Y^n}) + D(S_{X^n}\|P_{X^n})\}. \tag{71}$$

Here, the infimum is over S_{X^n} supported on $\mathcal{T}_n(P_X)$, and we have set S_{Y^n} by $S_{X^n}\to P_{Y^n|X^n}\to S_{Y^n}$. We will show that $\psi_{c,n}(P_{XY}) \geqslant n\phi_c(P_{XY}) - O(\ln n)$, which is useful for the converse proof,

For pedagogical purposes, let us first imagine that in the definition of $\psi_{c,n}$ above, $P_{X^nY^n}$ is replaced by some i.i.d. distribution $Q_{XY}^{\otimes n}$, and the constraint that S_{X^n} is supported on $\mathcal{T}_n(P_X)$ is dropped. Then we simply have the tensorization property $\psi_{c,n}(Q_{XY}) = n\psi_{c,1}(Q_{XY})$, which is also known as the single-letterization process in almost all information-theoretic converses (see for the case of WAK [49]). The proof of tensorization relies on the chain rule of the relative entropy and the fact that $Q_{X_{\setminus 1}Y_{\setminus 1}|X_1=x,Y_1=y} = Q_{XY}^{\otimes(n-1)}$, for all x and y, in the i.i.d. case.

Now for $P_{X^nY^n}$ which is the equiprobable distribution on $\mathcal{T}_n(P_{XY})$, the distribution $P_{X_{\setminus 1}Y_{\setminus 1}|X_1=x,Y_1=y}$ depends on x and y, which is why we cannot directly run the same single-letterization argument directly. However we will show that $\psi_{c,n}(P_{XY})$ is still approximated by $n\phi_c(P_{XY})$ up to a logarithmic error. To get a sense of the logarithmic error, consider the toy problem of approximating the entropy of the equiprobable distribution P_{X^n} on a type P_X:

$$nH(P_X) - |\mathcal{X}|\ln(n+1) \leqslant H(P_{X^n}) \leqslant nH(P_X). \tag{72}$$

Of course, for this toy problem it suffices to count the cardinality of $\mathcal{T}_n(P_X)$ to obtain (72), as was done in [10, Lemma 2.3]. Unfortunately, the counting argument does not seem to be easily extendable, so we try the following single-letterization method instead: Let $X^n \sim P_{X^n}$. For each $k = 1,\dots,n-1$, define the functions

$$\Delta_k(x_1,\dots,x_k) := P_{X_{k+1}|X^k=x^k} - P_{X_k|X^{k-1}=x^{k-1}}, \tag{73}$$

and the random variable (viewed as vector in $\mathbb{R}^{|\mathcal{X}|}$) $\Delta_k := \Delta_k(X_1,\dots,X_k)$. Then

$$H(X^n) = H(X_1) + H(X_2|X_1) + \dots + H(X_n|X^{n-1}) \tag{74}$$

$$= \sum_{k=0}^{n-1}\mathbb{E}[H(P_X + \Delta_1 + \dots + \Delta_k)]. \tag{75}$$

Now if P_X is fully supported on $\mathcal{X}$, we can choose $E > 0$ depending only on P_X large enough such that

$$H(S_X) \geqslant H(P_X) - \left\langle \ln\frac{1}{P_X},\, S_X - P_X\right\rangle - E\|S_X - P_X\|^2. \tag{76}$$

Using (76) to lower bound (75), and noting that $\Delta_1,\dots,\Delta_{n-1}$ is a martingale sequence, we have

$$H(X^n) \geqslant nH(P_X) - E\sum_{k=0}^{n-1}\mathbb{E}\|\Delta_1 + \dots, \Delta_k\|^2 \tag{77}$$

$$= nH(P_X) - E\sum_{k=0}^{n-1}(n-k)\mathbb{E}\|\Delta_k\|^2 \tag{78}$$

$$\geqslant nH(P_X) - E\sum_{k=1}^{n-1}(n-k)\cdot\frac{4}{(n-k)^2} \tag{79}$$

$$\geqslant nH(P_X) - 4E(1+\ln(n-1)) \tag{80}$$

where follows from the fact that $\|\Delta_k\| \leqslant |\Delta_k| \leqslant \frac{2}{n-k}$ with probability 1. Compare the logarithmic gaps in the left side of (72) and in (80). We now extend the above analysis to bound (71):

Lemma 2: Given Q_{XY} and $c \geqslant 1$, there exists $\lambda \in (0,1)$ and $E > 0$ such that for any $n \geqslant 1$ and n-type P_{XY} such that $|P_{XY} - Q_{XY}| < \lambda$, we have

$$\psi_{c,n}(P_{XY}) \geqslant n\phi_c(P_{XY}) - E\ln n. \tag{81}$$

Proof: Under the assumption that ϕ_c has bounded second derivatives in a neighborhood of Q_{XY}, there exists $\lambda \in (0,1)$ and $E' > 0$ large enough such that

$$\phi_c(S_{XY}) \geqslant \phi_c(P_{XY}) + \langle\nabla\,\phi_c|_{P_{XY}}, S_{XY} - P_{XY}\rangle$$
$$- E'\|S_{XY} - P_{XY}\|^2 \tag{82}$$

for any $P_{XY}\colon |P_{XY} - Q_{XY}| \leqslant \lambda$ and any S_{XY} in the probability simplex (the Taylor expansion proves (82) for S_{XY} in a neighborhood of P_{XY}. Then using the boundedness of $\phi_c(\cdot)$ we can extend (82) to all S_{XY} by choosing E' large enough). Here $\|\cdot\|$ denotes the ℓ_2 norm, although any norm admitting an inner product would work. Consider any S_{X^n} supported on $\mathcal{T}_n(P_X)$, and put $S_{X^nY^n} = S_{X^n}P_{Y^n|X^n}$. Let I be equiprobable on $\{1,\dots,n\}$ and independent of (X^n,Y^n)

under S. Let $X_{\backslash I}$ denote the coordinates excluding the I-th one. We have

$$H(S_{Y^n}) = H(S_{Y_I|I}|S_I) + H(S_{Y_{\backslash I}|IY_I}|S_{IY_I}) \tag{83}$$

$$\geqslant H(S_{Y_I|I}|S_I) + H(S_{Y_{\backslash I}|IX_IY_I}|S_{IX_IY_I}), \tag{84}$$

$$\begin{aligned} &D(S_{X^n}\|P_{X^n}) \\ &= D(S_{X_I|I}\|P_{X_I}|S_I) + D(S_{X_{\backslash I}|IX_I}\|P_{X_{\backslash I}|X_I}|S_{IX_I}) \quad (85)\\ &= D(S_{X_I|I}\|P_{X_I}|S_I)+D(S_{X_{\backslash I}|IX_IY_I}\|P_{X_{\backslash I}|X_IY_I}|S_{IX_IY_I}) \quad (86) \end{aligned}$$

where (86) follows from $X_{\backslash I} - IX_I - Y_I$ under S (this is because for any (i, x^n), we have $S_{Y_I|(IX_I)=(i,x_i)} = S_{Y_I|(I,X^n)=(i,x^n)} = S_{Y|X=x_i}$). Noting that $P_{X_I} = P_X$ and $S_{IX_IY_I} = S_{I|X_I}P_{XY}$, we can bound the weighted sum of the first terms in the above two expansions:

$$cH(S_{Y_I|I}|S_I) + D(S_{X_I|I}\|P_{X_I}|S_I) \geqslant \phi_c(P_{XY}). \tag{87}$$

To bound the weighted sum of the second terms in (84) and (86), for any (x, y), define

$$P_{XY}^{xy}(x', y') := \tfrac{1}{n-1}[nP_{XY}(x', y')-1_{(x',y')=(x,y)}], \quad \forall(x', y'). \tag{88}$$

That is, P_{XY}^{xy} denotes the $(n-1)$-type obtained by removing one pair (x, y) from sequences of the type n-type P_{XY}. Then

$$\begin{aligned} &cH(S_{Y_{\backslash I}|IX_IY_I}|S_{IX_IY_I})+D(S_{X_{\backslash I}|IX_IY_I}\|P_{X_{\backslash I}|X_IY_I}|S_{IX_IY_I}) \\ &\geqslant \sum_{x,y,i}\psi_{c,n-1}(P_{XY}^{xy})S_{IX_IY_I}(i, x, y) \quad (89)\\ &= \sum_{x,y}\psi_{c,n-1}(P_{XY}^{xy})P_{XY}(x, y). \quad (90) \end{aligned}$$

Summarizing and iterating,

$$\begin{aligned} \psi_{c,n}(P_{XY}) &\geqslant \phi_c(P_{XY}) + \mathbb{E}[\psi_{c,n-1}(P_{XY} + \Delta_1)] \quad (91)\\ &\geqslant \dots \quad (92)\\ &\geqslant \sum_{k=0}^{n-1}\mathbb{E}[\phi_c(P_{XY} + \Delta_1 + \dots + \Delta_k)] \quad (93) \end{aligned}$$

where we defined the sequence $\Delta_1, \Delta_2, \dots$ of random vectors in the following way: conditioned on $\Delta_1, \dots, \Delta_k$, denote the probability vector $S_k := P_{XY} + \sum_{i=1}^{k}\Delta_i$, and then $\Delta_{k+1} := S_k^{xy} - S_k$ with probability $S_k(x, y)$ for each (x, y). Thus $\Delta_1 + \dots + \Delta_k$ is a zero mean martingale, and

$$\begin{aligned} \psi_{c,n}(P_{XY}) &\geqslant n\phi_c(P_{XY}) - E'\sum_{k=1}^{n-1}\mathbb{E}\|\Delta_1 + \dots + \Delta_k\|^2 \quad (94)\\ &\geqslant n\phi_c(P_{XY}) - E'\sum_{k=1}^{n-1}(n-k)\mathbb{E}\|\Delta_k\|^2 \quad (95)\\ &\geqslant n\phi_c(P_{XY}) - E'\sum_{k=1}^{n-1}(n-k)\cdot\frac{4}{(n-k)^2} \quad (96)\\ &\geqslant n\phi_c(P_{XY}) - 4E'(1 + \ln(n-1)) \quad (97) \end{aligned}$$

where (94) uses (82), noting that the the linear term (second term on the right side of (82)) vanishes upon taking the expectation; (95) follows since the martingale property implies that $\mathbb{E}[\Delta_i|\Delta_{i'}] = 0$ for $i > i'$; (96) follows from the fact that $\|\Delta_k\| \leqslant |\Delta_k| \leqslant \frac{2}{n-k}$ with probability 1. Taking $E = 10E'$ completes the proof. □

Remark 4: The first-order term in (81) is tight, since otherwise we would get a converse result contradicting the achievability. The first-order tightness can also be argued directly: Suppose that $P_{UXY} = P_{U|X}P_{XY}$ where $P_{U|X}$ is a maximizer in the definition of ϕ_c. Assume without loss of generality that $\mathcal{U} := \{1, \dots, |\mathcal{U}|\}$. Split $\{1, \dots, n\}$ into blocks of sizes $n_1, \dots, n_{|\mathcal{U}|}$ where $n_u := nP_U(u)$ is assumed to be an integer (if not, perturb $P_{U|X}$ by $O(1/n)$ so that it is, and the proof will still work with a continuity argument). Denote $x^{(u)}$ and $y^{(u)}$ to be the sub-vectors corresponding to the u-blocks. Let S_{X^n} in the definition of $\psi_{c,n}(P_{XY})$ to be equiprobable on $\{x^n : x^{(u)} \in \mathcal{T}_{n_u}(P_{X|U=u}), \forall u\}$ where we assumed that $P_{X|U=u}$ is n_u-type (if not, perturb $P_{U|X}$ and P_X by $O(1/n)$, and continue the proof by an approximation argument). Computing the size of the type classes leads to

$$D(S_{X^n}\|P_{X^n}) = nI(U; X) + o(n). \tag{98}$$

Moreover, since it is easy to see that S_{X^n} and hence S_{Y^n}, is invariant under permutations within the same u-block, we have

$$\begin{aligned} &H(S_{Y^n}) \\ &= H_S(\hat{P}_{Y^{(1)}}, \dots, \hat{P}_{Y^{(|\mathcal{U}|)}}) \\ &\quad + \sum_{P^{(1)},\dots,P^{(|\mathcal{U}|)}} H_S(Y^n|P^{(1)}, \dots, P^{(|\mathcal{U}|)})\mathbb{P}_S(P^{(1)}, \dots, P^{(|\mathcal{U}|)}) \quad (99)\\ &= o(n)+ \\ &\quad \sum_{P^{(1)},\dots,P^{(|\mathcal{U}|)}} \ln\left|\left\{y^n : \hat{P}_{Y^{(1)}} = P^{(1)}, \dots, \hat{P}_{Y^{(|\mathcal{U}|)}} = P^{(|\mathcal{U}|)}\right\}\right| \\ &\quad \cdot \mathbb{P}_S(P^{(1)}, \dots, P^{(|\mathcal{U}|)}) \quad (100) \end{aligned}$$

Here, $H_S(\hat{P}_{Y^{(1)}}, \dots, \hat{P}_{Y^{(|\mathcal{U}|)}})$ denotes the entropy of the random variables $\hat{P}_{Y^{(1)}}, \dots, \hat{P}_{Y^{(|\mathcal{U}|)}}$ which are the empirical distributions of $Y^{(1)}$, ..., $Y^{(|\mathcal{U}|)}$ where $Y^n \sim S_{Y^n}$. The summation is over n_u-types $P^{(u)}$, and $\mathbb{P}_S(P^{(1)}, \dots, P^{(|\mathcal{U}|)})$ denotes the probability that $\hat{P}_{Y^{(u)}} = P^{(u)}$, for each u. Let (α_n) be a sequence in $\omega(\frac{1}{\sqrt{n}}) \cap o(1)$. We can verify that

$$\begin{aligned} &\mathbb{P}_S[|\hat{P}_{Y^{(u)}} - P_{Y|U=u}| \leqslant \alpha_n, \forall u] \\ &\geqslant \min_{x^n\in\mathrm{supp}(S_{X^n})} \mathbb{P}_S[|\hat{P}_{Y^{(u)}} - P_{Y|U=u}| \leqslant \alpha_n, \forall u|X^n = x^n] \quad (101)\\ &\geqslant 1 - o(1) \quad (102) \end{aligned}$$

by computing the number of configurations and standard approximations of combinatorial numbers. Since the logarithm in (100) is $O(n)$, we have

$$\begin{aligned} & H(S_{Y^n}) \\ &= o(n)+ \\ &\sum_{|\hat{P}_{Y(u)}-P_{Y|U=u}|\leqslant\alpha_n,\forall u} \ln\left|\left\{y^n : \hat{P}_{Y^{(1)}} = P^{(1)}, \dots, \hat{P}_{Y^{(|\mathcal{U}|)}} = P^{(|\mathcal{U}|)}\right\}\right| \\ &\cdot \mathbb{P}_S(P^{(1)}, \dots, P^{(|\mathcal{U}|)}) \qquad (103) \\ &= o(n)+ \\ &\sum_{|\hat{P}_{Y^{(u)}}-P_{Y|U=u}|\leqslant\alpha_n,\forall u} \ln\left|\left\{y^n : \hat{P}_{Y^{(u)}} = P_{Y|U=u}, \forall u\right\}\right| \\ &\cdot \mathbb{P}_S(P^{(1)}, \dots, P^{(|\mathcal{U}|)}) \qquad (104) \\ &= o(n) + nH(Y|U)\cdot(1-o(1)) \qquad (105) \\ &= o(n) + nH(Y|U) \qquad (106) \end{aligned}$$

where (104) is by continuity. This together with (98) shows the first-order tightness.

VI. Proof of the Achievability

In this section we prove Theorem 2. We first make a few preliminary observations about the optimization problem in the definition of ϕ_c:

Proposition 4: Let $c > 0$, and Q_{XY} be fully supported on the finite set $\mathcal{X}\times\mathcal{Y}$ (that is, $P_{XY}(x,y) > 0$ for each (x,y)). Let $P^\star_{U|X}$ be an infimizer for (1). Assume without loss of generality that $P^\star_U$ is fully supported on some finite set $\mathcal{U}$. Then

1) $P^\star_{U|X=x}$ is also fully supported on $\mathcal{U}$ for each x.
2) As long as $I(U;X) > 0$, $(U,X) \sim P^\star_{U|X}Q_X$, we have

$$\nabla_{P_{U|X}} I(U;X)\Big|_{P^\star_{U|X}} \neq 0. \qquad (107)$$

Proof: The proofs follow from the first-order optimality condition. For the first claim, note that if P_U is fully supported and $P_{U|X}(u|x) = 0$, we have $\frac{\partial}{\partial P_{U|X}(u|x)} I(U;X)\Big|_{P^\star_{U|X}} = -\infty$, whereas $\frac{\partial}{\partial P_{U|X}(u|x)} H(Y|U)\Big|_{P^\star_{U|X}}$ is finite since Q_{XY} (and hence Q_{UY}) is fully supported. This contradicts the optimality of $P^\star_{U|X}$. For the second claim, observe that $I(U;X) > 0$ under $P^\star_{U|X}Q_X$ implies the existence of some x such that $(\imath_{U;X}(u;x))_{u\in\mathcal{U}}$, which equals to $\nabla_{P_{U|X=x}} I(U;X)\Big|_{P^\star_{U|X}}$ up to an additive constant, cannot be a vector with constant coordinates. □

We next define the encoders and the decoder for each type of X^n, and perform the error analysis. A naive strategy, assuming the existence of an oracle that reveals the joint type $\hat{P}_{X^nY^n}$ to both the encoders and the decoder, is to design codes for each joint type. More precisely, let $(\hat{U},\hat{X},\hat{Y}) \sim P^\star_{U|X}\hat{P}_{X^nY^n}$, where we fix $P^\star_{U|X}$ to be an infimizer for (1), and suppose that $P_{\hat{X}\hat{Y}}$ is revealed to the encoders and the decoder. For any given $P_{\hat{U}\hat{X}}$, the codebook for the first encoder are generated i.i.d. and with i.i.d. coordinates distributed according to $P_{\hat{U}}$. By the type covering lemma, codebook of log-size $l_1(P_{\hat{X}}) = nI(\hat{U};\hat{X}) + o(\sqrt{n})$ is large enough so that a codeword with the joint type $P_{\hat{U}\hat{X}}$ can be selected with high (faster than any polynomial) probability. The second encoder may perform random binning and send $l_2(P_{\hat{X}\hat{Y}}) = nH(\hat{Y}|\hat{U}) + o(\sqrt{n})$ bits, so that the decoder can perform minimum conditional entropy decoding (see e.g. [9]) to recover Y^n with high probability. By performing the Taylor expansions on $I(\hat{U};\hat{X})$ and $H(\hat{Y}|\hat{U})$ and using the central limit style analysis, we can verify that the probability of $l_1(P_{\hat{X}}) + c\,l_2(P_{\hat{X}\hat{Y}}) > n\phi_c(Q_{XY}) + D\sqrt{n}$ indeed converges to $Q\left(\frac{D}{\sqrt{V}}\right)$. However, the challenge with this naive strategy is two-folded: first, of course, the joint type $P_{\hat{X}\hat{Y}}$ is not actually known to all parties; second, l_1 and l_2 individually vary with $P_{\hat{X}\hat{Y}}$, even though we may impose that $l_1 + c\,l_2$ is fixed.

A remedy is to perturb the "codeword distribution" $P_{U|X}$ according to $\hat{P}_{X^n}$, so that the compression lengths are at some fixed budgets $\bar{l}_1$ and $\bar{l}_2$ which do not vary with the type. We will see that (107) guarantees that we can find such a perturbation to bring $l_1(P_{\hat{X}})$ and $l_2(P_{\hat{X}\hat{Y}})$ to $\bar{l}_1$ and $\bar{l}_2$. The first-order optimality condition

$$\nabla_{P_{U|X}} I(U;X)\Big|_{P^\star_{U|X}} + c\cdot\nabla_{P_{U|X}} H(Y|U)\Big|_{P^\star_{U|X}} = 0, \qquad (108)$$

where $(U,X,Y) \sim P_{U|X}Q_{XY}$, ensures that we can always keep the balance $(l_1(P_{\hat{X}}) - \bar{l}_1) + c(l_2(P_{\hat{X}\hat{Y}}) - \bar{l}_2)$. We now describe the scheme in detail. Fix an arbitrary $\kappa \in (0,1/6)$.

Encoder 1: Define M_1 by

$$\log M_1 = nI(U^\star;X^\star) + n^{3\kappa} \qquad (109)$$

where

$$(U^\star,X^\star,Y^\star) \sim P^\star_{U|X}Q_{XY}. \qquad (110)$$

Encoder 1 constructs a codebook consisting of M_1 codewords i.i.d. according to $P_U^{\star\otimes n}$. Upon observing X^n, Encoder 1 will send the index of $\hat{P}_{X^n}$ (using $O(\log n)$ bits) and then send the index of the codeword. The codeword is selected in the following way: For each $\hat{P}_{X^n}$ satisfying

$$|\hat{P}_{X^n} - Q_X| \leqslant n^{-1/2+\kappa}, \qquad (111)$$

we define below a $P_{U|X}$ which is a perturbation of $P^\star_{U|X}$ (the purpose being that $I(\hat{P}_{X^n}, P_{U|X}) = I(U^\star;X^\star) + o(|\hat{P}_{X^n} - Q_X|)$ so that encoding is successful under the fixed budget (109)). Then upon observing X^n, the encoder selects the first codeword U^n such that (U^n,X^n) has type $P_{U|X}\hat{P}_{X^n}$ (if any).

To define such a $P_{U|X}$ associated with each $\hat{P}_{X^n}$, we can first pick a fixed $P'_{U|X}$ such that

$$\frac{\partial}{\partial t} I(Q_X,\, tP'_{U|X} + (1-t)P^\star_{U|X})\Big|_{t=0} \neq 0. \qquad (112)$$

This is possible in view of Proposition 4. Then take

$$P_{U|X} = tP'_{U|X} + (1-t)P^\star_{U|X}, \qquad (113)$$

rounded such that $P_{U|X}\hat{P}_{X^n}$ is n-type, where

$$t := -\frac{\left\langle \imath_{U^\star;X^\star},\, P^\star_{U|X}(\hat{P}_{X^n} - Q_X)\right\rangle}{\frac{\partial}{\partial t} I(Q_X,\, tP'_{U|X} + (1-t)P^\star_{U|X})\Big|_{t=0}}. \qquad (114)$$

Encoder 2: Each y^n sequence is mapped randomly to one of M_2 bins, where we defined M_2 by

$$\log M_2 = nH(Y^\star|U^\star) + \frac{D}{c}\sqrt{n} + n^{3\kappa}, \tag{115}$$

and the bin index is sent.

Decoding rule: The decoder receives $\widehat{P}_{X^n}$, and hence knows the previously agreed $P_{U|X}$ as long as (111) holds. The decoder selects the y^n sequence in the m_2-th bin that minimizes the empirical conditional entropy $H(\widehat{P}_{y^n|u^n}|\widehat{P}_{u^n})$ (see [9] for general reference on the minimum entropy decoder).

Error analysis:

- Let $\mathcal{E}_0$ be the event that (111) fails. We note that $\mathcal{E}_0^c$ implies that

$$|P_{U|X}\widehat{P}_{X^n} - P^\star_{U|X}Q_X| \leqslant n^{-1/2+\kappa}. \tag{116}$$

Indeed, by the definition (114) we see that[5] $t = O(n^{-1/2+\kappa})$, which in turn implies that $\|P_{U|X} - P^\star_{U|X}\| = O(n^{-1/2+\kappa})$ (the choice of the norm is immaterial here since finite alphabets are assumed), and hence (116) holds. By the Chebyshev inequality, we have the bound

$$\mathbb{P}[\mathcal{E}_0] = e^{-\Omega(n^{2\kappa})}. \tag{117}$$

- Let $\mathcal{E}_1 \subseteq \mathcal{E}_0^c$ be the event that no codeword is selected by Encoder 1. For each $\widehat{P}_{X^n} \in \mathcal{E}_0^c$, we use the Taylor expansion to show that

$$\begin{aligned} &D(P_{U|X}\|P^\star_U|\widehat{P}_{X^n}) \\ &= D(P_{U|X}\widehat{P}_{X^n}\|P^\star_U \times Q_X) - D(\widehat{P}_{X^n}\|Q_X) && (118)\\ &= D(P_{U|X}\widehat{P}_{X^n}\|P^\star_U \times Q_X) + O(n^{-1+2\kappa}) && (119)\\ &= I(U^\star;X^\star) + \left\langle \imath_{U^\star;X^\star},\, P_{U|X}\widehat{P}_{X^n} - P^\star_{U|X}Q_X \right\rangle \\ &\quad + O(n^{-1+2\kappa}) && (120)\\ &= I(U^\star;X^\star) + \left\langle \imath_{U^\star;X^\star},\, (P_{U|X} - P^\star_{U|X})Q_X \right\rangle \\ &\quad + \left\langle \imath_{U^\star;X^\star},\, P^\star_{U|X}(\widehat{P}_{X^n} - Q_X) \right\rangle + O(n^{-1+2\kappa}) && (121)\\ &= I(U^\star;X^\star) + O(n^{-1+2\kappa}) && (122)\end{aligned}$$

where
  - (119) uses (111).
  - (121), which is essentially Leibniz's rule for the first derivative, uses the fact that $|\widehat{P}_{X^n} - Q_X| \cdot \|P_{U|X} - P^\star_{U|X}\| = O(n^{-1+2\kappa})$.
  - (122) follows since the choice of $P_{U|X}$ in (113) ensures that the second and the third terms in (121) cancel.

Next, recall that each codeword is generated according to $P_U^{\star\otimes n}$. Conditioned on any $\widehat{P}_{X^n}$, a codeword and X^n has the joint type $P_{U|X}\widehat{P}_{X^n}$ with probability $e^{-nD(P_{U|X}\|P^\star_U|\widehat{P}_{X^n})-O(\log n)}$. Hence for any $\widehat{P}_{X^n}$ satisfying $\mathcal{E}_0^c$, using (122) and (109) we have

$$\begin{aligned} \mathbb{P}[\mathcal{E}_1|\widehat{P}_{X^n}] &\leqslant [1 - e^{-nD(P_{U|X}\|P^\star_U|\widehat{P}_{X^n})-O(\log n)}]^{M_1} && (123)\\ &\leqslant [1 - e^{-nI(U^\star;X^\star)-O(n^{2\kappa})}]^{M_1} && (124)\\ &\leqslant O(\exp_e(-e^{\kappa n})). && (125)\end{aligned}$$

The above steps are essentially the type covering argument; see e.g. [10, Section 2, Section 9]. Consequently,

$$\mathbb{P}[\mathcal{E}_1] \leqslant O(\exp_e(-e^{\kappa n})). \tag{126}$$

- Let $\mathcal{E}_2 \subseteq (\mathcal{E}_0 \cup \mathcal{E}_1)^c$ be the event that there exists some $y'^n \neq Y^n$ such that the conditional entropy for its empirical conditional distribution is smaller:

$$H(\widehat{P}_{y'^n|U^n}|\widehat{P}_{U^n}) < H(\widehat{P}_{Y^n|U^n}|\widehat{P}_{U^n}), \tag{127}$$

where U^n denotes the codeword selected by Encoder 1, and y'^n and Y^n are assigned to the same bin by Encoder 2. Since there are at most $e^{nH(\widehat{P}_{Y^n|U^n}|\widehat{P}_{U^n})}$ such y'^n sequences, and since each sequence is mapped to the same bin as Y^n with probability $1/M_2$, by the union bound we have

$$\mathbb{P}[\mathcal{E}_2|\mathcal{E}_0^c \cap \mathcal{E}_1^c] \leqslant \mathbb{E}\left[1 \wedge \frac{e^{nH(\widehat{P}_{Y^n|U^n}|\widehat{P}_{U^n})}}{M_2}\middle| \mathcal{E}_0^c \cap \mathcal{E}_1^c\right]. \tag{128}$$

We now bound the conditional entropy in (128). Let T_{Y^n} be the equiprobable distribution on $\mathcal{Y}^n$. Note that

$$\begin{aligned} &H(\widehat{P}_{Y^n|U^n}|\widehat{P}_{U^n}) \\ &= -D(\widehat{P}_{Y^nU^n}\|T_{Y^n} \times \widehat{P}_{U^n}) + n\log|\mathcal{Y}| && (129)\end{aligned}$$

is concave in $\widehat{P}_{Y^nU^n}$ (since the relative entropy is jointly convex), therefore,

$$\begin{aligned} &nH(\widehat{P}_{Y^n|U^n}|\widehat{P}_{U^n}) \\ &\leqslant nH(Y^\star|U^\star) + n\left\langle \imath_{Y^\star|U^\star},\, \widehat{P}_{U^nY^n} - P^\star_{UY}\right\rangle. && (130)\\ &= \sum_{i=1}^n \imath_{Y^\star|U^\star}(Y_i|U_i). && (131)\end{aligned}$$

Continuing (128) and using (115) and (131), we have

$$\begin{aligned} \mathbb{P}[\mathcal{E}_2|\mathcal{E}_0^c \cap \mathcal{E}_1^c] &\leqslant \mathbb{E}\left[1 \wedge e^{S_n - \frac{D}{c}\sqrt{n} - n^{3\kappa}}\middle| \mathcal{E}_0^c \cap \mathcal{E}_1^c\right] && (132)\\ &\leqslant e^{-n^{3\kappa}} + \mathbb{P}\left[S_n - \frac{D}{c}\sqrt{n} > 0 \middle| \mathcal{E}_0^c \cap \mathcal{E}_1^c\right] && (133)\end{aligned}$$

where we defined the random variable $S_n := \sum_{i=1}^n \imath_{Y^\star|U^\star}(Y_i|U_i) - nH(Y^\star|U^\star)$. Conditioned on each $\widehat{P}_{X^n} \in \mathcal{E}_0^c$ and under $\mathcal{E}_1^c$, notice that (U^n, X^n) has fixed empirical distribution $P_{U|X}\widehat{P}_{X^n}$ (determined by $\widehat{P}_{X^n}$), and hence the Y_i's for which $\{i\colon U_i = u\}$ are i.i.d. according to the conditional law in $P_{U|X}\widehat{P}_{X^n}Q_{Y|X}$ for each u, and therefore, conditioned on such a $\widehat{P}_{X^n}$

[5]In this section, the implicit constants in the $O(\cdot)$ and $\Omega(\cdot)$ notations depend on Q_{XY}, c and $P^\star_{U|X}$. That is, the implicit constants can be uniformly bounded in $\widehat{P}_{X^n}$ because we consider $\widehat{P}_{X^n}$ close to Q_X (111).

and $\mathcal{E}_1^c$, the distribution of $\frac{1}{\sqrt{n}}\left(S_n - \mathbb{E}[S_n|\widehat{P}_{X^n}, \mathcal{E}_1^c]\right)$ converges to that of a Gaussian distribution. We will apply the central limit theorem carefully later; for now let us compute the mean and variance of the conditional distribution. For each $\widehat{P}_{X^n} \in \mathcal{E}_0^c$ we have

$$
\begin{aligned}
&\mathbb{E}\left[S_n \middle| \widehat{P}_{X^n}, \mathcal{E}_1^c\right] \\
&= n\left\langle \imath_{Y^\star|U^\star},\, P_{U|X}\widehat{P}_{X^n}Q_{Y|X}\right\rangle - nH(Y^\star|U^\star) \quad (134)\\
&= n\left\langle \imath_{Y^\star|U^\star} + \frac{1}{c}\imath_{U^\star;X^\star},\, P_{U|X}\widehat{P}_{X^n}Q_{Y|X}\right\rangle - nH(Y^\star|U^\star) \\
&\quad - \frac{n}{c}I(U^\star;X^\star) - \frac{n}{c}\left\langle \imath_{U^\star;X^\star},\, P_{U|X}\widehat{P}_{X^n}\right\rangle + \frac{n}{c}I(U^\star;X^\star) \quad (135)\\
&= n\left\langle \imath_{Y^\star|U^\star} + \frac{1}{c}\imath_{U^\star;X^\star},\, P_{U|X}\widehat{P}_{X^n}Q_{Y|X} - P^\star_{U|X}Q_{XY}\right\rangle \\
&\quad + O(n^{2\kappa}) \quad (136)\\
&= n\left\langle \imath_{Y^\star|U^\star} + \frac{1}{c}\imath_{U^\star;X^\star},\, P^\star_{U|X}(\widehat{P}_{X^n} - Q_X)Q_{Y|X}\right\rangle \\
&\quad + O(n^{2\kappa}) \quad (137)\\
&= n\left\langle \mathbb{E}_{P^\star_{U|X}Q_{Y|X}}\left[\imath_{Y^\star|U^\star} + \frac{1}{c}\imath_{U^\star;X^\star}\middle| X\right],\, \widehat{P}_{X^n} - Q_X\right\rangle \\
&\quad + O(n^{2\kappa}). \quad (138)
\end{aligned}
$$

The steps above are explained as follows:

- (136) follows from bounding the last two terms in (135):

$$
\begin{aligned}
&\left\langle \imath_{U^\star;X^\star},\, P_{U|X}\widehat{P}_{X^n}Q_{Y|X}\right\rangle - I(U^\star;X^\star) \\
&= \left\langle \imath_{U^\star;X^\star},\, (P_{U|X} - P^\star_{U|X})Q_X\right\rangle \\
&\quad + \left\langle \imath_{U^\star;X^\star},\, P^\star_{U|X}(\widehat{P}_{X^n} - Q_X)\right\rangle + O(n^{-1+2\kappa}) \quad (139)\\
&= O(n^{-1+2\kappa}), \quad (140)
\end{aligned}
$$

  where (139) used the fact that $|\widehat{P}_{X^n} - Q_X|\cdot\|P_{U|X} - P^\star_{U|X}\| = O(n^{-1+2\kappa})$, and (140) follows from the fact that the definition of $P_{U|X}$ in (113) ensures that the first two terms in (139) cancel.
- To see (137), we note that

$$
\begin{aligned}
&\left\langle \imath_{Y^\star|U^\star} + \frac{1}{c}\imath_{U^\star;X^\star}, (P_{U|X} - P^\star_{U|X})(\widehat{P}_{X^n} - Q_X)Q_{Y|X}\right\rangle \\
&= O\left(\|P_{U|X} - P^\star_{U|X}\|\cdot|\widehat{P}_{X^n} - Q_X|\right) = O(n^{-1+2\kappa}) \quad (141)
\end{aligned}
$$

  and moreover, the first order optimality of $P^\star_{U|X}$ implies

$$
\begin{aligned}
&\left\langle \imath_{Y^\star|U^\star} + \frac{1}{c}\imath_{U^\star;X^\star},\, (P_{U|X} - P^\star_{U|X})Q_{XY}\right\rangle \\
&= O(\|P_{U|X} - P^\star_{U|X}\|^2) = O(n^{-1+2\kappa}). \quad (142)
\end{aligned}
$$

  These two inequalities show the step from (136) to (137) upon rearrangements.

For each $\widehat{P}_{X^n} \in \mathcal{E}_0^c$, the conditional variance is bounded as

$$
\begin{aligned}
&\frac{1}{n}\mathrm{Var}\left(S_n \middle| \widehat{P}_{X^n}, \mathcal{E}_1^c\right) \\
&= \mathbb{E}_{\widehat{P}_{X^n}P_{U|X}}\left[\mathrm{Var}_{Q_{Y|X}}\left(\imath_{Y^\star|U^\star}|UX\right)\right] \quad (143)\\
&= \mathbb{E}_{\widehat{P}_{X^n}P_{U|X}}\left[\mathrm{Var}_{Q_{Y|X}}\left(\imath_{Y^\star|U^\star} + \frac{1}{c}\imath_{U^\star;X^\star}\middle| UX\right)\right] \quad (144)\\
&= \mathbb{E}_{Q_X P^\star_{U|X}}\left[\mathrm{Var}_{Q_{Y|X}}\left(\imath_{Y^\star|U^\star} + \frac{1}{c}\imath_{U^\star;X^\star}\middle| UX\right)\right] \\
&\quad + O(n^{-1/2+\kappa}), \quad (145)
\end{aligned}
$$

where (143) follows since the distribution of S depends only on the empirical distribution (U^n, X^n), which is $P_{U|X}\widehat{P}_{X^n}$; (145) used (116) and the boundedness of $\imath_{Y^\star|U^\star} + \frac{1}{c}\imath_{U^\star;X^\star}$.

Finally, as alluded, for each type $\widehat{P}_{X^n} \in \mathcal{E}_0^c$ and under the event $\mathcal{E}_1^c$, we have that $\sum_{i:\, U_i = u} S_i$ is a sum of i.i.d. random variables for each u. We can therefore invoke the Berry-Esseén central limit theorem (see e.g. [12, Ch. XVI.5 Theorem 2]) to see that

$$
\begin{aligned}
&\mathbb{P}\left[\frac{1}{\sqrt{n}}S_n - \mathbb{E}\left[\frac{1}{\sqrt{n}}S_n\middle| \widehat{P}_{X^n}, \mathcal{E}_1^c\right] > \lambda \middle| \widehat{P}_{X^n}, \mathcal{E}_1^c\right] \\
&\leqslant \mathrm{Q}\left(\frac{\lambda}{\sqrt{\frac{1}{n}\mathrm{Var}\left(S_n \middle| \widehat{P}_{X^n}, \mathcal{E}_1^c\right)}}\right) + \xi_n, \quad \forall\lambda\in\mathbb{R} \quad (146)
\end{aligned}
$$

where ξ_n is some $o(1)$ sequence depending only on Q_{XY}, $P^\star_{U|X}$, and c (in particular, not depending on λ). Setting $\lambda = \frac{D}{c} - \mathbb{E}\left[\frac{1}{\sqrt{n}}S_n\middle| \widehat{P}_{X^n}, \mathcal{E}_1^c\right]$, we obtain from (146) and (133) that

$$
\begin{aligned}
&\mathbb{P}\left[\mathcal{E}_2|\mathcal{E}_0^c\cap\mathcal{E}_1^c\right] \leqslant e^{-n^{3\kappa}} + \\
&\mathbb{E}\left[\mathrm{Q}\left(\frac{\frac{D}{c} - \mathbb{E}\left[\frac{1}{\sqrt{n}}S_n\middle| \widehat{P}_{X^n}, \mathcal{E}_1^c\right]}{\sqrt{\frac{1}{n}\mathrm{Var}\left(S_n \middle| \widehat{P}_{X^n}, \mathcal{E}_1^c\right)}}\right)\middle| \mathcal{E}_0^c\cap\mathcal{E}_1^c\right] + \xi_n, \\
&\forall\lambda\in\mathbb{R}. \quad (147)
\end{aligned}
$$

Using $\lim_{t\to 1}\sup_{\lambda\in\mathbb{R}}|\mathrm{Q}(\lambda) - \mathrm{Q}(\lambda/t)| \leqslant \lim_{t\to 1}\sup_{\lambda\in\mathbb{R}}|\lambda - \frac{\lambda}{t}|e^{-\lambda^2/3} = 0$ which follows from the Gaussian density bound, we see that the second term on the right side of (147) is upper bounded by

$$
\mathbb{E}\left[\mathrm{Q}\left(\frac{\frac{D}{c} - \mathbb{E}\left[\frac{1}{\sqrt{n}}S_n\middle| \widehat{P}_{X^n}, \mathcal{E}_1^c\right]}{\sqrt{V}}\right)\middle| \mathcal{E}_0^c\cap\mathcal{E}_1^c\right] + o(1) \quad (148)
$$

where we defined

$$
V := \mathbb{E}_{Q_X P^\star_{U|X}}\left[\mathrm{Var}_{Q_{Y|X}}\left(\imath_{Y^\star|U^\star} + \frac{1}{c}\imath_{U^\star;X^\star}\middle| UX\right)\right] \quad (149)
$$

to be the first term on the right side of (145). Using the uniform continuity of $\mathrm{Q}(\cdot)$ and the result of (138), we can upper bound the first term in (148) by

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{Q}\left(\frac{\frac{D}{c}-T_n}{\sqrt{V}}\right)\middle|\mathcal{E}_0^c\cap\mathcal{E}_1^c\right]+o(1)\\
&\leqslant \frac{1}{\mathbb{P}[\mathcal{E}_0^c\cap\mathcal{E}_1^c]}\mathbb{E}\left[\mathrm{Q}\left(\frac{\frac{D}{c}-T_n}{\sqrt{V}}\right)\right]+o(1)\\
&\leqslant \mathbb{E}\left[\mathrm{Q}\left(\frac{\frac{D}{c}-T_n}{\sqrt{V}}\right)\right]+o(1) \qquad (150)
\end{aligned}$$

where we defined the zero mean random variable

$$T_n := \sqrt{n}\left\langle \mathbb{E}_{P^\star_{U|X}Q_{Y|X}}\left[\imath_{Y^\star|U^\star}+\frac{1}{c}\imath_{U^\star;X^\star}\middle| X\right], \hat{P}_{X^n}-Q_X\right\rangle \qquad (151)$$

which is the first term in (138) normalized by $1/\sqrt{n}$. Clearly T_n is a sum of i.i.d. random variables, and the central limit theorem applies. Since convergence in distribution implies the convergence of the expectation of a bounded continuous function, we see that

$$\mathbb{E}\left[\mathrm{Q}\left(\frac{\frac{D}{c}-T_n}{\sqrt{V}}\right)\right]=\mathbb{E}\left[\mathrm{Q}\left(\frac{\frac{D}{c}-\sqrt{V'}G}{\sqrt{V}}\right)\right] \qquad (152)$$

where G is a Gaussian random variable with zero mean and unit variance and

$$\begin{aligned}
V' &:= \mathrm{Var}_{Q_X}\left(\mathbb{E}_{P^\star_{U|X}Q_{Y|X}}\left[\imath_{Y^\star|U^\star}+\frac{1}{c}\imath_{U^\star;X^\star}\middle| X\right]\right) \qquad (153)\\
&= \mathrm{Var}_{Q_XP^\star_{U|X}}\left(\mathbb{E}_{Q_{Y|X}}\left[\imath_{Y^\star|U^\star}+\frac{1}{c}\imath_{U^\star;X^\star}\middle| UX\right]\right), \qquad (154)
\end{aligned}$$

which is the variance of T_n. The equality in (154) follows from Proposition 2. Now

$$\begin{aligned}
\mathbb{E}\left[\mathrm{Q}\left(\frac{\frac{D}{c}-\sqrt{V'}G}{\sqrt{V}}\right)\right] &= \mathbb{P}\left[G'>\frac{\frac{D}{c}-\sqrt{V'}G}{\sqrt{V}}\right] \qquad (155)\\
&= \mathrm{Q}\left(\frac{D/c}{\sqrt{V+V'}}\right). \qquad (156)
\end{aligned}$$

Returning to (147), we now have

$$\mathbb{P}\left[\mathcal{E}_2|\mathcal{E}_0^c\cap\mathcal{E}_1^c\right]\leqslant \mathrm{Q}\left(\frac{D/c}{\sqrt{V+V'}}\right)+o(1) \qquad (157)$$

and moreover, we have that

$$\begin{aligned}
V+V' &= \\
&\mathbb{E}_{Q_XP^\star_{U|X}}\left[\mathrm{Var}_{Q_{Y|X}}\left(\imath_{Y^\star|U^\star}+\frac{1}{c}\imath_{U^\star;X^\star}\middle| UX\right)\right]\\
&+\mathrm{Var}_{Q_XP^\star_{U|X}}\left(\mathbb{E}_{Q_{Y|X}}\left[\imath_{Y^\star|U^\star}+\frac{1}{c}\imath_{U^\star;X^\star}\middle| UX\right]\right) \qquad (158)\\
&= \mathrm{Var}\left(\imath_{Y^\star|U^\star}+\frac{1}{c}\imath_{U^\star;X^\star}\right) \qquad (159)
\end{aligned}$$

by the law of total variance. We thus have shown that $\mathbb{P}[\mathcal{E}_0],\mathbb{P}[\mathcal{E}_1]\to 0$, and $\limsup_{n\to\infty}\mathbb{P}[\mathcal{E}_2]\leqslant \mathrm{Q}\left(\frac{D}{\sqrt{\mathrm{Var}\left(\imath_{Y^\star|U^\star}+\frac{1}{c}\imath_{U^\star;X^\star}\right)}}\right)$, as desired.

Remark 5: Asymptotic analysis of expressions similar to (147) has appeared in the literature of second-order analysis in information theory; see [43, Lemma 18] and [40, Lemma 2].

VII. Appendix: Reverse Hypercontractivity for the Transposition Model

In this section we construct the "dominating operator" $\Lambda_{n,t}$ used in Section V through several stages. As alluded in Section V-B, the idea is, roughly speaking, to find an operator $\Lambda_{n,t}$ with the crucial property that the 0-norm of $\Lambda_{n,t}f$ is bounded below in terms of the 1-norm of f, where f is an arbitrary function (usually taken as the indicator of the decoding set). This is essentially a statement of the reverse hypercontractivity property of the operator $\Lambda_{n,t}$, and we can thus make use of the available results on the reverse hypercontractivity estimates of Markov semigroup operators, originally studied by probabilists with the goal of understanding the mixing properties of the Markov chains.

Generally speaking, suppose that $(X_t)_{t\geqslant 0}$ is a time-homogenous Markov chain. Note that t here is a continuous time index, which is unrelated to the index i of the source coordinate in the previous sections. The Markov semigroup operator $(T_t)_{t\geqslant 0}$ is defined in terms of

$$(T_tf)(x)=P_{X_t|X_0=x}(f), \quad \forall x,\ f. \qquad (160)$$

Intuitively, we see that if the chain mixes fast, then T_tf converges to a constant function fast. This implies that the 0-norm of T_tf increases fast, since the expectation of T_tf is time-invariant whereas the 0-norm is mostly sensitive to the small values of the function.

If X_t is a vector, then a natural Markov chain is the *Glauber dynamics*, i.e. whenever a Poisson clock clicks, the value of a randomly selected coordinate is resampled according to the conditional distribution given the values of other coordinates. This approach works well, for example, when the coordinates of the vector are i.i.d., as we previously solved [30]. In the present setting, however, we consider the case where the vector is equiprobable on a type class, and it is clear that the Glauber dynamics does not mix (X_t will always take the same value). An intuitively similar chain is the *transposition model*, where each time we randomly pick two coordinates instead and make a switch. Clearly this chain mixes (i.e., converges to the equiprobable distribution on the type class). We now explain the construction in detail.

A. The Transposition Model

Let $\mathcal{S}=\{1,\dots,n\}$. As a preliminary step, consider a reversible Markov chain where the state space Ω consists of the $n!$ permutations of the sequence $(1,2,\dots,n)$, and the generator is given by

$$L_nf := \frac{1}{n}\sum_{1\leqslant i,j\leqslant n}(f\sigma_{ij}-f) \qquad (161)$$

for any real-valued function f on Ω, where $f\sigma_{ij}$ denotes the composition of two mappings, and σ_{ij} denotes the transposition operator. That is, σ_{ij} switches the i-th and the j-th coordinates of a sequence for any $s^n \in \Omega$,

$$(\sigma_{ij}s^n)_k := \begin{cases} s_i & k = j; \\ s_j & k = i; \\ s_k & \text{otherwise.} \end{cases} \tag{162}$$

As an alternative interpretation of this Markov chain, whenever a Poisson clock of rate $\frac{1}{n}$ clicks, an index pair $(i,j) \in \{1,\dots,n\}^2$ is randomly selected and the corresponding coordinates are switched. Remark that the rate at which each coordinate changes its value roughly equals 1, which is the same as the semi-simple Markov Chain we used in [30]. Functional inequalities such as Poincaré, log-Sobolev, and modified log-Sobolev for such a Markov chain have been studied to bound its mixing time under various metrics. In particular, we recall the following upper bound on the modified log-Sobolev constant in [13], which was proved using a chain-rule and induction argument:

Theorem 4 ([13]): Let P be the equiprobable distribution on Ω. For any $n \geqslant 2$,

$$D(S\|P) \leqslant -\mathbb{E}\left[\left(L_n \log \frac{\mathrm{d}S}{\mathrm{d}P}\right)(X)\right], \quad \forall S \ll P, \tag{163}$$

where $X \sim S$.

It is known (e.g. [33, Theorem 1.10]) that for general time-reversible Markov chains, a modified log-Sobolev inequality is equivalent to a reverse hypercontractivity of the corresponding Markov semigroup operator $e^{L_n t} := \sum_{k=0}^{\infty} \frac{t^k}{k!} L_n^k$. We thus have

Corollary 1: In the transposition model, For any $q < p < 1$, $t \geqslant \ln \frac{1-q}{1-p}$, and $f \in \mathcal{H}_+(\Omega)$,

$$\|e^{L_n t} f\|_{L^q(\Omega)} \geqslant \|f\|_{L^p(\Omega)}. \tag{164}$$

We remark that the norms in (164) are with respect to the equiprobable measure P. By taking the limits, we have

$$\|f\|_{L^0(\Omega)} = \exp\left(P(\ln f)\right). \tag{165}$$

B. Reverse Hypercontractivity on Types

Now consider any finite $\mathcal{Y}$ and a Markov chain with state space $\mathcal{Y}^n$. With a slight abuse of notation, let L_n also denote the generator of this new Markov chain. Let P_Y be an n-type. Note that $\mathcal{T}_n(P_Y)$ is invariant under transposition and hence also invariant for the chain. We now prove a reverse hypercontractivity for the Markov semigroup operator for this new chain. Pick any map $\phi\colon \mathcal{S} \to \mathcal{Y}$ such that $|\phi^{-1}(y)| = nP_Y(y)$ for each y. Then the extension ϕ^n defines a function $\Omega \to \mathcal{T}_n(P_Y)$. Now for any $f \in \mathcal{H}_+(\mathcal{Y}^n)$, from (164) we have

$$\|e^{L_n t}(f\phi^n)\|_{L^q(\Omega)} \geqslant \|f\phi^n\|_{L^p(\Omega)}. \tag{166}$$

Theorem 5: (166) is equivalent to

$$\|e^{L_n t} f\|_{L^q(\mathcal{T}_n(P_Y))} \geqslant \|f\|_{L^p(\mathcal{T}_n(P_Y))} \tag{167}$$

where the norms are understood as the norms of $f \in \mathcal{H}_+(\mathcal{Y}^n)$ with the underlying measure the equiprobable distribution on $\mathcal{T}_n(P_Y)$.

Proof: $\|f\phi^n\|_{L^p(\Omega)} = \mathbb{E}^{\frac{1}{p}}\left[(f\phi^n(S^n))^p\right] = \mathbb{E}^{\frac{1}{p}}[f^p(Y^n)] = \|f\|_{L^p(\mathcal{T}_n(P_Y))}$. Here, $L^p(\mathcal{T}_n(P_Y))$ is with respect to the equiprobable measure on $\mathcal{T}_n(P_Y)$, and so the value of f on $\mathcal{Y}^n \backslash \mathcal{T}_n(P_Y)$ is immaterial. Moreover, from the definitions we can see that ϕ^n commutes with transposition, so $\left(e^{L_n t}(f\phi)\right)(s^n) = (e^{L_n t} f)(\phi^n(s^n))$ for any $s^n \in \Omega$, and the left sides of (166) and (167) are therefore also equal by the same argument. □

We remark that for P_Y not concentrated on a $y \in \mathcal{Y}$ and as $n \to \infty$, we don't lose too much tightness in the composition step argument, and the estimate in (167) is sharp. That is, the modified log-Sobolev constant is indeed of the constant order; the lower bound can be seen by taking linear functions in the corresponding Poincaré inequality, which is weaker than the modified log-Sobolev inequality.

C. Conditional Types: The Tensorization Argument

Let $\mathcal{X}$ and $\mathcal{Y}$ both be finite sets. For any $x^n \in \mathcal{X}^n$, define a linear operator $L_{x^n}\colon \mathcal{H}_+(\mathcal{Y}^n) \to \mathcal{H}_+(\mathcal{Y}^n)$ by

$$L_{x^n} f := \sum_{x \in \mathcal{X}} \frac{1}{n\hat{P}_{x^n}(x)} \sum_{i,j\colon x_i = x_j = x} (f\sigma_{ij} - f). \tag{168}$$

where we recall that $\hat{P}_{x^n}$ denotes the empirical distribution of x^n. Note that L_{x^n} is the generator of the Markov chain where independently for each $x \in \mathcal{X}$, the length $nP_X(x)$ subsequence of $\mathcal{Y}^n$ with indices $\{i\colon x_i = x\}$ is the transposition model in Section VII-B. Since L_{x^n} is the sum of $|\mathcal{X}|$ generators for transposition models, the Markov semigroup operator $e^{L_{x^n} t}$ is a tensor product, which satisfies the reverse hypercontractivity with the same constant, by the tensorization property (see e.g. [33]). Therefore for any n-type P_{XY}, $x^n \in \mathcal{T}_n(P_X)$, and $f\colon \mathcal{H}_+(\mathcal{Y}^n) \to \mathcal{H}_+(\mathcal{Y}^n)$,

$$\|e^{L_{x^n} t} f\|_{L^q(\mathcal{T}_{x^n}(P_{Y|X}))} \geqslant \|f\|_{L^p(\mathcal{T}_{x^n}(P_{Y|X}))}. \tag{169}$$

D. A Dominating Operator

The operator in (169) depends on x^n and hence cannot be used directly in the proof of Lemma 1. We now find an upper bound which is independent of x^n. Define the following self-adjoint linear operator $\tilde{L}_n\colon \mathcal{H}_+(\mathcal{Y}^n) \to \mathcal{H}_+(\mathcal{Y}^n)$ by

$$\tilde{L}_n f(y^n) = \frac{1}{n \min_x P_X(x)} \sum_{1 \leqslant i,j \leqslant n} f(\sigma_{ij} y^n). \tag{170}$$

Note that the summation includes the $i = j$ case, where σ_{ij} becomes the identity. From the general formula $\frac{\mathrm{d}}{\mathrm{d}t}(e^{Lt} f) = Le^{Lt} f$ we can see a comparison property: since $\tilde{L}_n f \geqslant L_{x^n} f$ for any $f \in \mathcal{H}_+(\mathcal{Y}^n)$, we have $e^{\tilde{L}_n t} f \geqslant e^{L_{x^n} t} f$ pointwise for any $t \geqslant 0$ and $f \in \mathcal{H}_+(\mathcal{Y}^n)$. Now consider

$$\Lambda_{n,t} := e^{\tilde{L}_n t}, \quad \forall t > 0 \tag{171}$$

which forms an operator semigroup (although not associated with a conditional expectation). Now $\Lambda_{n,t}$ is the operator we used in the proof of Lemma 1, and we need to lower/upper bound the norms of $\Lambda_{n,t} f$ in terms of the norms of f:

Lower bound: for any $f\colon \mathcal{Y}^n \to [0,1]$,

$$\begin{aligned} &\exp(P_{Y^n|X^n=x^n}(\ln \Lambda_{n,t} f)) \\ &= \|\Lambda_{n,t} f\|_{L^0(\mathcal{T}_{x^n}(P_{Y|X}))} \\ &\geqslant \|e^{L_{x^n} t} f\|_{L^0(\mathcal{T}_{x^n}(P_{Y|X}))} \quad (172) \\ &\geqslant \|f\|_{L^{1-e^{-t}}(\mathcal{T}_{x^n}(P_{Y|X}))} \\ &\geqslant P_{Y^n|X^n=x^n}^{\frac{1}{1-e^{-t}}}(f) \\ &\geqslant P_{Y^n|X^n=x^n}^{1+\frac{1}{t}}(f) \quad (173) \end{aligned}$$

where (173) follows from $e^t \geqslant 1+t$.

Upper bound (in fact, equality): For any $f\colon \mathcal{Y}^n \to [0,\infty)$,

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}t} P_{Y^n}(\Lambda_{n,t} f) \\ &= P_{Y^n}(\tilde{L}_n \Lambda_{n,t} f) \quad (174) \\ &= \tfrac{1}{n \min_x P_X(x)} \sum_{y^n \in \mathcal{T}_n(P_Y)} P_{Y^n}(y^n) \sum_{i,j} (\Lambda_{n,t} f)(\sigma_{ij} y^n) \quad (175) \\ &= \tfrac{n}{\min_x P_X(x)} \sum_{z^n \in \mathcal{T}_n(P_Y)} P_{Y^n}(z^n)(\Lambda_{n,t} f)(z^n) \quad (176) \\ &= \tfrac{n}{\min_x P_X(x)} P_{Y^n}(\Lambda_{n,t} f) \quad (177) \end{aligned}$$

where (176) used the change of variable $z^n = \sigma_{ij} y^n$ and the fact that P_{Y^n} is the equiprobable distribution on $\mathcal{T}_n(P_Y)$, to get the factor n^2 out front. Thus

$$P_{Y^n}(\Lambda_{n,t} f) = \exp_e\left(\frac{nt}{\min_x P_X(x)}\right) P_{Y^n}(f). \quad (178)$$

VIII. Discussion

Though we have focused on the example of the Wyner-Ahlswede-Körner (WAK) network, the techniques of the present paper have other potential applications. There are several distributed source type problems which are very similar to the WAK problem. For example using a tensor product semigroup for the stationary memoryless settings, [22, Section 4.4.4] proved an $O(\sqrt{n})$ second-order converse for common random generation with one-way rate limited communications. It appears straightforward to upgrade to dispersion bounds and obtain similar results as WAK, by following the same steps therein but using the techniques of the present paper.

Acknowledgment

The author would like to thank Professors Ramon van Handel and Sergio Verdú for their generous support and guidance on research along this line. Helpful discussions with Professors Ramon van Handel, Shun Watanabe, and Victoria Kostina are greatly appreciated. Insightful comments on the presentations of the paper by anonymous reviewers are graciously acknowledged. The majority of this work was done while Jingbo Liu was at Princeton University, Princeton, NJ 08544, USA.

References

[1] R. Ahlswede, P. Gács, and J. Kárner, "Bounds on conditional probabilities with applications in multi-user communication," *Zeitschrift Für Wahrscheinlichkeitstheorie Verwandte Gebiete*, vol. 34, no. 2, pp. 157–177, 1976.

[2] R. Ahlswede and J. Korner, "Source coding with side information and a converse for degraded broadcast channels," *IEEE Trans. Inf. Theory*, vol. 21, no. 6, pp. 629–637, Nov. 1975.

[3] R. Ahlswede and I. Csiszar, "Common randomness in information theory and cryptography. II. CR capacity," *IEEE Trans. Inf. Theory*, vol. 44, no. 1, pp. 225–240, Jan. 1998.

[4] R. Ahlswede and I. Csiszar, "Hypothesis testing with communication constraints," *IEEE Trans. Inf. Theory*, vol. 32, no. 4, pp. 533–542, Jul. 1986.

[5] R. Ahlswede and P. Gacs, "Spreading of sets in product spaces and hypercontraction of the Markov operator," *Ann. Probab.*, vol. 4, no. 6, pp. 925–939, Dec. 1976.

[6] V. Anantharam, A. Gohari, S. Kamath, and C. Nair, "On maximal correlation, hypercontractivity, and the data processing inequality studied by erkip and cover," 2013, *arXiv:1304.6133*. [Online]. Available: http://arxiv.org/abs/1304.6133

[7] V. Anantharam, A. A. Gohari, S. Kamath, and C. Nair, "On hypercontractivity and the mutual information between Boolean functions," in *Proc. 51st Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Oct. 2013, pp. 13–19.

[8] S. Beigi and C. Nair, "Equivalent characterization of reverse Brascamp-Lieb-type inequalities using information measures," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2016, pp. 1038–1042.

[9] I. Csiszar, "Linear codes for sources and source networks: Error exponents, universal coding," *IEEE Trans. Inf. Theory*, vol. 28, no. 4, pp. 585–592, Jul. 1982.

[10] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless System*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[11] A. E. Gamal and Y.-H. Kim, *Networks Information Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[12] W. Feller, *An Introduction to Probablity Theory Its Application*, vol. 2. Hoboken, NJ, USA: Wiley, 1971.

[13] J. Quastel and F. Gao, "Exponential decay of entropy in the random transposition and Bernoulli-Laplace models," *Ann. Appl. Probab.*, vol. 13, no. 4, pp. 1591–1600, Nov. 2003.

[14] M. Hayashi, "Second-order asymptotics in fixed-length source coding and intrinsic randomness," *IEEE Trans. Inf. Theory*, vol. 54, no. 10, pp. 4619–4637, Oct. 2008.

[15] M. Hayashi, "Information spectrum approach to second-order coding rate in channel coding," *IEEE Trans. Inf. Theory*, vol. 55, no. 11, pp. 4947–4966, Nov. 2009.

[16] A. Ingber and Y. Kochman, "The dispersion of lossy source coding," in *Proc. Data Compress. Conf. (DCC)*, Snowbird, UT, USA, Mar 2011, pp. 53–62.

[17] J. Kahn, G. Kalai, and N. Linial, "The influence of variables on Boolean functions," in *Proc. 29th Annu. Symp. Found. Comput. Sci.*, 1988, pp. 68–80.

[18] S. Kamath, "Reverse hypercontractivity using information measures," in *Proc. 53rd Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Sep. 2015, pp. 627–633.

[19] V. Kostina, "Lossy data compression: Nonasymptotic fundamental limits," Ph.D. dissertation, Dept. Elect. Eng., Princeton Univ., Princeton, NJ, USA, 2013.

[20] V. Kostina and S. Verdú, "Nonasymptotic noisy lossy source coding," *IEEE Trans. Inf. Theory*, vol. 62, no. 11, pp. 6111–6123, Nov. 2016.

[21] V. Kostina and S. Verdú, "Fixed-length lossy compression in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 58, no. 6, pp. 3309–3338, Jun. 2012.

[22] J. Liu, "Information theory from a functional viewpoint," Ph.D. dissertation, Dept. Elect. Eng., Princeton Univ., Princeton, NJ, USA, 2018.

[23] J. Liu, T. A. Courtade, P. Cuff, and S. Verdú, "Information-theoretic perspectives on Brascamp-Lieb inequality and its reverse," 2017, *arXiv:1702.06260*. [Online]. Available: http://arxiv.org/abs/1702.06260

[24] J. Liu, T. A. Courtade, P. Cuff, and S. Verdú, "Brascamp-Lieb inequality and its reverse: An information theoretic view," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2016, pp. 1048–1052.

[25] J. Liu, T. A. Courtade, P. Cuff, and S. Verdú, "Smoothing brascamp-lieb inequalities and strong converses for common randomness generation," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2016, pp. 1043–1047.

[26] J. Liu, P. Cuff, and S. Verdú, "Secret key generation with one communicator and a one-shot converse via hypercontractivity," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2015, pp. 710–714.
[27] J. Liu and A. Ozgur, "New converses for the relay channel via reverse hypercontractivity," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2019, pp. 2878–2882.
[28] J. Liu and A. Ozgur, "Capacity upper bounds for the relay channel via reverse hypercontractivity," *IEEE Trans. Inf. Theory*, vol. 66, no. 9, pp. 5448–5455, Sep. 2020.
[29] J. Liu, R. van Handel, and S. Verdú, "Second-order converses via reverse hypercontractivity," *Math. Statist. Learn.*, vol. 2, no. 2, pp. 103–163, 2019.
[30] J. Liu, R. van Handel, and S. Verdú, "Beyond the blowing-up lemma: Sharp converses via reverse hypercontractivity," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2017, pp. 943–947.
[31] E. Mossel, "A quantitative arrow theorem," *Probab. Theory Rel. Fields*, vol. 154, nos. 1–2, pp. 49–88, Oct. 2012.
[32] E. Mossel, R. O'Donnell, and K. Oleszkiewicz, "Noise stability of functions with low influences: Invariance and optimality," in *Proc. 46th Annu. IEEE Symp. Found. Comput. Sci. (FOCS)*, Oct. 2005, pp. 21–30.
[33] E. Mossel, K. Oleszkiewicz, and A. Sen, "On reverse hypercontractivity," *Geometric Funct. Anal.*, vol. 23, no. 3, pp. 1062–1097, Jun. 2013.
[34] C. Nair and Y. N. Wang, "Reverse hypercontractivity region for the binary erasure channel," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 938–942.
[35] Y. Oohama, "Exponent function for one helper source coding problem at rates outside the rate region," 2015, *arXiv:1504.05891*. [Online]. Available: http://arxiv.org/abs/1504.05891
[36] Y. Oohama, "Exponential strong converse for one helper source coding problem," *Entropy*, vol. 21, p. 567, Jun. 2019.
[37] Y. Polyanskiy. (2013). *Hypothesis Testing Via a Comparator (Extended)*. [Online]. Available: http://people.lids.mit.edu/yp/homepage/data/htstruct_journal.pdf
[38] Y. Polyanskiy, H. V. Poor, and S. Verdu, "New channel coding achievability bounds," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2008, pp. 1763–1767.
[39] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.
[40] J. Scarlett, "On the dispersions of the Gel'fand–Pinsker channel and dirty paper coding," *IEEE Trans. Inf. Theory*, vol. 61, no. 9, pp. 4569–4586, Sep. 2015.
[41] V. Strassen, "Asymptotic estimates in Shannon's information theory," in *Proc. 3rd Trans. Prague Conf. Inf. Theory*, 1962, pp. 689–723.
[42] V. Y. F. Tan, "Asymptotic estimates in information theory with non-vanishing error probabilities," *Found. Trends Commun. Inf. Theory*, vol. 10, no. 4, pp. 1–184, 2014.
[43] M. Tomamichel and V. Tan, "Second-order coding rates for channels with state," *IEEE Trans. Inf. Theory*, vol. 60, no. 8, pp. 4427–4448, May 2014.
[44] H. Tyagi and S. Watanabe, "Strong converse using change of measure arguments," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2018, pp. 1849–1853.
[45] R. V. Handel. *Probability in High Dimension*. Accessed: Dec. 21, 2016. [Online]. Available: https://web.math.princeton.edu/~rvan/APC550.pdf
[46] S. Watanabe, "A converse bound on Wyner-Ahlswede-Körner network via Gray-Wyner network," 2017, *arXiv:1704.02262*. [Online]. Available: http://arxiv.org/abs/1704.02262
[47] S. Watanabe, "Second-order region for Gray–Wyner network," *IEEE Trans. Inf. Theory*, vol. 63, no. 2, pp. 1006–1018, Feb. 2017.
[48] S. Watanabe, S. Kuzuoka, and V. Y. F. Tan, "Nonasymptotic and second-order achievability bounds for coding with side-information," *IEEE Trans. Inf. Theory*, vol. 61, no. 4, pp. 1574–1605, Apr. 2015.
[49] A. Wyner, "On source coding with side information at the decoder," *IEEE Trans. Inf. Theory*, vol. 21, no. 3, pp. 294–300, May 1975.
[50] A. Xu and M. Raginsky, "Converses for distributed estimation via strong data processing inequalities," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2015, pp. 2376–2380.
[51] L. Zhou, V. Y. F. Tan, L. Yu, and M. Motani, "Exponential strong converse for content identification with lossy recovery," 2017, *arXiv:1702.06649*. [Online]. Available: http://arxiv.org/abs/1702.06649

Jingbo Liu (Member, IEEE) received the B.E. degree from Tsinghua University, Beijing, China, in 2012, and the Ph.D. degree from Princeton University, Princeton, NJ, USA, in 2018, both in electrical engineering. After two years of Post-Doctoral Researcher with the MIT Institute for Data, Systems, and Society (IDSS), he joined the Department of Statistics, University of Illinois, Urbana–Champaign as an Assistant Professor. His research interests include information theory, statistical inference, high-dimensional probability and statistics, and the related fields. His undergraduate thesis received the Best Undergraduate Thesis Award at Tsinghua University in 2012. He gave a semi-plenary presentation at the 2015 IEEE International Symposium on Information Theory, Hong-Kong, China. He was a recipient of the Princeton University Wallace Memorial Fellowship in 2016. His Ph.D. thesis received the Bede Liu Best Dissertation Award of Princeton and the Thomas M. Cover Dissertation Award of the IEEE Information Theory Society in 2018.