Beyond the Blowing-Up Lemma: Sharp Converses via Reverse Hypercontractivity

Jingbo Liu, Ramon van Handel, and Sergio Verdú
Princeton University

***Abstract*—This paper proposes a general method for establishing non-asymptotic converses in information theory via reverse hypercontractivity of Markov semigroups. In contrast to the blowing-up approach for strong converses, the proposed approach is applicable to non-discrete settings, and yields the optimal order of the second-order term in the rate expansion (square root of the blocklength) in the regime of non-vanishing error probability.**

I. Introduction

In information theory, a converse result shows the impossibility of an operational task in a given parameter range. For example, in data transmission, if an equiprobable message $X^n \in \{x_1, \ldots, x_M\}$ is transmissible through a random transformation $P_{Y^n|X^n}$ with error probability ϵ, then Fano's inequality states that $I(X^n;Y^n) \geqslant (1-\epsilon)\log M - h(\epsilon)$. Fano's inequality has found great success in providing *weak converses* in both single-user and multiuser information theory. However, for many channels, a much sharper bound is possible: one expects that $I(X^n;Y^n) \geqslant \log M - o(n)$ even when the error probability ϵ is nonvanishing. Such *strong converses* are out of reach of Fano's inequality and require more delicate arguments.

Similar issues arise in problems throughout information theory, and various techniques have been proposed to prove strong converses (e.g. [1]–[5]) that are applicable in different settings. The most powerful among these is the *blowing-up* lemma (BUL) introduced in [3] and systematically exploited in the classic text [6]. While much simpler methods may be feasible in specific problems, the power of the BUL is that it provides a canonical approach to proving strong converses that is widely applicable to discrete memoryless source and channel coding problems, and that remains the only known method for proving strong converses in certain problems of multiuser information theory (including all settings in [6] with known single-letter rate region; see [6, Ch. 16/Discussion]). On the other hand, the BUL approach has some key weaknesses:

- it is fundamentally restricted to finite output alphabets; and
- it is suboptimal in that it fails to yield the correct $O(\sqrt{n})$ second-order term in the strong converse bound.

In contrast, other methods [1]–[5] based on binary hypothesis testing and information spectrum methods can surmount these issues, but have met with limited success in multiuser setups.

In this work we show that the above weaknesses of the BUL approach can be overcome in a very general setting. We thereby provide a canonical approach for proving strong converses that, in addition to all the advantages of BUL, yields a second-order term which is sharp both in n and in ϵ and is applicable to continuous settings. Our key insight is that the fundamental operation of BUL, the violent *blowing up* of a set, is inherently suboptimal: one should instead gently *smooth out* the indicator function of the set by applying to it a Markov semigroup, exploiting functional forms of the relevant information-theoretic inequalities. The role of BUL is now played by the *reverse hypercontractivity* phenomenon [7], [8].

In order to develop our new approach in as pedagogical a manner as possible, we begin by illustrating the method in two simple problems: binary hypothesis testing (Section II) and sharp Fano's inequalities (Section III). While both these problems are amenable to simpler methods, they are included to illustrate the simplicity and unified nature of our approach. The highlight of this paper is Section IV, where we obtain an analogue (57) of the change-of-measure problem of [6] with optimal second-order term. In contrast to Sections II–III, no existing method for proving strong converses except BUL is known to be applicable to this setting. Section V provides a glimpse at applications of our new method in multiuser information theory in discrete as well as Gaussian settings. Among these, the side information problem (Section V-B) was highlighted as an open problem in [9, Section 9.2]. Omitted proofs and further applications can be found in [10].

Notation. In this paper, $\mathcal{H}_+(\mathcal{Y})$ denotes the set of nonnegative measurable functions on $\mathcal{Y}$ and $\mathcal{H}_{[0,1]}(\mathcal{Y})$ is the subset of $\mathcal{H}_+(\mathcal{Y})$ with range in $[0,1]$. For a measure ν and $f \in \mathcal{H}_+(\mathcal{Y})$, we write $\nu(f) := \int f \mathrm{d}\nu$ and $\|f\|_p^p = \|f\|_{L^p(\nu)}^p = \int |f|^p \mathrm{d}\nu$, while the measure of a set is denoted as $\nu(\mathcal{A})$. A random transformation $Q_{Y|X}$, mapping measures on $\mathcal{X}$ to measures on $\mathcal{Y}$, is viewed as an operator mapping $f \in \mathcal{H}_+(\mathcal{Y})$ to $Q_{Y|X}(f) \in \mathcal{H}_+(\mathcal{X})$ according to $Q_{Y|X}(f)(x) := Q_{Y|X=x}(f)$.

II. Prelude: Binary Hypothesis Testing

Many converses in information theory (notably, the *meta-converse* [5]) rely on the analysis of a certain binary hypothesis test (BHT). The aim of this section is to introduce and illustrate our method in the simplest BHT problem.

A. Blowing-Up Method: A Review

Consider probability distributions P and Q on $\mathcal{Y}$. Let $f \in \mathcal{H}_{[0,1]}(\mathcal{Y})$ be the probability of deciding the hypothesis P upon observing y. Denote by $\pi_{P|Q} = Q(f)$ the probability that P is decided when Q is true and vice versa by $\pi_{Q|P} = 1 - P(f)$. By the data processing property of the relative entropy

$$D(P\|Q) \geqslant d(\pi_{Q|P}\|1-\pi_{P|Q}) \tag{1}$$
$$\geqslant (1-\pi_{Q|P})\log\tfrac{1}{\pi_{P|Q}} - h(\pi_{Q|P}), \tag{2}$$

where $d(\cdot\|\cdot)$ is the binary relative entropy function on $[0,1]^2$. In the special case of product measures $P \leftarrow P^{\otimes n}$, $Q \leftarrow Q^{\otimes n}$, and $\pi_{Q^{\otimes n}|P^{\otimes n}} \leqslant \epsilon \in (0,1)$, (2) yields

$$\pi_{P^{\otimes n}|Q^{\otimes n}} \geqslant \exp\left(-n\frac{D(P\|Q)}{1-\epsilon} - O(1)\right). \tag{3}$$

However, this bound is known to be suboptimal: the Chernoff-Stein Lemma implies $\pi_{P^{\otimes n}|Q^{\otimes n}} \geqslant \exp(-nD(P\|Q) - o(n))$, a strong converse, while (3) only yields a weak converse.

Author Emails: {jingbo,rvan,verdu}@princeton.edu
This work was supported in part by NSF Grants CCF-1016625, CCF-0939370, and DMS-1148711, by ARO Grants W911NF-15-1-0479 and W911NF-14-1-0094, and by the Center for Science of Information.

In the special case of *deterministic* tests ($f = 1_{\mathcal{A}}$ for some $\mathcal{A} \subseteq \mathcal{Y}^n$), (3) can be improved by means of a remarkable property enjoyed by product measures: a small blowup of a set of nonvanishing probability suffices to increase its probability to nearly 1 [3]. The following "modern version" is due to Marton [11]; see also [12, Lemma 3.6.2].

Blowing-up Lemma. *Denote the r-blowup of $\mathcal{A} \subseteq \mathcal{Y}^n$ by*

$$\mathcal{A}_r := \{v^n \in \mathcal{Y}^n : d_n(v^n, \mathcal{A}) \leqslant r\}, \tag{4}$$

where d_n is the Hamming distance on $\mathcal{Y}^n$. Then, for any $c > 0$

$$P^{\otimes n}(\mathcal{A}_r) \geqslant 1 - e^{-c^2} \text{ for } r = \sqrt{\tfrac{n}{2}}\left(\sqrt{\ln \tfrac{1}{P^{\otimes n}(\mathcal{A})}} + c\right). \tag{5}$$

Moreover, as every element of $\mathcal{A}_r$ is obtained from an element of $\mathcal{A}$ by changing at most r coordinates, a simple counting argument [3, Lemma 5] shows that

$$Q^{\otimes n}(\mathcal{A}_r) \leqslant C^r (r+1) \binom{n}{r} Q^{\otimes n}(\mathcal{A}) \tag{6}$$

$$= \exp(nh(r/n) + O(r)) Q^{\otimes n}(\mathcal{A}) \tag{7}$$

for $r \in \Omega(\sqrt{n}) \cap o(n)$, where $C = |\mathcal{Y}|/\min_y Q(y)$.

Theorem 1. *If $|\mathcal{Y}| < \infty$ and $\pi_{Q^{\otimes n}|P^{\otimes n}} \leqslant \epsilon \in (0,1)$, any deterministic test between $P^{\otimes n}$ and $Q^{\otimes n}$ on $\mathcal{Y}^n$ satisfies*

$$\pi_{P^{\otimes n}|Q^{\otimes n}} \geqslant \exp\left(-nD(P\|Q) - O(\sqrt{n}\log^{3/2} n)\right). \tag{8}$$

The proof consists of: using, in lieu of (1),

$$nD(P\|Q) \geqslant d\left(P^{\otimes n}(\mathcal{A}_r)\|Q^{\otimes n}(\mathcal{A}_r)\right); \tag{9}$$

applying (5) and (7); and taking the optimal $r = \Theta(\sqrt{n\log n})$.

Of course, there are far easier methods to prove the converse part of the Chernoff-Stein Lemma than through Theorem 1, which also improve the suboptimal sublinear term in the exponent of (8). Indeed, our aim is to lower bound $Q^{\otimes n}(\mathcal{A})$ given $P^{\otimes n}(\mathcal{A}) \geqslant 1 - \epsilon$. By the Neyman-Pearson lemma, the extremal $\mathcal{A}$ is a sublevel set of $\imath_{P^{\otimes n}\|Q^{\otimes n}} := \log \frac{\mathrm{d}P^{\otimes n}}{\mathrm{d}Q^{\otimes n}}$, which is a sum of i.i.d. random variables under $P^{\otimes n}$. The optimal rate of the second-order term in (8) is thus the central limit theorem rate $O(\sqrt{n})$. Such an optimal bound is fundamentally beyond the blowing-up method: simple examples show that both (5) and (7) are essentially sharp and the choice $r = \Theta(\sqrt{n\log n})$ is optimal. In other classical applications of BUL such as the settings in Section III and IV, the suboptimal $O(\sqrt{n}\log^{3/2} n)$ term emerges for a similar reason. Moreover, in [10] we argue that this sub-optimality cannot be overcome by generalizing BUL to other set transformations $\mathcal{A} \mapsto \tilde{\mathcal{A}}$: no such argument can yield a second-order term better than $O(\sqrt{n\log n})$.

B. New Approach via Reverse Hypercontractivity

In this work we propose a gentler alternative to BUL that achieves the optimal $O(\sqrt{n})$ second-order rate, while retaining the applicability of BUL to multiuser information theory problems. Instead of applying the data processing theorem as in (1) and (9), we note that, by the variational formula for the relative entropy (e.g. [12, (2.4.67)])[1],

$$D(P\|Q) \geqslant P(\ln g) - \ln Q(g), \quad \forall g \in \mathcal{H}_+(\mathcal{Y}). \tag{10}$$

Tempting as it is, the choice $g \leftarrow 1_{\mathcal{A}}$ in (10) is dismal since generally $P(\ln 1_{\mathcal{A}}) = -\infty$. Inspired by BUL, we seek a map $T\colon \mathcal{H}_+ \to \mathcal{H}_+$ so that $P(\ln T1_{\mathcal{A}})$ is finite, yet $Q(T1_{\mathcal{A}})$ is not much larger than $Q(\mathcal{A})$. That is, rather than blowing up the set $\mathcal{A}$, we seek a *positivity improving* map T that maps the indicator $1_{\mathcal{A}}$ to a *strictly positive* function. Markov semigroups appear to be tailor-made for this purpose. We say $(T_t)_{t\geqslant 0}$ is a *simple semigroup*[2] with stationary measure P if

$$T_t\colon \mathcal{H}_+(\mathcal{Y}) \to \mathcal{H}_+(\mathcal{Y}), \quad f \mapsto e^{-t}f + (1-e^{-t})P(f). \tag{11}$$

In the i.i.d. case $P \leftarrow P^{\otimes n}$ we consider their tensor product

$$T_t := [e^{-t} + (1-e^{-t})P]^{\otimes n} \tag{12}$$

The positivity-improving property of T_t is precisely quantified by the *reverse hypercontractivity* phenomenon discovered by Borell [7] and recently generalized by Mossel et al. [8].

Theorem 2. *[8] Let $(T_t)_{t\geqslant 0}$ be a simple semigroup (11) or an arbitrary tensor product of simple semigroups. Then for all*[3] *$0 \leqslant q < p < 1$, $f \in \mathcal{H}_+$ and $t \geqslant \ln \frac{1-q}{1-p}$,*

$$\|T_t f\|_q \geqslant \|f\|_p. \tag{13}$$

When T_t is defined by (12), for any $f \in \mathcal{H}_{[0,1]}(\mathcal{Y})$,

$$P^{\otimes n}(\ln T_t f) = \ln \|T_t f\|_{L^0(P^{\otimes n})} \tag{14}$$

$$\geqslant \ln \|f\|_{L^{1-e^{-t}}(P^{\otimes n})} \tag{15}$$

$$\geqslant \tfrac{1}{1-e^{-t}} \ln P^{\otimes n}(f) \tag{16}$$

$$\geqslant \left(\tfrac{1}{t} + 1\right) \ln P^{\otimes n}(f), \tag{17}$$

where (17) follows from $e^t \geqslant 1 + t$. On the other hand, instead of the counting argument (7), in the present setting we use

$$Q^{\otimes n}(T_t f) = Q^{\otimes n}((e^{-t} + (1-e^{-t})P)^{\otimes n} f) \tag{18}$$

$$\leqslant Q^{\otimes n}((e^{-t} + \alpha(1-e^{-t})Q)^{\otimes n} f) \tag{19}$$

$$= (e^{-t} + \alpha(1-e^{-t}))^n Q^{\otimes n}(f) \tag{20}$$

$$\leqslant e^{(\alpha-1)nt} Q^{\otimes n}(f), \tag{21}$$

for all $f \in \mathcal{H}_+(\mathcal{Y})$, where $\alpha := \|\frac{\mathrm{d}P}{\mathrm{d}Q}\|_\infty \geqslant 1$.

Theorem 3. *If $\pi_{Q^{\otimes n}|P^{\otimes n}} \leqslant \epsilon \in (0,1)$, any (possibly stochastic) test between $P^{\otimes n}$ and $Q^{\otimes n}$ satisfies*

$$\pi_{P^{\otimes n}|Q^{\otimes n}} \geqslant (1-\epsilon)\exp\left(-nD(P\|Q) - 2\sqrt{n\left\|\tfrac{\mathrm{d}P}{\mathrm{d}Q}\right\|_\infty \ln \tfrac{1}{1-\epsilon}}\right). \tag{22}$$

The proof consists of: $g \leftarrow T_t f$ in (10) where $f(y)$ is the probability that the test decides P upon observing y; applying (17) and (21); optimizing over $t > 0$.

Remark 1. It is instructive to compare the blowing-up and semigroup operations. Note that

$$1_{\mathcal{A}_r}(x) = \sup_{|\mathcal{S}|\leqslant r} \sup_{(z_i)_{i\in\mathcal{S}}} 1_{\mathcal{A}}((z_i)_{i\in\mathcal{S}}, (x_i)_{i\in\mathcal{S}^c}), \tag{23}$$

while expanding (12) gives (cf. (31) below)

$$T_t 1_{\mathcal{A}}(x) = \mathbb{E}[1_{\mathcal{A}}((X_i)_{i\in\mathcal{S}}, (x_i)_{i\in\mathcal{S}^c})] \tag{24}$$

where $X_i \sim P$ are i.i.d. and $\mathcal{S}$ is uniformly distributed over sets of size $|\mathcal{S}| \sim \mathrm{Binom}(n, 1-e^{-t})$. Thus the semigroup can be viewed as an analogue of blowing-up where, rather than *maximize* the indicator over subsets of r coordinates, we *average* over subsets of $\approx n(1-e^{-t})$ coordinates. Unlike the maximum, averaging increases *small* values of f (positivity-improving)

[1] While the bases of the information theoretic quantities and exp were arbitrary up to here, Henceforth they are natural.

[2] Readers who are unfamiliar with semigroups may safely ignore this terminology; while the semigroup property plays an important role in the proof of Theorem 2, it is not used directly in this paper.

[3] We define $\|f\|_{L^0(P)} := \lim_{q\downarrow 0} \|f\|_{L^q(P)} = \exp(P(\ln f))$.

while preserving the total mass $P^{\otimes n}(T_t f) = P^{\otimes n}(f)$, so that $Q^{\otimes n}(T_t f)$ does not increase too much.

We note that our new method draws on a similar philosophy as BUL, but enjoys the following advantages:

- Achieves the optimal $O(\sqrt{n \log \frac{1}{1-\epsilon}})$ second-order rate, which is sharp both as $n \to \infty$ and as $\epsilon \to 1$ [10];
- Purely measure-theoretic in nature: finiteness of the alphabet is sufficient, but not necessary, for Theorem 3. In fact, our approach extends readily to continuous settings such as Gaussian channels, cf. Section III-B. In contrast, while analogues of the blowing-up property (5) exist for many other measures, no analogue of the counting estimate (7) can hold in continuous settings as the blow-up of a set of measure zero may have positive measure.

III. Optimal Second-Order Fano's Inequality

A. Bounded Probability Density Case

Consider a random transformation $P_{Y|X}$. For any $x \in \mathcal{X}$, denote by $(T_{x,t})_{t \geqslant 0}$ the simple Markov semigroup:

$$T_{x,t} := e^{-t} + (1 - e^{-t}) P_{Y|X=x}. \tag{25}$$

Motivated by the steps (18)-(21), for $\alpha \in [1, \infty)$, $t \in [0, \infty)$ and a probability measure ν on $\mathcal{Y}$, define a linear operator $\Lambda^{\nu}_{\alpha,t} \colon \mathcal{H}_+(\mathcal{Y}^n) \to \mathcal{H}_+(\mathcal{Y}^n)$ by

$$\Lambda^{\nu}_{\alpha,t} := \prod_{i=1}^{n} (e^{-t} + \alpha(1 - e^{-t}) \nu^{(i)}), \tag{26}$$

where $\nu^{(i)} \colon \mathcal{H}(\mathcal{Y}^n) \to \mathcal{H}(\mathcal{Y}^n)$ is the linear operator that integrates the i-th coordinate with respect to ν. Since $\nu^{(i)} 1 = 1$,

$$\Lambda^{\nu}_{\alpha,t} 1 = \left(e^{-t} + \alpha(1 - e^{-t})\right)^n \leqslant e^{(\alpha - 1) n t}. \tag{27}$$

Lemma 4. *Fix $(P_{Y|X}, \nu, t)$. Suppose that*

$$\alpha := \sup_x \left\| \frac{\mathrm{d} P_{Y|X=x}}{\mathrm{d} \nu} \right\|_{\infty} \in [1, \infty); \tag{28}$$

$$T_{x^n,t} := \otimes_{i=1}^{n} T_{x_i,t}. \tag{29}$$

Then for $n \geqslant 1$ and $f \in \mathcal{H}_+(\mathcal{Y}^n)$,

$$\sup_{x^n} T_{x^n,t} f \leqslant \Lambda^{\nu}_{\alpha,t} f. \tag{30}$$

Proof. For any $x^n \in \mathcal{X}^n$, observe that

$$T_{x^n,t} f = \prod_{i=1}^{n} [e^{-t} + (1 - e^{-t}) P^{(i)}_{Y|X=x_i}] f \tag{31}$$

$$= \sum_{\mathcal{S} \subseteq \{1,\ldots,n\}} e^{-|\mathcal{S}^c| t} (1 - e^{-t})^{|\mathcal{S}|} \left(\prod_{i \in \mathcal{S}} P^{(i)}_{Y|X=x_i} \right) (f),$$

The result then follows from (26) and (28). □

Theorem 5. *Fix $P_{Y|X}$ and positive integers n and M. Suppose (28) holds for some probability measure ν on $\mathcal{Y}$. If there exists $c_1, \ldots, c_M \in \mathcal{X}^n$ and disjoint $\mathcal{D}_1, \ldots, \mathcal{D}_M \subseteq \mathcal{Y}^n$ such that*

$$\prod_{m=1}^{M} P^{\frac{1}{M}}_{Y^n|X^n=c_m}(\mathcal{D}_m) \geqslant 1 - \epsilon, \tag{32}$$

then

$$I(X^n; Y^n) \geqslant \ln M - 2\sqrt{(\alpha - 1) n \ln \tfrac{1}{1-\epsilon}} - \ln \tfrac{1}{1-\epsilon}, \tag{33}$$

where X^n is equiprobable on $\{c_1, \ldots, c_M\}$, Y^n is its output from $P_{Y^n|X^n} := P^{\otimes n}_{Y|X}$, and α is defined in (28).

Proof. Let $f_m := 1_{\mathcal{D}_m}$, $m = 1, \ldots, M$. Fix some $t > 0$ to be optimized later. Observe that

$$I(X^n; Y^n) = \tfrac{1}{M} \sum_{m=1}^{M} D(P_{Y^n|X^n=c_m} \| P_{Y^n}) \tag{34}$$

$$\geqslant \tfrac{1}{M} \sum_{m=1}^{M} P_{Y^n|X^n=c_m} (\ln \Lambda^{\nu}_{\alpha,t} f_m)$$

$$- \tfrac{1}{M} \sum_{m=1}^{M} \ln P_{Y^n}(\Lambda^{\nu}_{\alpha,t} f_m) \tag{35}$$

where (35) is from the variational formula (10). We can lower bound the first term of (35) by

$$\tfrac{1}{M} \sum_{m=1}^{M} \ln \|\Lambda^{\nu}_{\alpha,t} f_m\|_{L^0(P_{Y^n|X^n=c_m})}$$

$$\geqslant \tfrac{1}{M} \sum_{m=1}^{M} \ln \|T_{c_m,t} f_m\|_{L^0(P_{Y^n|X^n=c_m})} \tag{36}$$

$$\geqslant -\left(\ln \tfrac{1}{1-\epsilon}\right)\left(1 + \tfrac{1}{t}\right), \tag{37}$$

where (36)-(37) follows similarly as (14)-(17). For the second term in the right of (35), using Jensen's inequality and (27),

$$-\tfrac{1}{M} \sum_{m=1}^{M} \ln P_{Y^n}(\Lambda^{\nu}_{\alpha,t} f_m) \geqslant -\ln P_{Y^n}\left(\tfrac{1}{M} \sum_{m=1}^{M} \Lambda^{\nu}_{\alpha,t} f_m\right) \tag{38}$$

$$\geqslant \ln M - (\alpha - 1) n t. \tag{39}$$

The result then follows by optimizing t. □

Remark 2. A novel aspect of Theorem 5 is the "geometric average criterion" in (32), which is weaker than the maximal error criterion but stronger than the average error criterion. Under the average error criterion, one cannot expect that $I(X^n; Y^n) \geqslant \ln M - o(n)$ as it would contradict [13, Rem. 5].

When the output alphabet $\mathcal{Y}$ is finite, Theorem 5 applies with $\alpha = |\mathcal{Y}|$ by choosing ν to be equiprobable on $\mathcal{Y}$. However, unlike BUL-based approaches, Theorem 5 extends far beyond the finite alphabet setting. In particular, Theorem 5 applies to Gaussian, Poisson, or exponential channels under an amplitude constraint, for which it is readily seen that the bounded density assumption (28) holds for a suitable choice of ν.

B. Gaussian Case

In this section we consider Gaussian channels, i.e., $P_{Y|X=x} = \mathcal{N}(x, 1)$. As noted above, Theorem 5 applies immediately in this case under an amplitude constraint $|x| \leqslant x_{\max}$ (choose $\nu = N(0, 2)$, say). However, we will show in this section that the conclusion of Theorem 5 remains valid even in the absence of an amplitude constraint, in which case (28) is violated. To surmount this problem, it will be convenient to replace (29) in the Gaussian setting by the *Ornstein-Uhlenbeck semigroup* with stationary measure $\mathcal{N}(x^n, \mathbf{I}_n)$:

$$T_{x^n,t} f(y^n) := \mathbb{E}[f(e^{-t} y^n + (1 - e^{-t}) x^n + \sqrt{1 - e^{-2t}} V^n)] \tag{40}$$

for $f \in \mathcal{H}_+(\mathbb{R}^n)$, where $V^n \sim \mathcal{N}(0^n, \mathbf{I}_n)$. In this setting, (13) holds under the even weaker assumption $t \geqslant \frac{1}{2} \ln \frac{1-q}{1-p}$ [7].

The proof proceeds a little differently than in the discrete case. Here, the analogue of $\Lambda^{\nu}_{\alpha,t}$ in (26) is simply $T_{0^n,t}$. Instead of Lemma 4, we will exploit a simple change-of-variable formula: for any $f \geqslant 0$, $t > 0$, $x^n \in \mathbb{R}^n$, we have

$$P_{Y^n|X^n=x^n}(\ln T_{0^n,t} f) = P_{Y^n|X^n=e^{-t}x^n}(\ln T_{e^{-t}x^n,t} f), \tag{41}$$

which can be verified from the definition in (40). For later applications in broadcast channels, we consider a slight extension of the setting of Theorem 5 to allow stochastic encoders.

Theorem 6. *Let $P_{Y|X=x} = \mathcal{N}(x, \sigma^2)$. Assume that there exist*

$$\phi \colon \{1, \ldots, M\} \times \{1, \ldots, L\} \to \mathbb{R}^n, \tag{42}$$

and disjoint sets $(\mathcal{D}_w)_{w=1}^M$ *such that*

$$\prod_{w,v} P_{Y^n|X^n=\phi(w,v)}^{\frac{1}{ML}}(\mathcal{D}_w) \geqslant 1-\epsilon \tag{43}$$

where $\epsilon \in (0,1)$. *Then*

$$I(W;Y^n) \geqslant \ln M - \sqrt{2n \ln \tfrac{1}{1-\epsilon}} - \ln \tfrac{1}{1-\epsilon} \tag{44}$$

where (W,V) *is equiprobable on* $\{1,\dots,M\}\times\{1,\dots,L\}$, $X^n := \phi(W,V)$, *and* Y^n *is the output from* $P_{Y^n|X^n} := P_{Y|X}^{\otimes n}$.

Proof. By the scaling invariance of the bound it suffices to consider the case of $\sigma^2 = 1$. Let $f_w = 1_{\mathcal{D}_w}$ for $w \in \{1,\dots,M\}$. Put $\bar{X}^n = e^t X^n$ and $\bar{Y}^n$ the corresponding output from the same channel. Note that for each w,

$$\begin{aligned} & D(P_{\bar{Y}^n|W=w} \| P_{\bar{Y}^n}) \\ \geqslant & \tfrac{1}{L}\sum_{v=1}^{L} P_{Y^n|X^n=e^t\phi(w,v)}(\ln T_{0^n,t} f_w) - \ln P_{\bar{Y}^n}(T_{0^n,t} f_w) \\ = & \tfrac{1}{L}\sum_{v=1}^{L} P_{Y^n|X^n=\phi(w,v)}(\ln T_{\phi(w,v),t} f_w) - \ln P_{\bar{Y}^n}(T_{0^n,t} f_w) \end{aligned} \tag{45}$$

where the key step (45) used (41). The summand in the first term of (45) can be bounded using reverse hypercontractivity for the Ornstein-Uhlenbeck semigroup (40) as

$$P_{Y^n|X^n=\phi(w,v)}(\ln T_{\phi(w,v),t} f_w) \geqslant \tfrac{1}{1-e^{-2t}} \ln P_{Y^n|X^n=\phi(w,v)}(f_w)$$

thus from the assumption (43),

$$\tfrac{1}{ML}\sum_{w,v} P_{Y^n|X^n=e^t\phi(w,v)}(\ln T_{0^n,t} f_w) \geqslant -\tfrac{1}{1-e^{-2t}} \ln \tfrac{1}{1-\epsilon}.$$

On the other hand, by Jensen's inequality,

$$\tfrac{1}{M}\sum_{w=1}^{M} \ln P_{\bar{Y}^n}(T_{0^n,t} f_w) \leqslant \ln P_{\bar{Y}^n}\left(T_{0^n,t}\tfrac{1}{M}\sum_{w=1}^{M} f_w\right) \leqslant \ln \tfrac{1}{M}$$

where the last step uses $\sum_{w=1}^M f_w \leqslant 1$. Thus taking $\frac{1}{M}\sum_{w=1}^M$ on both sides of (45), we find

$$I(W;\bar{Y}^n) \geqslant \ln M - \tfrac{1}{1-e^{-2t}} \ln \tfrac{1}{1-\epsilon}. \tag{46}$$

Moreover, let $G^n \sim \mathcal{N}(0^n, \mathbf{I}_n)$ be independent of X^n,

$$\begin{aligned} h(\bar{Y}^n) &= h(e^t X^n + G^n) && (47)\\ &= h(X^n + e^{-t} G^n) + nt && (48)\\ &\leqslant h(Y^n) + nt, && (49) \end{aligned}$$

where (49) can be seen from the entropy power inequality. On the other hand for each w we have

$$\begin{aligned} & h(\bar{Y}^n|W=w) - h(Y^n|W=w) \\ = & I(\bar{Y}^n; X^n|W=w) - I(Y^n; X^n|W=w) \geqslant 0 \end{aligned} \tag{50}$$

where (50) can be seen from [14, Theorem 1]. The result follows from (46), (49), (50) and by optimizing t. □

IV. Optimal Second-Order Change-of-Measure

We now revisit a change-of-measure problem considered in [3] in proving the strong converse of source coding with side information (cf. Section V-B). For a "combinatorial" formulation, which amounts to assigning an equiprobable distribution on the strongly typical set, see [6, Theorem 15.10]. Variations on such an idea have been applied to the "source and channel networks" in [6, Chapter 16] under the name "image-size characterizations". Given a random transformation $Q_{Y|X}$, measures ν on $\mathcal{Y}$ and μ_n on $\mathcal{X}^n$, we want to lower bound the $\nu^{\otimes n}$-measure of $\mathcal{A} \subseteq \mathcal{Y}^n$ in terms of the μ_n-measure of its "ϵ-preimage" under $Q_{Y^n|X^n} := Q_{Y|X}^{\otimes n}$. More precisely, for given $c > 0$, $\epsilon \in (0,1)$ find an upper bound on

$$\sup_{\mathcal{A}} \left\{\ln \mu_n[x^n \colon Q_{Y^n|X^n=x^n}(\mathcal{A}) > 1-\epsilon] - c \ln \nu^{\otimes n}(\mathcal{A})\right\}.$$

Definition 7. Fix μ on $\mathcal{X}$, ν on $\mathcal{Y}$, and $Q_{Y|X}$. For $c \in (0,\infty)$

$$\mathrm{d}(\mu, Q_{Y|X}, \nu, c) := \sup_{P_X \colon P_X \ll \mu} \{cD(P_Y\|\nu) - D(P_X\|\mu)\} \tag{51}$$

with $P_X \to Q_{Y|X} \to P_Y$.[4] Equivalently (see [15] or [10])

$$(51) = \sup_{f \in \mathcal{H}_+(\mathcal{Y})} \left\{\ln \mu(e^{cQ_{Y|X}(\ln f)}) - c \ln \nu(f)\right\}. \tag{52}$$

Obeserve that given Q_{XY}, the largest $c > 0$ for which $d(Q_X, Q_{Y|X}, Q_Y, c) = 0$ is the reciprocal of the *strong data processing constant*; see the references in [15]. If in definition (51) for $\mathrm{d}(\mu_n, Q_{Y|X}^{\otimes n}, \nu^{\otimes n}, c)$ we choose P_{X^n} to be μ_n conditioned on $\mathcal{B} := \{x^n \colon Q_{Y^n|X^n=x^n}(\mathcal{A}) > 1-\epsilon\}$, i.e., $P_{X^n}(\mathcal{C}) := \frac{\mu_n(\mathcal{B}\cap\mathcal{C})}{\mu_n(\mathcal{B})}$, $\forall \mathcal{C}$, then standard computations (essentially [3, (19)-(21)]; or see [10]) show that when $|\nu| = 1$,

$$\begin{aligned} & \ln \mu_n[x^n \colon Q_{Y^n|X^n=x^n}(\mathcal{A}) > 1-\epsilon] - c(1-\epsilon)\ln \nu^{\otimes n}(\mathcal{A}) \\ \leqslant & \mathrm{d}(\mu_n, Q_{Y|X}^{\otimes n}, \nu^{\otimes n}, c) + c\ln 2. \end{aligned} \tag{53}$$

Note the undesired second ϵ in (53), which would result in a weak converse. For finite $\mathcal{Y}$, [3] used the blowing-up lemma to strengthen (53). For the same reason discussed in Section II, even using modern results on concentration of measure, one can only obtain $O(\sqrt{n}\log^{3/2} n)$ in the second-order term:

$$\begin{aligned} & \ln \mu_n[x^n \colon Q_{Y^n|X^n=x^n}(\mathcal{A}) > 1-\epsilon] - c\ln \nu^{\otimes n}(\mathcal{A}) \\ \leqslant & \mathrm{d}(\mu_n, Q_{Y|X}^{\otimes n}, \nu^{\otimes n}, c) + O(\sqrt{n}\log^{3/2} n). \end{aligned} \tag{54}$$

If $\mu_n = Q_X^{\otimes n}$, then by tensorization ([15] or [10])

$$\mathrm{d}(Q_X^{\otimes n}, Q_{Y|X}^{\otimes n}, \nu^{\otimes n}, c) = n\,\mathrm{d}(Q_X, Q_{Y|X}, \nu, c), \tag{55}$$

we see the right side of (53) grows linearly with slope $\mathrm{d}(Q_X, Q_{Y|X}, \nu, c)$, which is larger than desired (i.e. when applied to the source coding problem in Section V-B would only result in an outer bound). Luckily, when $|\mathcal{X}| < \infty$, it was noted in [3] that if μ_n is the restriction of $Q_X^{\otimes n}$ on the Q_X-strongly typical set[5], then the linear growth rate of $\mathrm{d}(\mu_n, Q_{Y|X}^{\otimes n}, \nu^{\otimes n}, c)$ is the following (desired) quantity:

Definition 8. Given Q_X, $Q_{Y|X}$, measure ν on $\mathcal{Y}$ and $c \in (0,\infty)$, define

$$\begin{aligned} & \mathrm{d}^\star(Q_X, Q_{Y|X}, \nu, c) \\ := & \sup_{P_{UX} \colon P_X = Q_X} \left\{cD(P_{Y|U}\|\nu|P_U) - D(P_{X|U}\|Q_X|P_U)\right\} \end{aligned}$$

where $P_{UXY} := P_{UX}Q_{Y|X}$, and $D(P_{X|U}\|Q_X|P_U) := \int D(P_{X|U=\cdot}\|Q_X)\mathrm{d}P_U$ denotes the conditional relative entropy.

It follows from Definitions 7, 8 that $\mathrm{d}^\star(Q_X, Q_{Y|X}, \nu, c) \leqslant \mathrm{d}(Q_X, Q_{Y|X}, \nu, c)$. For general alphabets, we can extend the idea and let $\mu_n = Q_X^{\otimes n}\big|_{\mathcal{C}_n}$ for some $\mathcal{C}_n$ with $1 - Q_X^{\otimes n}(\mathcal{C}_n) \leqslant \delta$ for some $\delta \in (0,1)$ independent of n; this will not affect its information-theoretic applications in the non-vanishing error probability regime. We show in the discrete and the Gaussian cases that we can choose $\mathcal{C}_n$ so that

$$\mathrm{d}(\mu_n, Q_{Y|X}^{\otimes n}, \nu^{\otimes n}, c) \leqslant n\,\mathrm{d}^\star(Q_X, Q_{Y|X}, \nu, c) + O(\sqrt{n}). \tag{56}$$

[4] In Definitions 7 and 8 we adopt the convention $\infty - \infty = -\infty$.

[5] The restriction $\mu|_{\mathcal{C}}$ of a measure μ on a set $\mathcal{C}$ is $\mu|_{\mathcal{C}}(\mathcal{D}) := \mu(\mathcal{C}\cap\mathcal{D})$.

Using the semigroup method, we can also improve the $O(\sqrt{n}\log^{3/2} n)$ term in (54) to $O(\sqrt{n})$ in the discrete and the Gaussian cases. This combined with (56) implies

$$\begin{aligned}&\ln \mu_n[x^n\colon Q_{Y^n|X^n=x^n}(\mathcal{A}) > 1-\epsilon] - c\ln \nu^{\otimes n}(\mathcal{A})\\ &\leqslant n\,\mathrm{d}^\star(Q_X, Q_{Y|X}, \nu, c) + O(\sqrt{n}).\end{aligned} \tag{57}$$

While the original proof [3] used a data processing argument, just as (2), to get (53), the present approach uses the functional inequality (52), just as (10). In the discrete case we have:

Theorem 9. *Consider Q_X a probability measure on a finite set $\mathcal{X}$, ν a probability measure on $\mathcal{Y}$, and $Q_{Y|X}$. Let*

$$\beta_X := 1/\min_x Q_X(x) \in [1,\infty), \tag{58}$$

$$\alpha := \sup_x \left\| \tfrac{\mathrm{d}Q_{Y|X=x}}{\mathrm{d}\nu} \right\|_\infty \in [1,\infty). \tag{59}$$

Let $c \in (0,\infty)$, $\eta,\delta \in (0,1)$ and $n > 3\beta_X \ln \frac{|\mathcal{X}|}{\delta}$. We can choose some set $\mathcal{C}_n$ with $Q_X^{\otimes n}(\mathcal{C}_n) \geqslant 1-\delta$, such that for $\mu_n := Q_X^{\otimes n}|_{\mathcal{C}_n}$ we have

$$\begin{aligned}&\ln \mu_n[x^n\colon Q_{Y^n|X^n=x^n}(f) \geqslant \eta] - c\ln \nu^{\otimes n}(f)\\ &\leqslant n\mathrm{d}^\star(Q_X, Q_{Y|X}, \nu, c) + A\sqrt{n} + c\ln\tfrac{1}{\eta}\end{aligned} \tag{60}$$

for any $f \in \mathcal{H}_{[0,1]}(\mathcal{Y}^n)$, where

$$A := \ln(\alpha^c \beta_X^{c+1})\sqrt{3\beta_X \ln\tfrac{|\mathcal{X}|}{\delta}} + 2c\sqrt{(\alpha-1)\ln\tfrac{1}{\eta}}. \tag{61}$$

The proof of Theorem 9 relies on some ideas in the proof of Theorem 5; in particular the properties (27) and (30) for the operator $\Lambda^\nu_{\alpha,t}$ play a critical role. For a counterpart of Theorem 9 for Gaussian sources, see [10].

V. Applications

A. Output Distribution of Channel Codes; Broadcast Channels

Consider a stationary memoryless channel $P_{Y|X}$ with capacity C. If $|\mathcal{Y}| < \infty$ then by the steps [13, (64)-(66)] and our Theorem 5 we conclude that an (n, M, ϵ) code under the maximal error criterion with deterministic encoders satisfies

$$D(P_{Y^n}\|P^\star_{Y^n}) \leqslant nC - \ln M + 2\sqrt{|\mathcal{Y}| n \ln\tfrac{1}{1-\epsilon}} + \ln\tfrac{1}{1-\epsilon}. \tag{62}$$

This implies that the Burnashev condition

$$\sup_{x,x'} \left\| \tfrac{\mathrm{d}P_{Y|X=x}}{\mathrm{d}P_{Y|X=x'}} \right\|_\infty < \infty \tag{63}$$

in [13, Theorem 6] (cf. [12, Theorem 3.6.6]) is not necessary. Using the blowing-up lemma, [13, Theorem 7] bounded $D(P_{Y^n}\|P^\star_{Y^n})$ without requiring (63), but with a suboptimal $O(\sqrt{n}\log^{3/2} n)$ second-order term. Also note that in our approach $|\mathcal{Y}| < \infty$ can be weakened to a bounded density assumption (28), and the maximal error criterion assumption can be weakened to the geometric average criterion (32).

Consider a Gaussian broadcast channel where the SNR in the two component channels are $S_1, S_2 \in (0,\infty)$. Suppose there exists an (n, M_1, M_2, ϵ)-maximal error code. Using Theorem 6 and the same steps in the proof of the weak converse (see e.g. [16, Theorem 5.3]), we immediately obtain

$$\ln M_1 \leqslant n\mathsf{C}(\alpha S_1) + \sqrt{2n\ln\tfrac{1}{1-\epsilon}} + \ln\tfrac{1}{1-\epsilon}; \tag{64}$$

$$\ln M_2 \leqslant n\mathsf{C}\left(\tfrac{(1-\alpha)S_2}{\alpha S_2+1}\right) + \sqrt{2n\ln\tfrac{1}{1-\epsilon}} + \ln\tfrac{1}{1-\epsilon}, \tag{65}$$

for some $\alpha \in [0,1]$, where $\mathsf{C}(t) := \frac{1}{2}\ln(1+t)$. An alternative proof of the strong converse via information spectrum, yielding less precise (e.g. suboptimal dependence on ϵ and the signal to noise ratio) bounds on the sublinear terms, is given in [17]. Converses under the average error criterion can be obtained by codebook expurgation (e.g. [16, Problem 8.11]).

An $O(\sqrt{n})$ second-order counterpart for the discrete broadcast channel is given in [10].

B. Source Coding with Compressed Side Information

Consider a stationary memoryless discrete source with per-letter distribution Q_{XY}. Let $\epsilon \in (0,1)$ and $n \geqslant 3\beta_X \ln\frac{2(1+\epsilon)|\mathcal{X}|}{1-\epsilon}$, where β_X is defined in (58). Suppose that there exist encoders $f\colon \mathcal{X}^n \to \mathcal{W}_1$ and $g\colon \mathcal{Y}^n \to \mathcal{W}_2$ and decoder $V\colon \mathcal{W}_1 \times \mathcal{W}_2 \to \hat{\mathcal{Y}}^n$ such that

$$\mathbb{P}[Y^n \neq \hat{Y}^n] \leqslant \epsilon, \tag{66}$$

Using Theorem 9 plus essentially the same arguments as [3, Theorem 3], we can show that for any $c \in (0,\infty)$,

$$\begin{aligned}&\ln|\mathcal{W}_1| + c\ln|\mathcal{W}_2|\\ &\geqslant n \inf_{U\colon U-X-Y}\{cH(Y|U) - I(U;X)\}\\ &\quad - \sqrt{n}\left(\ln(|\mathcal{Y}|^c\beta_X^{c+1})\sqrt{3\beta_X\ln\tfrac{4|\mathcal{X}|}{1-\epsilon}} + 2c\sqrt{|\mathcal{Y}|\ln\tfrac{2}{1-\epsilon}}\right)\\ &\quad - (1+c)\ln\tfrac{2}{1-\epsilon}.\end{aligned} \tag{67}$$

Note that the first term on the right side of (67) corresponds to the rate region (see e.g. [16, Theorem 10.2]). Using the BUL [3], one can only bound the second term as $O(\sqrt{n}\log^{3/2} n)$, which is suboptimal. A counterpart for Gaussian sources under the quadratic distortion is given in [10].

References

[1] C. E. Shannon, R. G. Gallager, and E. Berlekamp, "Lower bounds to error probability for coding on discrete memoryless channels, I," *Information and Control*, vol. 10, pp. 65–103, 1967.

[2] J. Wolfowitz, "Notes on a general strong converse," *Information and Control*, vol. 12, pp. 1–4, 1968.

[3] R. Ahlswede, P. Gács, and J. Körner, "Bounds on conditional probabilities with applications in multi-user communication," *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 1976.

[4] S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Trans. Inf. Theory*, vol. 40, pp. 1147–1157, July 1994.

[5] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

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

[7] C. Borell, "Positivity improving operators and hypercontractivity," *Mathematische Zeitschrift*, vol. 180, no. 3, pp. 225–234, 1982.

[8] E. Mossel, K. Oleszkiewicz, and A. Sen, "On reverse hypercontractivity," *Geometric and Functional Analysis*, vol. 23, no. 3, pp. 1062–1097, 2013.

[9] V. Y. Tan, "Asymptotic estimates in information theory with non-vanishing error probabilities," *Foundations and Trends in Communications and Information Theory*, vol. 11, no. 1-2, pp. 1–184, 2014.

[10] J. Liu, R. van Handel, and S. Verdú, "Beyond the blowing-up lemma: Sharp converses via reverse hypercontractivity." http://www.princeton.edu/~jingbo/preprints/ISITBU.pdf.

[11] K. Marton, "A simple proof of the blowing-up lemma," *IEEE Trans. Inf. Theory*, vol. 24, pp. 857–866, 1986.

[12] M. Raginsky and I. Sason, "Concentration of measure inequalities in information theory, communications, and coding," *Foundations and Trends in Communications and Information Theory*, vol. 10, no. 1-2, pp. 1–250, 2013 (revision: 2014).

[13] Y. Polyanskiy and S. Verdú, "Empirical distribution of good channel codes with nonvanishing error probability," *IEEE Trans. Inf. Theory*, vol. 60, no. 1, pp. 5–21, 2014.

[14] D. Guo, S. Shamai, and S. Verdú, "Mutual information and minimum mean-square error in Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1261–1282, 2005.

[15] J. Liu, T. A. Courtade, P. Cuff, and S. Verdú, "Brascamp-Lieb inequality and its reverse: An information theoretic view," in *Proc. of IEEE International Symposium on Information Theory*, pp. 1048–1052, July 2016, Barcelona, Spain.

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

[17] S. L. Fong and V. Y. Tan, "A proof of the strong converse theorem for Gaussian broadcast channels via the Gaussian Poincaré inequality," in *Proc. of IEEE International Symposium on Information Theory*, pp. 170–174, July 2016, Barcelona, Spain.