# A Strong Entropy Power Inequality

Thomas A. Courtade, *Member, IEEE*

***Abstract*— When one of the random summands is Gaussian, we sharpen the entropy power inequality (EPI) in terms of the strong data processing function for Gaussian channels. Among other consequences, this 'strong' EPI generalizes the vector extension of Costa's EPI to non-Gaussian channels in a precise sense. This leads to a new reverse EPI and, as a corollary, sharpens Stam's uncertainty principle relating entropy power and Fisher information (or, equivalently, Gross' logarithmic Sobolev inequality). Applications to network information theory are also given, including a short self-contained proof of the rate region for the two-encoder quadratic Gaussian source coding problem and a new outer bound for the one-sided Gaussian interference channel.**

***Index Terms*— Entropy power inequality, Costa's EPI, Stam's inequality, reverse EPI, strong data processing, gaussian source coding.**

## I. Introduction and Main Result

FOR a random vector $X$ with density $f$ on $\mathbb{R}^d$, the differential entropy of $X$ is defined by

$$h(X) = -\int_{\mathbb{R}^d} f(x)\log f(x)dx, \tag{1}$$

where we adopt the convention that the logarithm is computed with respect to the natural base. The celebrated Entropy Power Inequality (EPI) put forth by Shannon [2] and rigorously established by Stam [3] and Blachman [4] asserts that for $X, W$ independent and $Y = X + W$,

$$e^{\frac{2}{d}h(Y)} \geq e^{\frac{2}{d}h(X)} + e^{\frac{2}{d}h(W)}. \tag{2}$$

Under the assumption that $W$ is Gaussian, our main result is the following improvement of (2):

*Theorem 1: Suppose $X, W$ are independent random vectors in $\mathbb{R}^d$, and moreover that $W$ is Gaussian. Define $Y = X + W$. For any $V$ satisfying $X \to Y \to V$,*

$$e^{\frac{2}{d}(h(Y)-I(X;V))} \geq e^{\frac{2}{d}(h(X)-I(Y;V))} + e^{\frac{2}{d}h(W)}. \tag{3}$$

The notation $X \to Y \to V$ in Theorem 1 indicates that the random variables $X$, $Y$ and $V$ form a Markov chain, in that order (this and other notations are detailed in Section II). In case the integral (1) does not exist, or if $X$ does not have density, then we adopt the convention that $h(X) = -\infty$. In this case, the inequality (3) is a trivial consequence of the data processing inequality. Likewise, the inequality is also degenerate when $h(X) = \infty$. So, as with the classical EPI, Theorem 1 is only informative when the entropy $h(X)$ exists and is finite.

Let us briefly remark on equality cases of Theorem 1 in comparison to the classical EPI. First, we recall that the Shannon-Stam-Blachman EPI (2) attains equality precisely when $X$ and $W$ are Gaussian with proportional covariances [3], [5]. In contrast, by judiciously choosing the auxiliary random variable $V$, equality can be achieved in (3) for any given $X$ with finite entropy. In fact, choosing $V = X + W$ renders both sides equal to $e^{\frac{2}{d}h(W)}$. Furthermore, in analogy to the classical EPI, it is a simple calculation to see that equality is also achieved in (3) whenever the random variables $X$, $Y$ and $V$ are jointly Gaussian with proportional covariances.

For $X, Y$ as in the statement of Theorem 1, we may associate a function $\Psi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ to their joint law $P_{XY}$ as follows

$$\Psi(t) := \sup_{V: X\to Y\to V} \{I(X;V) : I(Y;V) \leq t\}, \tag{4}$$

where the supremum ranges over all random variables $V$ satisfying the Markov relation $X \to Y \to V$. The function $\Psi$ is generally referred to as a *strong data processing function* because it sharpens the classical data processing inequality for mutual information in a certain way. Indeed, definitions ensure $I(X;V) \leq \Psi(I(Y;V)) \leq I(Y;V)$ for all $V$ satisfying $X \to Y \to V$, and $\Psi$ is the (pointwise) smallest function for which the first inequality holds. With $\Psi$ defined in this way, Theorem 1 may be recast in terms of $\Psi$ as

$$e^{\frac{2}{d}(h(Y)-\Psi(t))} \geq e^{\frac{2}{d}(h(X)-t)} + e^{\frac{2}{d}h(W)} \quad \forall t \geq 0. \tag{5}$$

Hence, Theorem 1 strengthens the EPI when one of the variables is Gaussian, and does so precisely in terms of the strong data processing function $\Psi$. According to this, we find it fitting to refer to (3) as a 'strong' entropy power inequality, giving this paper its title.

It should be noted that Calmon, Polyanskiy and Wu [6], [7] have recently considered a somewhat related problem where they bound the best-possible data processing function defined according to

$$F_I(t,\gamma) := \sup_{U,X: U\to X\to Y} \{I(Y;U) : I(X;U) \leq t\}, \tag{6}$$

where $Y = X + W$, $W \sim N(0, I)$ is independent of $(X, U)$, and the supremum is over all joint distributions $P_{UX}$ such that $\mathbb{E}[X^2] \leq \gamma$. Since the Markov assumptions in (4) and (6) differ, and Calmon *et al.* further optimize over the distribution $P_X$, the present results and those of [6], [7] are not comparable.

Manuscript received June 1, 2016; revised October 2, 2017; accepted October 12, 2017. Date of publication December 14, 2017; date of current version March 15, 2018. This work was supported by the Center for Science of Information NSF under Grant CCF-1528132 and Grant CCF-0939370. This paper was presented in part at the 2016 International Symposium on Information Theory [1]

The author is with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720 USA.

Communicated by M. Madiman, Associate Editor for Probability and Statistics.

Digital Object Identifier 10.1109/TIT.2017.2779745

Various improvements of the EPI have appeared previously in the literature. However, none are highly similar to (3). For instance, strengthened EPIs for random vectors with log-concave densities have recently been proposed by Toscani [8] and Courtade, Fathi and Pananjady [9]. In another direction, Madiman and Barron have established an EPI for subsets of random variables [10], which generalizes the monotonicity of entropy along the central limit theorem [11]. The reader is referred to the recent survey [12] for a general overview of EPI-related results.

Among those inequalities appearing in the literature, we consider Costa's EPI [13] to be most comparable to Theorem 1. In the following section, we clarify this relationship by demonstrating that Theorem 1 generalizes Costa's EPI [13] and its vector extension [14], which enjoy applications ranging from interference channels to secrecy capacity (e.g., [14]–[17]). However, Theorem 1 goes considerably further than generalizing Costa's EPI. In the context of functional inequalities, we will see that Theorem 1 leads to new reverse entropy power and Fisher information inequalities, which in turn can be applied to sharpen the Gaussian logarithmic Sobolev inequality. On the other hand, in the context of coding theorems, we will see that Theorem 1 leads to a concise proof of the converse for the rate region of the quadratic Gaussian two-encoder source-coding problem [18], [19] (a result which seems beyond the reach of Costa's EPI). Applications to one-sided interference channels and strong data processing inequalities are also briefly discussed.

The restriction of $W$ to be Gaussian in Theorem 1 should not be a severe limitation in practice. In applications of the EPI, it is often the case that one of the variables is Gaussian. As noted by Rioul [20], examples include the scalar Gaussian broadcast channel problem [21] and its generalization to the multiple-input multiple-output case [22], [23]; the secrecy capacity of the Gaussian wiretap channel [24] and its multiple access extension [25]; determination of the corner points for the scalar Gaussian interference channel problem [15], [16]; the scalar Gaussian source multiple-description problem [26]; and characterization of the rate-distortion regions for several multiterminal Gaussian source coding schemes [19], [27], [28]. Furthermore, the EPI with one Gaussian variable is equivalent to the Gaussian logarithmic Sobolev inequality [29], which has a number of important applications in analysis (e.g,. [30], [31]). It is tempting to conjecture that (3) holds even when the distribution of $W$ is non-Gaussian, and we have found no evidence to suggest the contrary. This is a very interesting question, which we leave as an open problem.

### A. Organization

The remainder of this paper is organized as follows: Section II defines notation and Section III explores various applications of Theorem 1. For the reader interested only in applications of the main result, Sections IV and beyond can be safely omitted as they are dedicated to the proof of Theorem 1.

As for the proof of Theorem 1, it is rather long and we know of no simpler proof at this time. Section IV is intended to familiarize the reader with the basic strategy of the proof. In particular, Section IV-A is devoted to a heuristic discussion of the crux of our argument, in which we sketch the basic ideas involved in proving a one-dimensional version of Theorem 1, but gloss over several technical issues that need to be dealt with. Section V shows how to recover Theorem 1 from its one-dimensional counterpart by leveraging some of the machinery developed in Section IV.

Section VI and the two appendices are for readers interested in the details of the proof. Specifically, Section VI revisits the heuristic discussion of Section IV-A, and makes it rigorous by treating the subtle points with more care. In order to arrive at its main conclusion, Section VI requires two key technical estimates whose proofs are essentially orthogonal to the rest of the argument. Appendices A and B are dedicated to establishing these separate results.

## II. Notation

This section establishes basic notation, most of which is standard in the information theory literature. We write $P_{XY}, P_X$ and $P_Y$ to denote the joint and respective marginal probability distributions associated to a pair of random variables $(X, Y)$. We write $X \sim P_X$ to denote $X$ has distribution $P_X$, and similarly $(X, Y) \sim P_{XY}$ to indicate that $(X, Y)$ have joint law $P_{XY}$, and so forth. We shall use $Y|\{X = x\}$ to denote the random variable $Y$ conditional on $\{X = x\}$, and denote its law by $P_{Y|\{X=x\}}$. Note that $Y|\{X = x\}$ is uniquely defined in the sense that different versions of the same are equal $P_X$-a.e. $x$; as such, we will not distinguish between these versions. It will be often convenient to factor joint distributions into products of marginals and conditionals, so that $(X, Y) \sim P_X P_{Y|X}$ denotes that $X, Y$ are jointly distributed such that $X \sim P_X$ and $Y|\{X = x\} \sim P_{Y|\{X=x\}}$. This notation is particularly handy to describe conditional independence. For example, any joint distribution $P_{XYV}$ may be factored as $P_{XY} P_{V|XY}$. So, if we write $(X, Y, V) \sim P_{XY} P_{V|Y}$, this implies that the conditional law $P_{V|XY} = P_{V|Y}$ does not depend on $X$, and therefore $V$ and $X$ are conditionally independent given $Y$. As already introduced previously, conditional independence (i.e., Markov) structure can be compactly represented as $X \to Y \to V$ to denote $P_{XYV} = P_X P_{Y|X} P_{V|Y}$. When additional conditioning is needed, we write $X \to Y \to V|Q$ to indicate that $P_{QXYV} = P_Q P_{X|Q} P_{Y|XQ} P_{V|YQ}$. In other words, $X \to Y \to V$ form a Markov chain, conditioned on $Q$. So, for example, if we write

$$\inf_{V: X \to Y \to V|Q} \Big\{ I(Y; V|Q) - \lambda I(X; V|Q) \Big\},$$

this denotes an infimum taken over all distributions $P_{V|YQ}$, where the information measures in the argument of the infimum are evaluated with respect to the joint distribution $P_{QXYV} = P_Q P_{X|Q} P_{Y|XQ} P_{V|YQ}$. Reader familiarity with information measures (entropy, mutual information, etc.) and their calculus is assumed. For the unfamiliar reader, definitions may be found in any standard textbook, e.g., [32].

A sequence of random variables $X_1, X_2, \ldots$ indexed by $n \in \mathbb{N}$ will be denoted by the shorthand $\{X_n\}$, and convergence

of $\{X_n\}$ in distribution to a random variable $X_*$ is written $X_n \xrightarrow{\mathcal{D}} X_*$.

We write $W \sim N(\mu, \Sigma)$ to indicate that $W$ has Gaussian distribution with mean $\mu$ and covariance $\Sigma$. We will often be interested in the *Gaussian channel* parametrized by the scalar quantity $\varrho > 0$ (the *signal-to-noise ratio*), in which $Y = \sqrt{\varrho}X + Z$, where $Z \sim N(0,1)$ is independent of the input $X$. For this particular situation, we reserve the notation $\mathbb{G}^{\varrho}_{Y|X}$ to denote the conditional law of $Y$ given $X$. Hence, writing $(Q_n, X_n, Y_n) \sim P_{Q_n X_n}\mathbb{G}^{\varrho}_{Y_n|X_n}$ is compact notation for $(Q_n, X_n) \sim P_{Q_n,X_n}$ and $Y_n = \sqrt{\varrho}X_n + Z$, where $Z \sim N(0,1)$ is independent of $(Q_n, X_n)$.

For a conditional law $P_{Y|X}$ and a random variable $X \sim P_X$, we sometimes employ the compact notation $P_{Y|X} : X \mapsto Y$ to indicate that $Y$ is a random variable with law obtained by composing $P_{Y|X}$ with $P_X$. In particular, given a sequence $\{X_n\}$, we may define a sequence $\{Y_n\}$ via $P_{Y|X} : X_n \mapsto Y_n$, which means that the conditional law of $Y_n$ given $X_n$ is equal to $P_{Y|X}$ for each $n$.

Without loss of generality, it is assumed throughout that all logarithms are base-$e$, so that all information measures are in units of *nats*. Exceptions to this convention occur in Sections III-C and III-D, and are explicitly noted. Finally, $\|\cdot\|$ is used to denote Euclidean length on $\mathbb{R}^d$.

## III. Applications

Here we present several representative applications of Theorem 1.

### A. Generalized Costa's Entropy Power Inequality

Costa's EPI [13] (see [33]–[35] for alternate proofs) states that, for independent $d$-dimensional random vectors $X \sim P_X$, $W \sim N(0, \Sigma)$ and $|\alpha| \leq 1$

$$e^{\frac{2}{d}h(X+\alpha W)} \geq (1-\alpha^2)e^{\frac{2}{d}h(X)} + \alpha^2 e^{\frac{2}{d}h(X+W)}. \tag{7}$$

This result was generalized to matrix weightings by Liu, Liu, Poor and Shamai using perturbation and I-MMSE arguments [14]. We demonstrate below that this generalization follows as an easy corollary to Theorem 1 by taking $V$ equal to $X$ contaminated by additive Gaussian noise. In this sense, Theorem 1 may be interpreted as a further generalization of Costa's EPI, where the additive noise is non-Gaussian.

*Theorem 2 [14]: Let $X \sim P_X$ and $W \sim N(0, \Sigma)$ be independent random vectors in $\mathbb{R}^d$. For a positive semidefinite matrix $A \preceq I$ that commutes with $\Sigma$,*

$$e^{\frac{2}{d}h(X+A^{1/2}W)} \geq |I-A|^{1/d}e^{\frac{2}{d}h(X)} + |A|^{1/d}e^{\frac{2}{d}h(X+W)}, \tag{8}$$

*where $|\cdot|$ denotes determinant.*

*Remark 1: The original claim by Liu, Liu, Poor and Shamai in [14] does not contain the hypothesis that $A$ and $\Sigma$ commute. However, the inequality can fail if this commutativity property does not hold [36].*

*Proof:* We assume $\Sigma$ is positive definite; the general case follows by approximation. Let $W_1, W_2$ denote two independent copies of $W$, and put $Y = X + A^{1/2}W_1$ and $V = Y + (I-A)^{1/2}W_2$. Using the fact that $A$ and $\Sigma$ commute, it holds that $V = X + W$ in distribution so that $I(X;V) = h(X+W) - h(W)$. Similarly, $I(Y;V) = h(X+W) - h((I-A)^{1/2}W)$. Now, (8) follows from Theorem 1 since

$$\begin{aligned} &e^{\frac{2}{d}(h(X+A^{1/2}W)-h(X+W)+h(W))} \\ &\quad= e^{\frac{2}{d}(h(Y)-I(X;V))} \\ &\quad\geq e^{\frac{2}{d}(h(X)-I(Y;V))} + e^{\frac{2}{d}h(A^{1/2}W_1)} \\ &\quad= e^{\frac{2}{d}(h(X)-h(X+W)+h((I-A)^{1/2}W))} + |A|^{1/d}e^{\frac{2}{d}h(W)} \\ &\quad= |I-A|^{1/d}e^{\frac{2}{d}(h(X)-h(X+W)+h(W))} + |A|^{1/d}e^{\frac{2}{d}h(W)}. \end{aligned}$$

Multiplying both sides by $e^{\frac{2}{d}(h(X+W)-h(W))}$ completes the proof. □

Costa's EPI may be interpreted as a concavity property enjoyed by entropy powers. The proof of Theorem 2 suggests a generalization of this property to non-Gaussian additive noise. Indeed, we have the following general result:

*Theorem 3: Let $X \sim P_X, Y \sim P_Y$ and $W \sim N(0, \Sigma)$ be independent random vectors in $\mathbb{R}^d$ with finite second moments. Then*

$$e^{\frac{2}{d}(h(X+W)+h(Y+W))} \geq e^{\frac{2}{d}(h(X)+h(Y))} + e^{\frac{2}{d}(h(X+Y+W)+h(W))}.$$

*Proof:* This is an immediate consequence of Theorem 1 by letting $V = X + Y + W$ and rearranging exponents. □

We remark that the assumption of finite second moments in Theorem 3 is only needed to ensure that all entropies are less than $+\infty$ (precluding the indeterminate $\infty - \infty$ in the exponent) and can generally be relaxed. If any of the entropies are equal to $-\infty$, the claim is trivial.

We also mention that Madiman observed the following related inequality on submodularity of differential entropy [37], which can be proved via data processing: if $X, Y, W$ are independent (one-dimensional) random variables, then

$$e^{2(h(X+W)+h(Y+W))} \geq e^{2(h(X+Y+W)+h(W))}. \tag{9}$$

When $W$ is Gaussian, Theorem 3 sharpens inequality (9) by reducing the LHS by a factor of $e^{2(h(X)+h(Y))}$.

### B. Reverse Convolution Inequalities and Refinement of Stam-Gross LSI

Theorem 3 admits several interesting corollaries that are connected to reversing the convolution inequalities for entropy and Fisher information. We explore these connections briefly in this section, but remark that our treatment is far from exhaustive.[1] To start, define the entropy power $N(X)$ and the Fisher information $J(X)$ of a random vector $X$ in $\mathbb{R}^d$ with density $f$ (with respect to Lebesgue measure) as follows:

$$N(X) := \frac{1}{2\pi e}e^{\frac{2}{d}h(X)} \quad J(X) := \mathbb{E}\left[\frac{\|\nabla f(X)\|^2}{f^2(X)}\right].$$

If the Fisher information does not exist (e.g., if $f$ is not sufficiently smooth), we adopt the convention $J(X) = \infty$.

[1] A more comprehensive treatment of this topic and the connections to Costa's EPI can be found in [35].

In his classic 1959 proof of the entropy power inequality [3], Stam observed the following uncertainty principle for $d$-dimensional $X$ with finite second moments:

$$\mathsf{p}(X) := \tfrac{1}{d} N(X) J(X) \geq 1. \tag{10}$$

As noted by Costa and Cover [38], $\mathsf{p}(X)$ may be interpreted as a notion of surface area associated to $X$; indeed, (10) is derived from the EPI in the same way that the isoperimetric inequality is derived from the Brunn-Minkowski inequality. Since (10) is proved using de Bruijn's identity and the special case of Shannon's EPI when one summand is Gaussian, Theorem 3 naturally leads to a sharpening of (10). This strengthening takes the form of a reverse EPI, which upper bounds $N(X+Y)$ in terms of only marginal entropies and Fisher informations.

*Theorem 4: If $X$ and $Y$ are independent $d$-dimensional random vectors with finite second moments, then*

$$N(X)N(Y)\left(J(X)+J(Y)\right) \geq dN(X+Y). \tag{11}$$

*Proof:* We may assume $J(X) < \infty$ and $J(Y) < \infty$, else there is nothing to prove. To begin, let $Z \sim N(0, I)$ be independent of $X, Y$ and recall de Bruijn's identity [3]: $\frac{d}{dt} h(X + \sqrt{t}Z) = \frac{1}{2} J(X + \sqrt{t}Z)$. In particular, we have

$$\frac{d}{dt} N(X+\sqrt{t}Z)\Big|_{t=0} = \frac{1}{d} N(X) J(X). \tag{12}$$

Identifying $W = \sqrt{t}Z$ in Theorem 3 and rearranging, we find

$$\frac{N(X+\sqrt{t}Z)N(Y+\sqrt{t}Z) - N(X)N(Y)}{t} \geq N(X+Y+\sqrt{t}Z) \geq N(X+Y).$$

Letting $t \to 0$ and applying (12) proves the claim. □

By recalling the definition of $\mathsf{p}(\cdot)$ in (10), we obtain the following corollary, which reverses the convolution inequalities for entropy powers and Fisher information:

*Corollary 1: Let $X, Y$ be independent with finite second moments, and choose $\theta$ to satisfy $\theta/(1-\theta) = N(Y)/N(X)$. Then*

$$N(X+Y) \leq \left(N(X)+N(Y)\right)\left(\theta \mathsf{p}(X) + (1-\theta)\mathsf{p}(Y)\right) \tag{13}$$

*and*

$$\frac{1}{J(X+Y)} \leq \left(\frac{1}{J(X)} + \frac{1}{J(Y)}\right) \mathsf{p}(X)\mathsf{p}(Y). \tag{14}$$

*Proof:* Inequality (13) is the same as (11), but rewritten in terms of $\mathsf{p}(\cdot)$. Likewise, (14) follows immediately from (11) and (10), applied to the sum $X+Y$. □

We remind the reader that the convolution inequalities for entropy power and Fisher information may, respectively, be written as

$$N(X+Y) \geq \left(N(X)+N(Y)\right)$$

and

$$\frac{1}{J(X+Y)} \geq \frac{1}{J(X)} + \frac{1}{J(Y)}.$$

So, we may conclude sharpness of the "reverse" estimates in Corollary 1. Precisely, we see that if both $X$ and $Y$ each nearly saturate (10), then the convolution inequalities for entropy power and Fisher information will also be nearly saturated. Notably, the reverse EPI (13) is nontrivial whenever the Fisher informations $J(X)$, $J(Y)$ exist and are finite. This should be contrasted with the reverse EPI of Bobkov and Madiman [39], which holds only for convex measures and must be stated in terms of volume preserving maps (similar to Milman's reverse Brunn-Minkowski inequality for convex bodies [40]). The reader is referred again to the recent paper [12] which surveys known reverse EPIs, all of which apparently require convexity properties of the involved measures.

We have already seen above that Theorem 3 is a natural generalization of Costa's EPI. However, we note that its particularization to (13) continues to imply Costa's EPI. Indeed, if $Z \sim N(0, I)$, then $N(\sqrt{t}Z) = t$, $\mathsf{p}(\sqrt{t}Z) = 1$ and de Bruijn's identity (12) together with (13) yields

$$N(X+\sqrt{t}Z) \leq N(X) + t\left(\frac{d}{dt} N(X+\sqrt{t}Z)\Big|_{t=0}\right), \tag{15}$$

which is equivalent to concavity of $t \mapsto N(X+\sqrt{t}Z)$.

In addition to generalizing Costa's inequality, (13) also improves the uncertainty principle (10). To see this, let $X, X_*$ be i.i.d. random vectors on $\mathbb{R}^d$. Then, it follows immediately from (13) that

$$\mathsf{p}(X) \geq \exp\left[\frac{2}{d}\left(h\left(\frac{1}{\sqrt{2}}(X+X_*)\right) - h(X)\right)\right]. \tag{16}$$

The RHS of (16) is strictly greater than one by the EPI, unless $X$ is Gaussian. The quantity $h\left(\frac{1}{\sqrt{2}}(X+X_*)\right) - h(X)$ is referred to as the *entropy jump* associated to $X$, and can be lower bounded by a linear function of the relative entropy $D(X\|Z)$ when $\mathrm{Cov}(X) = I$ and $X$ satisfies regularity conditions (i.e., $X$ satisfies a Poincaré inequality [41]–[43] and, for $d \geq 2$, has log-concave density [44]).

In closing this section, we briefly discuss how Theorem 4 leads to a sharpening of the Gaussian logarithmic Sobolev inequality (LSI). In 1975, Gross rediscovered (10) in a different form by establishing the LSI for the standard Gaussian measure $\gamma$ on $\mathbb{R}^d$ [29]. In particular, he showed that for every $g$ on $\mathbb{R}^d$ with gradient in $L^2(\gamma)$,

$$\int_{\mathbb{R}^d} g^2 \log g^2 \, d\gamma \leq 2\int_{\mathbb{R}^d} \|\nabla g\|^2 \, d\gamma + \left(\int_{\mathbb{R}^d} g^2 d\gamma\right) \log\left(\int_{\mathbb{R}^d} g^2 \, d\gamma\right). \tag{17}$$

In the same paper, Gross also proved that (17) is equivalent to the hypercontractivity of the Ornstein-Uhlenbeck semigroup [45]. Despite the fact that Stam's inequality (10) preceded Gross' LSI (17) by over a decade, it wasn't until the 1990's that Carlen [46] recognized that they were completely equivalent (a concise proof can be found in [47]). By viewing $g^2$ as the probability density (with respect to the standard Gaussian measure $\gamma$) associated to a random vector $X$, it is well known that the LSI (17) may be equivalently written in information-theoretic terms as

$$\frac{1}{2} I(X\|Z) \geq D(X\|Z), \tag{18}$$

where $I(X\|Z)$ denotes the relative Fisher information of $X$ with respect to $Z \sim N(0, I)$, and $D(X\|Z)$ is the relative entropy (with units of *nats*). There have been several recent works that attempt to give quantitative bounds on the deficit $\delta_{\mathsf{LSI}}(X) := \frac{1}{2}I(X\|Z) - D(X\|Z)$ (e.g., [48]–[50]). By recalling the expressions for relative entropy and Fisher information in terms of their non-relative counterparts

$$D(X\|Z) = \frac{d}{2}\log(2\pi e) - h(X) + \frac{1}{2}\mathbb{E}\|X\|^2 - \frac{d}{2}$$
$$I(X\|Z) = J(X) - 2d + \mathbb{E}\|X\|^2,$$

we may take logarithms in (16) and use the inequality $\log x \leq x - 1$ to obtain the following concise bound on simplification:

*Theorem 5: Let $X, X_*$ be i.i.d. random vectors on $\mathbb{R}^d$ with finite second moments. For $\delta_{\mathsf{LSI}}$ defined as above,*

$$\delta_{\mathsf{LSI}}(X) \geq h\left(\tfrac{1}{\sqrt{2}}(X + X_*)\right) - h(X). \tag{19}$$

Thus, we see that $\delta_{\mathsf{LSI}}(X)$ controls the entropy jump associated to $X$. Note that inequality (19) does not impose regularity conditions on $X$ (beyond finiteness of second moment). This should be contrasted with the quantitative bound in [49], which assumes that $X$ satisfies a Poincaré inequality. Inequality (19) also has the desirable property that it is dimension-free in the sense that it is additive on independent components of $X$. The disadvantage of (19) is that it is presently unknown how the entropy jump associated to $X$ is quantitatively related to the distance of $X$ from Gaussian, except under regularity assumptions [9], [41], [44]. This situation is similar to the HSI inequality of Ledoux, Nourdin and Peccati which strengthens (18), provided the so-called Stein discrepancy associated to $X$ is finite. As with entropy jumps, finiteness of Stein discrepancy is presently only ensured under regularity assumptions such as positive spectral gap [51].

### C. Conditional Strong EPI and Converse for the Two-Encoder Quadratic Gaussian Source Coding Problem

A conditional version of the EPI is often useful in applications (e.g., [52]). Theorem 1 easily generalizes along these lines. In particular, due to joint convexity of $(u, v) \mapsto \log(e^u + e^v)$ in $(u, v)$, we immediately obtain via Jensen's inequality the following corollary:

*Corollary 2: Suppose $X, W$ are random vectors in $\mathbb{R}^d$, conditionally independent given $Q$, and moreover that $W$ is conditionally Gaussian given $Q$. Define $Y = X + W$. For any $V$ satisfying $X \to Y \to V|Q$,*

$$e^{\frac{2}{d}(h(Y|Q) - I(X;V|Q))} \geq e^{\frac{2}{d}(h(X|Q) - I(Y;V|Q))} + e^{\frac{2}{d}h(W|Q)}. \tag{20}$$

Let's now see an example of how this may be applied to establish nontrivial converse results in network information theory. Toward this end, let us mention that characterizing the rate region for the two-encoder quadratic Gaussian source coding problem was a longstanding open problem until its ultimate resolution by Wagner, Tavildar and Viswanath in their tour de force [18], which established that a separation-based scheme [53], [54] was optimal. Wagner *et al.*'s work built upon Oohama's earlier solution to the one-helper problem [19] and the independent solutions to the Gaussian CEO problem due to Prabhakaran, Tse and Ramachandran [28] and Oohama [27] (see [52] for a self-contained treatment). Since Wagner *et al.*'s original proof of the sum-rate constraint, other proofs have been proposed based on estimation-theoretic arguments and semidefinite programming (e.g., [55]), however all known proofs are quite complex. Below, we show that the converse result for the entire rate region is a direct consequence of Corollary 2, thus unifying the results of [18] and [19] under a common and succinct inequality.

*Theorem 6 [18]: Let $\rho \in [-1, 1]$ and let $X \sim N(0, I)$ be a $d$-dimensional standard Gaussian random vector. Define $Y = \rho X + \sqrt{1-\rho^2}Z$, where $Z \sim N(0, I)$ is independent of $X$. Let $\phi_X : \mathbb{R}^d \to \{1, \ldots, 2^{dR_X}\}$ and $\phi_Y : \mathbb{R}^d \to \{1, \ldots, 2^{dR_Y}\}$, and define*

$$d_X := \frac{1}{d}\mathbb{E}\left[\|X - \mathbb{E}[X|\phi_X(X), \phi_Y(Y)]\|^2\right]$$
$$d_Y := \frac{1}{d}\mathbb{E}\left[\|Y - \mathbb{E}[Y|\phi_X(X), \phi_Y(Y)]\|^2\right].$$

*Then*

$$R_X \geq \frac{1}{2}\log_2\left(\frac{1}{d_X}\left(1 - \rho^2 + \rho^2 2^{-2R_Y}\right)\right) \tag{21}$$
$$R_Y \geq \frac{1}{2}\log_2\left(\frac{1}{d_Y}\left(1 - \rho^2 + \rho^2 2^{-2R_X}\right)\right) \tag{22}$$
$$R_X + R_Y \geq \frac{1}{2}\log_2\frac{(1-\rho^2)\beta(d_X d_Y)}{2 d_X d_Y}, \tag{23}$$

*where $\beta(D) := 1 + \sqrt{1 + \frac{4\rho^2 D}{(1-\rho^2)^2}}$.*

For convenience, we assume throughout the remainder of this subsection that all information quantities have units of *bits* (i.e., their defining logarithms are taken to be base-2).

The key ingredient for establishing Theorem 6 is the following simple consequence of Corollary 2:

*Proposition 1: For $X, Y$ as above and any $U, V$ satisfying $U \to X \to Y \to V$,*

$$2^{-\frac{2}{d}(I(X;U,V) + I(Y;U,V))} \geq 2^{-\frac{2}{d}I(X,Y;U,V)}\left(1 - \rho^2 + \rho^2 2^{-\frac{2}{d}I(X,Y;U,V)}\right). \tag{24}$$

*Proof:* By identifying $Q \leftarrow U$, an application of Corollary 2 directly yields

$$\begin{aligned} 2^{\frac{2}{d}(h(Y|U) - I(X;V|U))} &\geq 2^{\frac{2}{d}(h(\rho X|U) - I(Y;V|U))} + 2^{\frac{2}{d}h(\sqrt{1-\rho^2}Z)} \\ &= \rho^2 2^{\frac{2}{d}(h(X|U) - I(Y;V|U))} + (2\pi e)(1-\rho^2). \end{aligned}$$

Since $2^{-\frac{2}{d}h(Y)} = 2^{-\frac{2}{d}h(X)} = \frac{1}{2\pi e}$, multiplying through by $\frac{1}{2\pi e}2^{-\frac{2}{d}I(X,Y;U,V)}$ and rearranging exponents establishes the claim. □

*Proof of Theorem 6:* Put $U = \phi_X(X)$ and $V = \phi_Y(Y)$. The left- and right-hand sides of (24) are monotone decreasing in $\frac{1}{d}(I(X;U,V) + I(Y;U,V))$ and $\frac{1}{d}I(X,Y;U,V)$, respectively. Therefore, if

$$\frac{1}{d}(I(X;U,V) + I(Y;U,V)) \geq \frac{1}{2}\log_2\frac{1}{D}$$

and

$$\frac{1}{d}I(X,Y;U,V) \le R$$

for some pair $(R, D)$, then we have

$$D \ge 2^{-2\ R}\left(1-\rho^2+\rho^2 2^{-2\ R}\right),$$

which is a quadratic inequality with respect to the term $2^{-2\ R}$. This is easily solved using the quadratic formula to obtain:

$$2^{-2R} \le \frac{2D}{(1-\rho^2)\beta(D)} \Longrightarrow R \ge \frac{1}{2}\log_2\frac{(1-\rho^2)\beta(D)}{2D}.$$

Jensen's inequality and the maximum-entropy property of Gaussians imply

$$\frac{1}{d}I(X;U,V) \ge \frac{1}{2}\log_2\frac{1}{d_X}$$

and

$$\frac{1}{d}I(Y;U,V) \ge \frac{1}{2}\log_2\frac{1}{d_Y},$$

so that

$$\frac{1}{d}(I(X;U,V)+I(Y;U,V)) \ge \frac{1}{2}\log_2\frac{1}{d_X d_Y},$$

establishing (23) since

$$\frac{1}{d}I(X,Y;U,V) \le \frac{1}{d}\left(H(U)+H(V)\right) \le R_X + R_Y.$$

Similarly,

$$\begin{aligned} 2^{2\ R_X+\log_2\ d_X} &\ge 2^{\frac{2}{d}(I(X;U|V)-I(X;U,V))} \\ &= 2^{-\frac{2}{d}I(X;V)} \\ &\ge (1-\rho^2)+\rho^2 2^{-\frac{2}{d}I(Y;V)} \\ &\ge (1-\rho^2)+\rho^2 2^{-2\ R_Y}, \end{aligned}$$

where the second inequality is a consequence of Theorem 1 and the fact that $h(X) = h(Y) = h(Z)$ for the variables as defined. Rearranging (and symmetry) yields (21)-(22). □

We remark that Proposition 1 (a special case of Corollary 2) first appeared in [56] by the author and Jiao. In fact, Proposition 1 yields a stronger result than the converse for the two-terminal Gaussian source coding problem; it shows that the rate regions coincide for the problems when distortion is measured under quadratic loss and logarithmic loss [57], [58]. On a different note, we remark that the quadratic Gaussian sum-rate bound has an alternate proof that avoids using the EPI, favoring estimation-theoretic arguments instead [55].

### D. One-Sided Gaussian Interference Channel

The one-sided Gaussian interference channel (IC) (or Z-Gaussian IC) is a discrete memoryless channel, with input-output relationship given by

$$Y_1 = X_1 + W \tag{25}$$

$$Y_2 = \alpha Y_1 + X_2 + W_2, \tag{26}$$

where $X_i$ and $Y_i$ are the channel inputs and observations corresponding to Encoder $i$ and Decoder $i$, respectively, for $i = 1, 2$. Here, $W \sim N(0,1)$ and $W_2 \sim N(0, 1-\alpha^2)$ are independent of each other and of the channel inputs $X_1, X_2$. We have assumed $|\alpha| < 1$ since the setting where $|\alpha| \ge 1$ is referred to as the *strong interference* regime, and the capacity is known to coincide with the Han-Kobayashi inner bound [15], [52], [59], [60]. Observe that we have expressed the one-sided Gaussian IC in *degraded form*, which has capacity region identical to the corresponding non-degraded version as proved by Costa [15]. Despite receiving significant attention from researchers over several decades, the capacity region $\mathscr{C}(\alpha, P_1, P_2)$ of the one-sided Gaussian IC remains unknown in the regime of $|\alpha| < 1$ described above.

Having already discussed connections between Costa's EPI (7) and Theorem 1 above, we remark that Costa's EPI was motivated by the Gaussian IC [15]. Since Theorem 1 generalizes Costa's result, the one-sided Gaussian IC presents itself as a natural application. Toward this end, we establish a new multi-letter outer bound to give a simple demonstration of how Theorem 1 might be applied to the one-sided Gaussian IC.

*Theorem 7: Let all information quantities have units of bits.* $(R_1, R_2) \in \mathscr{C}(\alpha, P_1, P_2)$ *only if*

$$R_1 \le \frac{1}{2}\log_2(1+P_1) \tag{27}$$

$$R_2 \le \frac{1}{2}\log_2(1+P_2) \tag{28}$$

*and*

$$\begin{aligned} 2^{-2R_2+\epsilon_n} \ge 2^{-\frac{2}{n}I(X_1^n,X_2^n;Y_2^n)} \sup_{V:Y_1^n\to Y_0^n\to V} \Big\{ &\alpha^2 2^{2R_1-\frac{2}{n}I(Y_0^n;V|Y_1^n)} \\ &+(1-\alpha^2)2^{\frac{2}{n}I(Y_1^n;V)}\Big\}, \end{aligned} \tag{29}$$

*for some* $\epsilon_n \to 0$ *and independent* $X_1^n, X_2^n$ *satisfying the power constraints* $\mathbb{E}[\|X_i^n\|^2] \le nP_i,\ i = 1, 2$.

*Proof:* For consistency, we assume throughout the proof that information quantities have units of bits. The only non-trivial inequality to prove is (29). Thus, we begin by noting that Corollary 2 implies

$$\begin{aligned} &2^{\frac{2}{n}(h(Y_2^n|X_2^n)-I(Y_1^n;V|X_2^n))} \\ &\quad\ge 2^{\frac{2}{n}(h(\alpha Y_1^n|X_2^n)-I(Y_2^n;V|X_2^n))} + 2^{\frac{2}{n}h(W_2^n|X_2^n)} \\ &\quad= \alpha^2 2^{\frac{2}{n}(h(Y_1^n)-I(Y_2^n;V|X_2^n))} + (1-\alpha^2)2^{\frac{2}{n}h(W^n)} \end{aligned}$$

for all $V$ such that $Y_1^n \to Y_2^n \to V|X_2^n$. Since $h(W^n) = h(Y_2^n|X_1^n, X_2^n) = h(Y_1^n|X_1^n)$, $I(Y_2^n; V|X_2^n) = I(Y_0^n; V, X_2^n)$ and $I(Y_1^n; V|X_2^n) = I(Y_1^n; V, X_2^n)$, this can be rewritten as

$$\begin{aligned} &2^{-\frac{2}{n}I(X_2^n;Y_2^n)+\frac{2}{n}I(X_1^n,X_2^n;Y_2^n)} \\ &\quad\ge \sup_{V:Y_1^n\to Y_0^n\to V} \Big\{\alpha^2 2^{\frac{2}{n}I(X_1^n;Y_1^n)-\frac{2}{n}I(Y_0^n;V|Y_1^n)} \\ &\qquad\qquad + (1-\alpha^2)2^{\frac{2}{n}I(Y_1^n;V)}\Big\}. \end{aligned}$$

Therefore,

$$\begin{aligned} &2^{-2(R_2-\epsilon_n)} \\ &\quad\ge 2^{-\frac{2}{n}I(X_2^n;Y_2^n)} \qquad (30) \\ &\quad\ge 2^{-\frac{2}{n}I(X_1^n,X_2^n;Y_2^n)} \sup_{V:Y_1^n\to Y_0^n\to V} \Big\{\alpha^2 2^{\frac{2}{n}I(X_1^n;Y_1^n)-\frac{2}{n}I(Y_0^n;V|Y_1^n)} \\ &\qquad + (1-\alpha^2)2^{\frac{2}{n}I(Y_1^n;V)}\Big\} \qquad (31) \end{aligned}$$

$$\geq 2^{-\frac{2}{n}I(X_1^n,X_2^n;Y_2^n)} \sup_{V:Y_1^n\to Y_0^n\to V} \Big\{\alpha^2 2^{2(R_1-\epsilon_n)-\frac{2}{n}I(Y_0^n;V|Y_1^n)} + (1-\alpha^2)2^{\frac{2}{n}I(Y_1^n;V)}\Big\}, \quad (32)$$

where (30) and (32) hold for $\epsilon_n \to 0$ due to Fano's inequality. Multiplying both sides by $2^{2\epsilon_n}$ proves the claim. □

The Han-Kobayashi achievable region [52], [60] evaluated for Gaussian inputs (without power control) can be expressed as the set of rate pairs $(R_1, R_2)$ satisfying (27), (28) and

$$2^{-2R_2} \geq \frac{\alpha^2 P_2 2^{2R_1}}{(P_2+1-\alpha^2)(1+\alpha^2 P_1+P_2)} + \frac{1-\alpha^2}{P_2+1-\alpha^2}. \quad (33)$$

Interestingly, this takes a similar form to (29); however, it is known that transmission without power control is suboptimal for the Gaussian Z-interference channel in general [61], [62]. Nevertheless, it may be possible to identify a random variable $V$ in (29), possibly depending on $X_2^n$, which ultimately improves known bounds. We leave this for future work.

### E. Relationship to Strong Data Processing

Strong data processing inequalities and their connection to hypercontractivity have garnered much attention recently (e.g,. [6], [7], [63]–[68]). We have already seen the connection to Theorem 1 in the discussion surrounding (5). We briefly mention here that, for $\Psi(t) \equiv \Psi(t, P_{XY})$ as defined in (4), inequality (5) may be rearranged to provide the following explicit upper bound on the strong data processing function $\Psi$ which may be of independent interest:

*Corollary 3: Let $X \sim P_X$ and $W \sim N(0, I)$ be independent random vectors in $\mathbb{R}^d$. For $Y = X + W$,*

$$\Psi(t, P_{XY}) \leq I(X;Y) - \frac{d}{2}\log\left(1 + \frac{1}{2\pi e}e^{\frac{2}{d}(h(X)-t)}\right).$$

*Moreover, equality is achieved if $X$ is Gaussian with covariance proportional to the identity matrix $I$.*

## IV. Outline of Proof of Main Results

The goal of this section is to provide an overview of the key ideas involved in establishing our main results. To this end, we focus attention on the scalar version of Theorem 1, stated as Theorem 8 below. The rationale for this is that the multidimensional result of Theorem 1 will be proved as a corollary of its one-dimensional counterpart in Section V, using some of the machinery that we develop in this section.

*Theorem 8: Let $X$ be random variable on $\mathbb{R}$, and let $W \sim N(0, \sigma^2)$ be independent of $X$. For $Y = X + W$ and any $V$ satisfying $X \to Y \to V$,*

$$e^{2(h(Y)-I(X;V))} \geq e^{2(h(X)-I(Y;V))} + e^{2h(W)}. \quad (34)$$

In order to give the intuition behind our proof of (34), we begin by defining the following quantities whose motivation will soon be apparent:

*Definition 1: Let $X \sim P_X$ be a random variable on $\mathbb{R}$. For $(Y, X) \sim \mathbb{G}^{\varrho}_{Y|X} P_X$, define the family of functionals (of $P_X$)*

$$\mathsf{s}_\lambda(X,\varrho) := -h(X) + \lambda h(Y) + \inf_{V:X\to Y\to V}\Big\{I(Y;V) - \lambda I(X;V)\Big\}$$

*parameterized by $\lambda \geq 1$ and $\varrho > 0$. Similarly, let $(X, Q) \sim P_{XQ}$ be a pair of random variables on $\mathbb{R}\times\mathcal{Q}$. For $(Y, X, Q) \sim \mathbb{G}^{Q}_{Y|X} P_{XQ}$, define the functional (of $P_{XQ}$)*

$$\mathsf{s}_\lambda(X,\varrho|Q) := -h(X|Q) + \lambda h(Y|Q) + \inf_{V:X\to Y\to V|Q}\Big\{I(Y;V|Q) - \lambda I(X;V|Q)\Big\}.$$

Noting that the term $e^{2\,h(W)}$ is constant in (34), one approach toward proving Theorem 8 would be to simultaneously minimize the exponent $h(Y) - I(X; V)$, while maximizing the exponent $h(X) - I(Y; V)$ over all valid choices of $X, V$. If, for all such choices, the LHS of (34) exceeds the RHS of (34), the theorem would be proved. Modulo rescaling of random variables, the functional $\mathsf{s}_\lambda(X, \varrho)$ makes this intuition precise by capturing this tradeoff of exponents, as parameterized by $\lambda$. This is apparent by observing that

$$\mathsf{s}_\lambda(X,1) = \inf_{V:X\to Y\to V}\Big\{\lambda(h(Y)-I(X;V)) - (h(X)-I(Y;V))\Big\},$$

where $X, Y$ are as in the statement of Theorem 8 for $\sigma^2 = 1$.

Since $\mathsf{s}_\lambda(X, \varrho)$ has the optimization over valid choices of $V$ built into its definition, the stated approach of extremizing exponents in (34) reduces to characterizing the infimum of $\mathsf{s}_\lambda(X, \varrho)$, taken over all distributions $P_X$ with finite variance (in fact, variance at most 1 suffices). Such an optimization appears difficult to execute directly, so we instead turn to the relaxation

$$\mathsf{V}_\lambda(\varrho) := \inf_{P_{XQ}:\,\mathbb{E}[X^2]\leq 1} \mathsf{s}_\lambda(X,\varrho|Q). \quad (35)$$

This is indeed a relaxation of infimizing $\mathsf{s}_\lambda(X, \varrho)$ over the set of distributions $P_X$ with $\mathbb{E}X^2 \leq 1$ because problem (35) is equivalent to finding the infimum of the lower convex envelope of the functional $P_X \mapsto \mathsf{s}_\lambda(X, \varrho)$, over the set of distributions $P_X$ with $\mathbb{E}X^2 \leq 1$. Similar relaxation strategies have been fruitfully applied elsewhere in information theory (cf. [69]–[71]).

Having described the overall objective, our strategy for characterizing $\mathsf{V}_\lambda(\varrho)$ will be to show that in (35) it suffices to minimize over Gaussian $X$, independent of $Q$. Stated precisely, we aim to show:

*Theorem 9:*

$$\mathsf{V}_\lambda(\varrho) = \inf_{0\leq\gamma\leq 1} \mathsf{s}_\lambda(X_\gamma,\varrho), \quad \textit{where } X_\gamma \sim N(0,\gamma). \quad (36)$$

With Theorem 9 in hand, it is a matter of calculus to explicitly compute $\mathsf{V}_\lambda(\varrho)$ as a function of $\varrho$ and $\lambda$, yielding Theorem 8 as a corollary. Thus, we dedicate the following subsection to provide a heuristic discussion of how Theorem 9 will be established. This is intended to orient the reader and provide context for the proof details that follow in later sections. The computations for deducing Theorem 8 from Theorem 9 are detailed in subsection IV-B.

*Remark 2: A somewhat technical point is that the value of the optimization problem* (35) *potentially depends on the set $\mathcal{Q}$ (in which the random variable $Q$ takes values), which we have not explicitly specified in Definition 1. However, this*

*is not problematic since any non-singleton set $\mathcal{Q}$ will suffice. Indeed, the standard dimensionality reduction argument can be applied: Consider the image of the map $P_{XQ} \mapsto (\mathbb{E}X^2, \mathsf{s}_\lambda(X, \varrho|Q))$, taken over all distributions $P_{XQ}$ on $\mathbb{R}\times\mathcal{Q}$ with $\mathbb{E}X^2 \leq 1$ and, say for concreteness, $\mathcal{Q} = \mathbb{R}$. As usual, this image is a convex set in $\mathbb{R}^2$ by definition. Hence, by the Fenchel-Caratheodory-Bunt theorem [72, Theorem 18$(ii)$], for any point $(a, b)$ in this set, there is a distribution $P_{XQ}$ on $\mathbb{R} \times \mathcal{Q}$, with $P_Q$ supported on at most two points in $\mathcal{Q}$, that achieves the values $\mathbb{E}X^2 = a$ and $\mathsf{s}_\lambda(X, \varrho|Q) = b$. Since we will often consider random variables $Q$ taking values on different spaces, the set $\mathcal{Q}$ will be implicitly defined as the set in which the random variable $Q$ takes values. However, taking $\mathcal{Q} = \{1, 2, 3\}$ is generally sufficient for our purposes as will be made clear in Section VI-B.*

### A. Heuristic Proof of Theorem 9

Having argued that Theorem 9 is the crux, we now give a heuristic discussion of the main ideas that will be carried forward later into the full arguments. In order to minimize interruption to the flow of the argument, remarks on related literature are collected at the conclusion of this subsection.

Establishing sufficiency of Gaussian $X$ in optimization problem (35) appears nontrivial from the outset. However, the first crucial idea is that we may exploit a "doubling" property enjoyed by the functional $\mathsf{s}_\lambda$. In particular, we have the following simple lemma, established in Section VI-A:

*Lemma 1: Let $P_{XQ}$ be a distribution on $\mathbb{R} \times \mathcal{Q}$, let $(Y, X, Q) \sim \mathbb{G}^{\varrho}_{Y|X} P_{XQ}$, and let $(Y_1, X_1, Q_1)$ and $(Y_2, X_2, Q_2)$ denote two independent copies of $(Y, X, Q)$. Define*

$$X_+ = \frac{X_1 + X_2}{\sqrt{2}}, \qquad X_- = \frac{X_1 - X_2}{\sqrt{2}} \tag{37}$$

*and, in a similar manner, define $Y_+, Y_-$. Letting $\mathbf{Q} = (Q_1, Q_2)$, we have for $\lambda \geq 1$*

$$2\mathsf{s}_\lambda(X, \varrho|Q) \geq \mathsf{s}_\lambda(X_+, \varrho|X_-, \mathbf{Q}) + \mathsf{s}_\lambda(X_-, \varrho|Y_+, \mathbf{Q}) \tag{38}$$

*and*

$$2\mathsf{s}_\lambda(X, \varrho|Q) \geq \mathsf{s}_\lambda(X_+, \varrho|Y_-, \mathbf{Q}) + \mathsf{s}_\lambda(X_-, \varrho|X_+, \mathbf{Q}). \tag{39}$$

In view of the central limit theorem, the doubling operation (37) should produce random variables $X_+$, and $X_-$ that are "more Gaussian" than $X$. The essential content of Lemma 1 is that the doubling operation (37) also improves the value of the $\mathsf{s}_\lambda$ functional in the sense of (38) and (39). So, the general idea will be to exploit this property to show that Gaussian $X$ minimizes the functional $\mathsf{s}_\lambda(\cdot, \varrho)$.

Let us now make this intuition precise, while still glossing over some of the challenging technicalities that need to be dealt with. In particular, let us assume that the infimum in (35) is attained, and let $\mathcal{P}$ denote the subset of distributions $P_{XQ}$ achieving the corresponding minimum, with the additional property that $\mathbb{E}X = 0$. The assumption of centered $X$ is only for convenience, since the functional $\mathsf{s}_\lambda(\cdot, \varrho)$ is invariant to translation of the mean of its argument (an immediate consequence of the same translation invariance enjoyed by entropy and mutual information).

Now, suppose there is a distribution $P_{XQ} \in \mathcal{P}$ having the extremal property that, for $(Y, X, Q) \sim \mathbb{G}^{\varrho}_{Y|X} P_{XQ}$ and any other $(Y', X', Q') \sim \mathbb{G}^{\varrho}_{Y'|X'} P_{X'Q'}$ with $P_{X'Q'} \in \mathcal{P}$,

$$h(Y|Q) - h(X|Q) \leq h(Y'|Q') - h(X'|Q'). \tag{40}$$

We will argue below that by choosing $P_{XQ}$ in this manner, the corresponding conditional law $P_{X|\{Q=q\}}$ is Gaussian with variance not depending on $q$.

Toward this end, let $X_+, X_-, Y_+, Y_-, \mathbf{Q}$ be as in Lemma 1, constructed from independent copies of $(Y, X, Q) \sim \mathbb{G}^{\varrho}_{Y|X} P_{XQ}$. Since the transformation $(X_1, X_2) \mapsto (X_+, X_-)$ is variance-preserving, each of the four quantities $\mathsf{s}_\lambda(X_+, \varrho|X_-, \mathbf{Q})$, $\mathsf{s}_\lambda(X_-, \varrho|Y_+, \mathbf{Q})$, $\mathsf{s}_\lambda(X_+, \varrho|Y_-, \mathbf{Q})$ and $\mathsf{s}_\lambda(X_-, \varrho|X_+, \mathbf{Q})$ is at least $\mathsf{V}_\lambda(\varrho)$ by definition (i.e., (35)). However, the assumption that $P_{XQ} \in \mathcal{P}$ combined with Lemma 1 establishes the reverse inequality, giving equality. E.g.,

$$\mathsf{s}_\lambda(X_+, \varrho|Y_-, \mathbf{Q}) = \mathsf{s}_\lambda(X_-, \varrho|Y_+, \mathbf{Q}) = \mathsf{V}_\lambda(\varrho). \tag{41}$$

Thus, roughly speaking, the set $\mathcal{P}$ is closed under the doubling operation. It is tempting at this point to imagine applying the doubling operation ad infinitum, leading to Gaussianity by the central limit theorem. However, this approach runs into the problem that the auxiliary alphabet $\mathcal{Q}$ grows at each stage, leading to technical difficulties down the line. One could potentially control this growth by applying the dimensionality reduction operation described in Remark 2. Unfortunately, it is not clear that such dimensionality reduction will preserve any additional "Gaussianity" gained through doubling. The need to circumvent this problem is what motivates the additional extremal property (40) which we assumed of $P_{XQ}$ and have not yet exploited.

To see how the argument works, let $B \sim \text{Bernoulli}(1/2)$ be a Bernoulli random variable taking values on $\{+, -\}$, independent of $X_+, X_-, Y_+, Y_-, \mathbf{Q}$. Define $\bar{B}$ to be the complement of $B$ in $\{+, -\}$, meaning that $\{B = +\} \Leftrightarrow \{\bar{B} = -\}$, and vice-versa. Now, let us define a new pair of random variables $(X', Q')$ as follows: $X' = X_B$ and $Q' = (B, Y_{\bar{B}}, \mathbf{Q})$. By construction and (41),

$$\begin{aligned}\mathsf{s}_\lambda(X', \varrho|Q') &= \frac{1}{2}\mathsf{s}_\lambda(X_+, \varrho|Y_-, \mathbf{Q}) + \frac{1}{2}\mathsf{s}_\lambda(X_-, \varrho|Y_+, \mathbf{Q}) \\ &= \mathsf{V}_\lambda(\varrho).\end{aligned}$$

We also have $\mathbb{E}X'^2 = \mathbb{E}X^2$, so that $P_{X'Q'} \in \mathcal{P}$, where $P_{X'Q'}$ denotes the law of $(X', Q')$. Unraveling definitions, the following crucial identity may be obtained:

$$\begin{aligned}h(Y|Q) - h(X|Q) = h(Y'|Q') - h(X'|Q') \\ + \frac{1}{2} I(X_+; X_-|Y_+, Y_-, \mathbf{Q}).\end{aligned} \tag{42}$$

Therefore, assumption (40) together with non-negativity of mutual information implies that we must have

$$\begin{aligned}I(X_1 + X_2; X_1 - X_2|Y_1, Y_2, Q_1, Q_2) \\ = I(X_+; X_-|Y_+, Y_-, \mathbf{Q}) = 0.\end{aligned} \tag{43}$$

Let us now take a final leap of faith, and pretend for sake of simplicity that $Y_1, Y_2$ are absent in the conditioning

in (43). This would imply that $I(X_+; X_-|Q_1, Q_2) = 0$. Equivalently, $X_+|\{Q_1, Q_2 = q_1, q_2\}$ and $X_-|\{Q_1, Q_2 = q_1, q_2\}$ are independent for $P_Q \times P_Q$-a.e. $q_1, q_2$. However, $X_1|\{Q_1, Q_2 = q_1, q_2\}$ and $X_2|\{Q_1, Q_2 = q_1, q_2\}$ are independent by construction. At this point, a classical characterization of the normal law due to Bernstein [73] is helpful:

*Lemma 2 [74, Theorem 5.1.1]: If $A_1, A_2$ are independent random variables such that $A_1 + A_2$ and $A_1 - A_2$ are independent, then $A_1$ and $A_2$ are normal, with identical variances.*

Since $X_i|\{Q_1, Q_2 = q_1, q_2\}$ is equal in distribution to $X_i|\{Q_i = q_i\}$, for $i = 1, 2$ (by independence of the copies $(X_1, Q_1)$, $(X_2, Q_2)$), Bernstein's theorem allows us to conclude from the above that $X_1|\{Q_1 = q_1\}$ and $X_2|\{Q_2 = q_2\}$ are normal with identical variances. Moreover, freedom of choosing $q_1$ and $q_2$ lets us further surmise that these variances do not depend on the specific choice of $q_1, q_2$. So, we are left to conclude that $X|\{Q = q\} \sim N(\mu_q, \sigma_X^2)$, where $\sigma_X^2$ does not depend on $q$, but the mean $\mu_q$ may. However, since the functional $\mathsf{s}_\lambda(\cdot, \varrho)$ is invariant to translations of the mean, we arrive at the statement of Theorem 9.

The above heuristic argument provides a roadmap for the complete proof that follows. In order to make the proof rigorous, there are three main technical issues to be dealt with: First, it is not clear a priori that the infimum in (35) is achieved. We remedy this by adapting the above argument to work for distributions which closely approach the infimum in (35), and are *nearly* extremal in the sense of (40). This allows us to infer the existence of a sequence of distributions which approach the infimum (35), but also have mutual informations of the form (43) that asymptotically vanish. Second, the application of Bernstein's theorem above relied on equality in (43), and also neglected the conditioning on the variables $Y_1, Y_2$. To correct these issues, we develop an information-theoretic variation of Bernstein's theorem that enables us to conclude that the aforementioned asymptotically optimal sequence of distributions converges weakly to Gaussian. Finally, to pass from this weakly convergent sequence to Theorem 9, we need to establish a local semicontinuity property enjoyed by the functional $P_X \mapsto \mathsf{s}_\lambda(X, \varrho)$. Generally speaking, these three points are separately dealt with in Section VI, and Appendices A and B, respectively.

*Literature Notes:* At this point, we would like to make a few remarks on literature relevant to the above proof. Although considerably different in the details, the doubling argument presented above draws inspiration from Geng and Nair's exposition in [70], which established Gaussian optimality for different information functionals by appealing to rotational invariance of the optimizers. They refer to their proof technique as "factorization of concave envelopes". Another argument appealing to rotational invariance appeared in [75], providing an alternate proof to the fact that Gaussian inputs maximize mutual information in Gaussian channels. In addition to these relatively recent appearances in the information theory literature, this general strategy has been successfully employed for establishing extremality of Gaussian kernels in functional inequalities [46], [76], [77]. Our reading of the literature suggests that Lieb [76] deserves credit for inventing the general approach; indeed, Carlen's contemporaneous work [46] attributes the idea to Lieb and coined the "doubling trick" terminology. The fact that this doubling trick works well for establishing Gaussian optimality in both information-theoretic and functional inequalities is not coincidental. Indeed, there is a precise duality between these different classes of inequalities; interested readers are referred to [78]–[81] for further discussion.

In a different direction, we remark that the usual statement of Bernstein's theorem does not comment on the identical variances of $A_1, A_2$ as stated in Lemma 2. However, assuming without loss of generality that $A_1, A_2$ are zero-mean, the observation that $A_1$ and $A_2$ have identical variances is immediate since $\mathbb{E}[A_1^2] - \mathbb{E}[A_2^2] = \mathbb{E}[(A_1 - A_2)(A_1 + A_2)] = 0$. This fact was explicitly noted by Geng and Nair [70].

### B. From Theorem 9 to Theorem 8

Taking Theorem 9 for granted, we now detail the computations needed to derive Theorem 8 from it. To begin, we establish the following explicit characterization of $\mathsf{V}_\lambda(\varrho)$.

*Theorem 10:*

$$\mathsf{V}_\lambda(\varrho) = \begin{cases} \frac{1}{2}\left[\lambda \log\left(\frac{\lambda 2\pi e}{\lambda - 1}\right) - \log\left(\frac{2\pi e}{\lambda - 1}\right) + \log(\varrho)\right] & \text{if } \varrho \geq \frac{1}{\lambda - 1} \\ \frac{1}{2}\left[\lambda \log\left(2\pi e(1 + \varrho)\right) - \log\left(2\pi e\right)\right] & \text{if } \varrho \leq \frac{1}{\lambda - 1}. \end{cases}$$

We require the following lemma, which is a simple consequence of the conditional EPI, and an equivalent formulation of an inequality observed by Oohama [19].

*Lemma 3: Let $X \sim N(0, \gamma)$ and $Z \sim N(0, 1)$ be independent, and define $Y = \sqrt{\varrho} X + Z$. For $\lambda \geq 1$,*

$$\inf_{V: X \to Y \to V} \Big( I(Y; V) - \lambda I(X; V) \Big) = \begin{cases} \frac{1}{2}\left[\log\left((\lambda - 1)\gamma\,\varrho\right) - \lambda \log\left(\frac{\lambda - 1}{\lambda}(1 + \gamma\,\varrho)\right)\right] \\ \quad \text{if } \gamma\,\varrho \geq \frac{1}{\lambda - 1} \\ 0 \quad \text{if } \gamma\,\varrho \leq \frac{1}{\lambda - 1}. \end{cases}$$

*Proof:* Let $V$ be such that $X \to Y \to V$, and let $X|\{V = v\}, Y|\{V = v\}$ denote the random variables conditioned on $\{V = v\}$. Since $X, Y$ are jointly Gaussian and $V \to Y \to X$, we have $X|\{V = v\} = \rho Y|\{V = v\} + W$, where $\rho := \frac{\gamma\sqrt{\varrho}}{1 + \gamma\,\varrho}$ and $W \sim N\left(0, \gamma - \frac{\gamma^2 \varrho}{1 + \gamma\,\varrho}\right)$ is independent of $Y|\{V = v\}$. By the entropy power inequality, it holds that

$$\begin{aligned} e^{2h(X|V=v)} &\geq e^{2h(\rho Y|V=v)} + e^{2\,h(W)} \\ &= \frac{\gamma^2 \varrho}{(1 + \gamma\,\varrho)^2} e^{2h(Y|V=v)} + 2\pi e\left(\gamma - \frac{\gamma^2 \varrho}{1 + \gamma\,\varrho}\right). \end{aligned}$$

Rearranging, applying Jensen's inequality to the convex function $t \mapsto \frac{1}{2}\log\left(\frac{\gamma^2 \varrho}{(1+\gamma\,\varrho)^2} e^{2t} + 2\pi e\left(\gamma - \frac{\gamma^2 \varrho}{1+\gamma\,\varrho}\right)\right)$, and rearranging again, yields

$$e^{-2\,I(X;V)} \geq \frac{1 + \gamma\,\varrho\, e^{-2I(Y;V)}}{1 + \gamma\,\varrho}.$$

It follows that

$$\begin{aligned} &I(Y;V)-\lambda I(X;V)\\ &\geq I(Y;V)+\frac{\lambda}{2}\log\left(1+\gamma\,\varrho\,e^{-2I(Y;V)}\right)-\frac{\lambda}{2}\log(1+\gamma\,\varrho)\\ &\geq \begin{cases}\frac{1}{2}\left[\log\left((\lambda-1)\gamma\,\varrho\right)-\lambda\log\left(\frac{\lambda-1}{\lambda}\,(1+\gamma\,\varrho)\right)\right]\\ \quad \text{if } \gamma\,\varrho>\frac{1}{\lambda-1}\\ 0\quad \text{if } \gamma\,\varrho\leq\frac{1}{\lambda-1},\end{cases}\end{aligned}$$

where the second inequality follows by minimizing over the scalar quantity $I(Y;V)\geq 0$. When $\gamma\,\varrho\leq\frac{1}{\lambda-1}$, this is trivially achieved by setting $V=\mathsf{constant}$. On the other hand, if $\gamma\,\varrho>\frac{1}{\lambda-1}$, then it is easy to see that the lower bound is achieved by taking $V=Y+W$, where $W\sim N(0,\frac{1+\gamma\,\varrho}{\gamma\,\varrho(\lambda-1)-1})$ is independent of $Y$. □

*Proof of Theorem 10:* Let $X_\gamma\sim N(0,\gamma)$. Recalling the definition of $\mathsf{s}_\lambda(\cdot\,,\varrho)$, Lemma 3 implies

$$\begin{aligned}&\mathsf{s}_\lambda(X_\gamma,\varrho)\\ &=\begin{cases}\frac{1}{2}\left[\lambda\log\left(\frac{\lambda 2\pi e}{\lambda-1}\right)-\log\left(\frac{2\pi e}{\lambda-1}\right)+\log(\varrho)\right] & \text{if } \gamma\,\varrho\geq\frac{1}{\lambda-1}\\ \frac{1}{2}\left[\lambda\log\left(2\pi e\,(1+\gamma\,\varrho)\right)-\log\left(2\pi e\gamma\right)\right] & \text{if } \gamma\,\varrho\leq\frac{1}{\lambda-1}.\end{cases}\end{aligned}$$

Differentiating with respect to the quantity $\gamma$, we find that $\frac{1}{2}\left[\lambda\log\left(2\pi e\,(1+\gamma\,\varrho)\right)-\log\left(2\pi e\gamma\right)\right]$ is decreasing in $\gamma$ provided $\gamma\,\varrho\leq\frac{1}{\lambda-1}$. Therefore, taking $\gamma=1$ minimizes $\mathsf{s}_\lambda(X_\gamma,\varrho)$ over the interval $\gamma\in[0,1]$. Taken together with Theorem 9, the claim is proved. □

Given the explicit characterization of $\mathsf{V}_\lambda(\varrho)$ afforded by Theorem 10, we may now prove Theorem 8.

*Proof of Theorem 8:* We first establish (34) under the additional assumption that $\mathbb{E}[X^2]<\infty$, and generalize at the end via a truncation argument. Toward this goal, since mutual information is invariant to scaling, it is sufficient to prove that, for $Y=\sqrt{\varrho}X+Z$ with $\mathbb{E}[X^2]\leq 1$ and $Z\sim N(0,1)$ independent of $X$, we have

$$e^{2(h(Y)-I(X;V))}\geq\varrho\,e^{2(h(X)-I(Y;V))}+e^{2h(Z)} \tag{44}$$

for $V$ satisfying $X\to Y\to V$. Multiplying both sides by $\sigma^2$ and choosing $\varrho:=\frac{\mathrm{Var}(X)}{\sigma^2}$ gives the desired inequality (34) when $\mathbb{E}[X^2]<\infty$. Thus, to prove (44), observe by definition of $\mathsf{V}_\lambda(\varrho)$ that for all $\lambda\geq 1$

$$-h(X)+I(Y;V)\geq\lambda(I(X;V)-h(Y))+\mathsf{V}_\lambda(\varrho). \tag{45}$$

Inequality (44) is now immediately verified by substituting into (45) the unique $\lambda\geq 1$ which satisfies

$$\frac{\lambda}{\lambda-1}=\frac{1}{2\pi e}e^{-2(I(X;V)-h(Y))} \tag{46}$$

and simplifying. Note that there always exists a unique $\lambda\geq 1$ that solves (46) since

$$\begin{aligned}\frac{1}{2\pi e}e^{-2(I(X;V)-h(Y))}&=e^{-2h(Z)}e^{-2(I(X;V)-h(Y))}\\ &=e^{2\,I(X;Y|V)}\geq 1.\end{aligned}$$

Now, we eliminate the assumption that $\mathbb{E}[X^2]<\infty$. Toward this end, let $X$ have density, let $W$ be Gaussian independent of $X$, and consider $V$ satisfying $X\to Y\to V$, where $Y=X+W$. Define $X_n$ to be the random variable $X$ conditioned on the event $\{|X|\leq n\}$, let $Y_n=X_n+W$ and define $V_n$ via $P_{V|Y}:Y_n\mapsto V_n$. Since $X_n$ is bounded, $\mathbb{E}[X_n^2]<\infty$ so that

$$e^{2(h(Y_n)-I(X_n;V_n))}\geq e^{2(h(X_n)-I(Y_n;V_n))}+e^{2h(W)}.$$

The dominated convergence theorem can be used to show that $\lim_{n\to\infty}h(X_n)=h(X)$, provided $h(X)$ exists and is finite, which covers all cases requiring proof in view of the comments following Theorem 1. Moreover, since $X_n\xrightarrow{\mathcal{D}}X$, Lemma 10 (see Appendix A) asserts that $\lim_{n\to\infty}h(X_n+W)=h(X+W)$, so that $h(Y_n)\to h(Y)$. It is easy to see that $(X_n,V_n)\xrightarrow{\mathcal{D}}(X,V)$, so $\liminf_{n\to\infty}I(X_n;V_n)\geq I(X;V)$ by lower semicontinuity of relative entropy. Finally, the fact that $\mathbb{1}_{\{|X|\leq n\}}\to Y\to V$, combined with the chain rule for mutual information implies

$$\begin{aligned}I(Y;V)&=I(\mathbb{1}_{\{|X|\leq n\}},Y;V)\\ &\geq I(Y;V|\mathbb{1}_{\{|X|\leq n\}})\\ &\geq I(Y_n;V_n)\mathbb{P}\{|X|\leq n\},\end{aligned}$$

giving $\limsup_{n\to\infty}I(Y_n;V_n)\leq I(Y;V)$. Putting these observations together, we have established

$$e^{2(h(Y)-I(X;V))}\geq e^{2(h(X)-I(Y;V))}+e^{2h(W)}$$

in absence of second moment constraints on $X$, as desired. □

## V. Extension to Random Vectors

The vector generalization of the classical EPI is usually proved by a combination of conditioning, Jensen's inequality and induction (e.g., [52, Problem 2.9]). The same argument does not appear to readily apply in generalizing Theorem 8 to its vector counterpart in Theorem 1 due to complications arising from the Markov constraint $X\to Y\to V$. However, the desired generalization may be established by noting an additivity property enjoyed by the functional $\mathsf{s}_\lambda$ of Definition 1, appropriately generalized to probability distributions on $\mathbb{R}^d$.

For a random vector $X\sim P_X$ in $\mathbb{R}^d$, let the conditional law $P_{Y|X}$ be defined via the Gaussian channel $Y=\Gamma^{1/2}X+Z$, where $Z\sim N(0,I)$ is independent of $X$ and $\Gamma$ is a $d\times d$ diagonal matrix with nonnegative diagonal entries. Analogous to the scalar case, $(X,Y)\sim P_X P_{Y|X}$, we define the family of functionals of $P_X$

$$\begin{aligned}\mathsf{s}_\lambda(X,\Gamma):=&-h(X)+\lambda h(Y)\\ &+\inf_{V:X\to Y\to V}\left\{I(Y;V)-\lambda I(X;V)\right\}\end{aligned}$$

parameterized by $\lambda\geq 1$. Similarly, for $(Y,X,Q)\sim P_{Y|X}P_{XQ}$, define

$$\begin{aligned}\mathsf{s}_\lambda(X,\Gamma|Q):=&-h(X|Q)+\lambda h(Y|Q)\\ &+\inf_{V:X\to Y\to V|Q}\left\{I(Y;V|Q)-\lambda I(X;V|Q)\right\},\end{aligned}$$

and consider the optimization problem

$$\mathsf{V}_\lambda(\Gamma)=\inf_{P_{XQ}:\,\mathbb{E}[X_i^2]\leq 1,i\in[d]}\mathsf{s}_\lambda(X,\Gamma|Q). \tag{47}$$

*Theorem 11: If* $\Gamma = \mathrm{diag}(\varrho_1, \varrho_2, \ldots, \varrho_d)$, *then*

$$\mathsf{V}_\lambda(\Gamma) = \sum_{i=1}^{d} \mathsf{V}_\lambda(\varrho_i).$$

*Proof:* Consider $(Y, X, Q) \sim P_{Y|X} P_{XQ}$ defined as above, and let $\Gamma$ be a block diagonal matrix with blocks given by $\Gamma = \mathrm{diag}(\Gamma_1, \Gamma_2)$. Partition $X = (X_1, X_2)$ and $Z = (Z_1, Z_2)$ such that $Y_i = \Gamma_i^{1/2} X_i + Z_i$ for $i = 1, 2$. Then, for any $V$ such that $X \to Y \to V|Q$, it follows from Lemma 4 (see Section VI-B) that

$$\mathsf{s}_\lambda(X, \Gamma|Q) \geq \mathsf{s}_\lambda(X_1, \Gamma_1|X_2, Q) + \mathsf{s}_\lambda(X_2, \Gamma_2|Y_1, Q).$$

Therefore, by definitions, $\mathsf{V}_\lambda(\Gamma) \geq \mathsf{V}_\lambda(\Gamma_1) + \mathsf{V}_\lambda(\Gamma_2)$. The reverse direction of this inequality is immediate; the infimum in (47) cannot decrease if we restrict attention to measures of product form $P_{XQ} = P_{X_1\ Q_1} P_{X_2\ Q_2}$, on which $\mathsf{s}_\lambda$ is additive. The general case then follows by induction. □

Evidently, Theorem 11 relates $\mathsf{V}_\lambda(\Gamma)$ for matrix-valued parameter $\Gamma$ to the quantity $\mathsf{V}_\lambda(\varrho)$, for scalar $\varrho$. However, we have already seen a complete characterization of the latter quantity in Theorem 10. Thus, we are now positioned to establish the first stated result of this paper, Theorem 1.

*Proof of Theorem 1:* Define $Y = X + W$ for convenience, where $W \sim N(0, \Sigma_W)$ is independent of $X$. As in the scalar setting, we establish the claim first under the constraint $\mathbb{E}[\|X\|^2] < \infty$. The general result follows by a truncation argument exactly as in the scalar setting. Moreover, we may assume $\Sigma_W \succ 0$, else the inequality reduces to $h(Y) + I(Y; V) \geq h(X) + I(X; V)$, which is trivially true by the data processing inequality and the fact that conditioning reduces entropy.

Thus, due to positive definiteness of $\Sigma_W$ and invariance of mutual information under one-one transformations, we may multiply both sides of (3) by $|\Sigma_W|^{-1/d}$ to obtain the equivalent inequality

$$e^{\frac{2}{d}(h(\Sigma_W^{-1/2}Y) - I(\Sigma_W^{-1/2}X; V))} \geq e^{\frac{2}{d}(h(\Sigma_W^{-1/2}X) - I(\Sigma_W^{-1/2}Y; V))} + e^{\frac{2}{d}h(\Sigma_W^{-1/2}W)}.$$

However, $\Sigma_W^{-1/2} W \sim N(0, I)$ and, $\mathbb{E}[\|\Sigma_W^{-1/2} X\|^2] < \infty$ provided $\mathbb{E}[\|X\|^2] < \infty$, so we may assume without loss of generality that $W \sim N(0, I)$ in establishing (3).

To simplify further, put $\varrho := \max_{1 \leq i \leq d} \mathbb{E}[X_i^2]$. Note that we may assume $\varrho > 0$, else the claimed inequality is trivial since $h(X) = -\infty$ and $h(Y) - I(X; V) \geq h(Y) - I(X; Y) = h(W)$. Therefore, (3) is equivalent to

$$e^{\frac{2}{d}(h(Y) - I(X;V))} \geq \varrho\, e^{\frac{2}{d}(h(X) - I(Y;V))} + e^{\frac{2}{d}h(Z)}$$

holding for $X \to Y \to V$, where $Y = \sqrt{\varrho} X + Z$, $Z \sim N(0, I)$ is independent of $X$, and $\max_{1 \leq i \leq d} \mathbb{E}[X_i^2] \leq 1$. This is established exactly as in the proof of Theorem 8 (starting from equation (44)), since $\mathsf{V}_\lambda(\varrho \cdot I) = d\, \mathsf{V}_\lambda(\varrho)$ by Theorem 11. □

## VI. Proof of Theorem 9

This section is dedicated to the proof of Theorem 9. Our agenda will be to revisit the heuristic discussion given in Section IV-A and fill in the details where needed. In particular, we first establish the doubling property of the functional $\mathsf{s}_\lambda$, and then proceed to adapt the the heuristic argument so that it applies to near-extremal distributions.

### A. Proof of Lemma 1

In order to establish the doubling property asserted by Lemma 1, we require the following simple observation, holding for all random variables with joint distribution of a prescribed form.

*Lemma 4: Let* $\mathbf{X} = (X_+, X_-)$, $\mathbf{Y} = (Y_+, Y_-)$, *and* $Q$ *have joint distribution of the form* $P_{\mathbf{XY}Q} = P_{X_+X_-Q} P_{Y_+|X_+} P_{Y_-|X_-}$. *If* $V$ *satisfies* $\mathbf{X} \to \mathbf{Y} \to V|Q$, *then for* $\lambda \geq 1$, *we have*

$$\begin{aligned} &I(Y_+, Y_-; V|Q) - h(X_+, X_-|Q) \\ &\quad - \lambda\,(I(X_+, X_-; V|Q) - h(Y_+, Y_-|Q)) \\ &\geq I(Y_+; V|X_-, Q) - h(X_+|X_-, Q) \\ &\quad - \lambda\,(I(X_+; V|X_-, Q) - h(Y_+|X_-, Q)) \\ &\quad + I(Y_-; V|Y_+, Q) - h(X_-|Y_+, Q) \\ &\quad - \lambda\,(I(X_-; V|Y_+, Q) - h(Y_-|Y_+, Q)). \end{aligned} \tag{48}$$

*Moreover,* $X_+ \to Y_+ \to V|(X_-, Q)$ *and* $X_- \to Y_- \to V|(Y_+, Q)$.

*Proof:* The second claim is straightforward. Indeed, using $P_{\mathbf{XY}Q} = P_{X_+X_-Q} P_{Y_+|X_+} P_{Y_-|X_-}$, we can factor the joint distribution of $(\mathbf{X}, \mathbf{Y}, V, Q)$ as $P_{\mathbf{XY}VQ} = P_{X_+X_-Q} P_{Y_+|X_+} P_{Y_-|X_-} P_{V|Y_+Y_-Q} = P_{X_+X_-Q} P_{Y_+|X_+} P_{Y_-V|Y_+X_-Q}$. Marginalizing over $Y_-$, we find that $X_+ \to Y_+ \to V|(X_-, Q)$. The symmetric Markov chain follows similarly by writing $P_{\mathbf{XY}V,\ Q} = P_{X_+X_-Y_+Q} P_{Y_-|X_-} P_{V|Y_+Y_-Q}$ and marginalizing over $X_+$.

To prove the claimed inequality, note the following identity which does not make use of any particular structure of the joint distribution:

$$\begin{aligned} &I(Y_+, Y_-; V|Q) - h(X_+, X_-|Q) \\ &\quad = I(Y_+; V|Q) + I(Y_-; V|Q, Y_+) \\ &\qquad - h(X_-|Q) - h(X_+|Q, X_-) \\ &\quad = I(Y_+; V|Q) + I(Y_-; V|Q, Y_+) \\ &\qquad - h(X_-|Q, Y_+) - h(X_+|Q, X_-) - I(X_-; Y_+|Q) \\ &\quad = I(Y_+; V|Q, X_-) + I(Y_-; V|Q, Y_+) \\ &\qquad - h(X_-|Q, Y_+) - h(X_+|Q, X_-) - I(X_-; Y_+|Q, V), \end{aligned}$$

A symmetric argument gives

$$\begin{aligned} &I(X_+, X_-; V|Q) - h(Y_+, Y_-|Q) \\ &\quad = I(X_-; V|Q, Y_+) + I(X_+; V|Q, X_-) \\ &\qquad - h(Y_+|Q, X_-) - h(Y_-|Q, Y_+) - I(X_-; Y_+|Q, V). \end{aligned}$$

Therefore,

$$\begin{aligned} &I(Y_+, Y_-; V|Q) - h(X_+, X_-|Q) \\ &\qquad - \lambda\,(I(X_+, X_-; V|Q) - h(Y_+, Y_-|Q)) \\ &\quad = I(Y_+; V|X_-, Q) - h(X_+|X_-, Q) \\ &\qquad - \lambda\,(I(X_+; V|X_-, Q) - h(Y_+|X_-, Q)) \end{aligned}$$

$$
\begin{aligned}
&+ I(Y_-; V|Y_+, Q) - h(X_-|Y_+, Q) \\
&- \lambda\left(I(X_-; V|Y_+, Q) - h(Y_-|Y_+, Q)\right) \\
&+ (\lambda - 1) I(X_-; Y_+|V, Q),
\end{aligned}
$$

which proves the inequality (48) since $\lambda \geq 1$ and mutual information is non-negative. □

We are now in a position to establish Lemma 1, stated in Section IV-A.

*Proof of Lemma 1:* Let all quantities be as defined in the statement of the lemma. The crucial observation is that the unitary transformation $(Y_1, Y_2) \mapsto (Y_+, Y_-)$ preserves the Gaussian nature of the channel. That is, if $Y_i = \sqrt{\varrho} X_i + Z_i$, then $Y_+ = \sqrt{\varrho} X_+ + \frac{1}{\sqrt{2}}(Z_1 + Z_2)$ and $Y_- = \sqrt{\varrho} X_- + \frac{1}{\sqrt{2}}(Z_1 - Z_2)$, where the pair $(\frac{1}{\sqrt{2}}(Z_1 + Z_2), \frac{1}{\sqrt{2}}(Z_1 - Z_2))$ is equal in distribution to $(Z_1, Z_2)$.

Thus, consider an arbitrary $V$ satisfying $(X_+, X_-) \to (Y_+, Y_-) \to V|\mathbf{Q}$. By Lemma 4 and the above observation, we have

$$
\begin{aligned}
&I(Y_1, Y_2; V|\mathbf{Q}) - h(X_1, X_2|\mathbf{Q}) \\
&\quad - \lambda\left(I(X_1, X_2; V|\mathbf{Q}) - h(Y_1, Y_2|\mathbf{Q})\right) \\
&= I(Y_+, Y_-; V|\mathbf{Q}) - h(X_+, X_-|\mathbf{Q}) \\
&\quad - \lambda\left(I(X_+, X_-; V|\mathbf{Q}) - h(Y_+, Y_-|\mathbf{Q})\right) \\
&\geq I(Y_+; V|X_-, Q) - h(X_+|X_-, \mathbf{Q}) \\
&\quad - \lambda\left(I(X_+; V|X_-, \mathbf{Q}) - h(Y_+|X_-, \mathbf{Q})\right) \\
&\quad + I(Y_-; V|Y_+, \mathbf{Q}) - h(X_-|Y_+, \mathbf{Q}) \\
&\quad - \lambda\left(I(X_-; V|Y_+, \mathbf{Q}) - h(Y_-|Y_+, \mathbf{Q})\right) \\
&\geq \mathsf{s}_\lambda(X_+, \varrho|X_-, \mathbf{Q}) + \mathsf{s}_\lambda(X_-, \varrho|Y_+, \mathbf{Q}).
\end{aligned}
$$

This proves (38) since

$$
\begin{aligned}
&\inf_{V: X \to Y \to V|\mathbf{Q}} \Big\{ I(Y_1, Y_2; V|\mathbf{Q}) - h(X_1, X_2|\mathbf{Q}) \\
&\quad - \lambda\left(I(X_1, X_2; V|\mathbf{Q}) - h(Y_1, Y_2|\mathbf{Q})\right) \Big\} \\
&\leq \sum_{i=1}^{2} \inf_{V: X_i \to Y_i \to V|Q_i} \Big\{ I(Y_i; V|Q_i) - h(X_i|Q_i) \\
&\quad - \lambda\left(I(X_i; V|Q_i) - h(Y_i|Q_i)\right) \Big\} \\
&= 2\,\mathsf{s}_\lambda(X, \varrho|Q),
\end{aligned}
$$

where the inequality follows since the infimum is taken over a smaller set. □

### *B. Existence of Sequences Satisfying $\lim_{n\to\infty} \mathsf{s}_\lambda(X_n, \varrho|Q_n) = \mathsf{V}_\lambda(\varrho)$ That Converge Weakly to Gaussian*

Define $\mathcal{Q}_3 := \{1, 2, 3\}$. The following definition will also be convenient.

*Definition 2: For given $\varrho > 0$ and $\lambda \geq 1$, a sequence $\{X_n, Q_n\}$ is said to be* admissible *if, for each $n \geq 1$, $(X_n, Q_n)$ takes values in $\mathbb{R} \times \mathcal{Q}_3$, and the following two conditions hold:*

$$
\lim_{n\to\infty} \mathsf{s}_\lambda(X_n, \varrho|Q_n) = \mathsf{V}_\lambda(\varrho) \tag{49}
$$

$$
\mathbb{E}[X_n^2] \leq 1 \quad n \geq 1. \tag{50}
$$

From Remark 2, it is clear that the set of admissible sequences is nonempty, even under the restriction that the variables $Q_n$ take values in the three-point set $\mathcal{Q}_3$. In fact, two points would suffice to ensure the existence of admissible sequences, but the additional degree of freedom afforded by the third point will be needed in our proof. Associate to each $(X_n, Q_n)$ taking values in $\mathbb{R} \times \mathcal{Q}_3$ the random variable $Y_n$ defined by $Y_n = \sqrt{\varrho} X_n + Z$, where $Z \sim N(0, 1)$ is independent of $(X_n, Q_n)$. We define the following quantity:

$$
h^\star(\varrho) = \inf \Big\{ \liminf_{n\to\infty} \left(h(Y_n|Q_n) - h(X_n|Q_n)\right) : \text{the sequence } \{X_n, Q_n\} \text{ is admissible} \Big\}.
$$

This definition is meant to capture the extremal property (40) in the discussion of Section IV-A, appropriately modified for admissible sequences. In particular, we will aim to show that admissible sequences $\{X_n, Q_n\}$ which are extremal in the sense that $\liminf_{n\to\infty} (h(Y_n|Q_n) - h(X_n|Q_n)) = h^\star(\varrho)$ will be guaranteed to approach normality in a precise sense. First, we observe that degeneracy is avoided in the sense that $h^\star(\varrho)$ is always finite.

*Proposition 2: It holds that $|h^\star(\varrho)| < \infty$.*

*Proof:* Let $(Y_n, X_n, Q_n)$ be as above, with $\{X_n, Q_n\}$ being an admissible sequence. Note first that $h(Y_n|Q_n) - h(X_n|Q_n) \geq 0$ since conditioning reduces entropy, so $h^\star(\varrho) \geq 0$. On the other hand, since $\{X_n, Q_n\}$ is admissible, we have $\mathsf{s}_\lambda(X_n, \varrho|Q_n) < \mathsf{V}_\lambda(\varrho) + 1$ for $n$ sufficiently large by definition, provided $\mathsf{V}_\lambda(\varrho) > -\infty$. However, this is always the case. Indeed, for $\lambda \geq 1$, we use again the inequality $h(Y_n|Q_n) - h(X_n|Q_n) \geq 0$ to observe

$$
\begin{aligned}
-h(X_n|Q_n) + \lambda h(Y_n|Q_n) &\geq (\lambda - 1) h(Y_n|Q_n) \\
&\geq (\lambda - 1) h(Z) \\
&= (\lambda - 1)\frac{1}{2}\log(2\pi e),
\end{aligned} \tag{51}
$$

where $Z \sim N(0, 1)$ and (51) follows from the definition of $Y_n$ and the fact that conditioning reduces entropy. Hence,

$$
\begin{aligned}
&\mathsf{s}_\lambda(X_n, \varrho|Q_n) \\
&\quad = -h(X_n|Q_n) + \lambda h(Y_n|Q_n) \\
&\qquad + \inf_{V: X_n \to Y_n \to V|Q_n} \Big\{ I(Y_n; V|Q_n) - \lambda I(X_n; V|Q_n) \Big\} \\
&\quad \geq (\lambda - 1)\frac{1}{2}\log(2\pi e) \\
&\qquad + \inf_{V: X_n \to Y_n \to V|Q_n} \Big\{ I(Y_n; V|Q_n) - \lambda I(X_n; V|Q_n) \Big\} \\
&\quad \geq (\lambda - 1)\frac{1}{2}\log(2\pi e) - (\lambda - 1) I(X_n; Y_n|Q_n),
\end{aligned} \tag{52}
$$

where (52) follows from the data processing inequalities $I(Y_n; V|Q_n) \geq I(X_n; V|Q_n)$ and $I(X_n; Y_n|Q_n) \geq I(X_n; V|Q_n)$. Of course, $I(X_n; Y_n|Q_n) \leq \frac{1}{2}\log(1 + \varrho)$ by the maximum entropy property of Gaussians (since $\mathbb{E}X_n^2 \leq 1$ by admissibility), so that $\mathsf{V}_\lambda(\varrho) > -\infty$ as claimed.

Therefore, if $n$ is sufficiently large so that $\mathsf{s}_\lambda(X_n, \varrho|Q_n) < \mathsf{V}_\lambda(\varrho) + 1$, there is necessarily some $V_n$ satisfying

$X_n \to Y_n \to V_n|Q_n$ for which

$$\begin{aligned} h(Y_n|Q_n) - h(X_n|Q_n) &\le \mathsf{V}_\lambda(\varrho) + 1 + \lambda I(X_n; V_n|Q_n) - I(Y_n; V_n|Q_n) \\ &\quad - (\lambda-1)h(Y_n|Q_n) \\ &\le \mathsf{V}_\lambda(\varrho) + 1 + (\lambda-1) I(X_n; V_n|Q_n) \\ &\quad - (\lambda-1)h(Y_n|Q_n) \qquad (53) \\ &\le \mathsf{V}_\lambda(\varrho) + 1 + (\lambda-1) I(X_n; Y_n|Q_n) \\ &\quad - (\lambda-1)h(Y_n|Q_n) \\ &= \mathsf{V}_\lambda(\varrho) + 1 - (\lambda-1)h(Y_n|X_n), \qquad (54) \end{aligned}$$

where (53) and (54) are both due to the data processing inequality. Since $\mathsf{V}_\lambda(\varrho) < \infty$ trivially and $h(Y_n|X_n) = h(Z) = \frac{1}{2}\log 2\pi e$, we conclude that $h^\star(\varrho) < \infty$, establishing the claim. □

At this point, we may essentially repeat the heuristic discussion of Section IV-A to obtain the following conclusion that applies to distributions that are near-extremal in a precise sense.

*Lemma 5: Fix $\epsilon > 0$. Consider a distribution $P_{XQ}$ on $\mathbb{R} \times \mathcal{Q}_3$ such that $\mathbb{E}X = 0$, $\mathbb{E}X^2 \le 1$,*

$$\mathsf{s}_\lambda(X, \varrho|Q) \le \mathsf{V}_\lambda(\varrho) + \epsilon \qquad (55)$$

*and*

$$h(Y|Q) - h(X|Q) \le h^\star(\varrho) + \epsilon, \qquad (56)$$

*where $(Y, X, Q) \sim \mathbb{G}^\varrho_{Y|X} P_{XQ}$. There exists a distribution $P_{X'Q'}$ on $\mathbb{R} \times \mathcal{Q}_3$ such that, for $(Y', X', Q') \sim \mathbb{G}^\varrho_{Y'|X'} P_{X'Q'}$, we have $\mathbb{E}X'^2 \le 1$,*

$$\mathsf{s}_\lambda(X', \varrho|Q') \le \mathsf{V}_\lambda(\varrho) + 2\epsilon$$

*and*

$$\begin{aligned} &h(Y'|Q') - h(X'|Q') \\ &\quad + \frac{1}{2} I(X_1 + X_2; X_1 - X_2|Y_1, Y_2, Q_1, Q_2) \le h^\star(\varrho) + \epsilon, \end{aligned}$$

*where $(Y_1, X_1, Q_1)$ and $(Y_2, X_2, Q_2)$ denote independent copies of $(Y, X, Q)$.*

*Proof:* Let $X_+, X_-, Y_+, Y_-, \mathbf{Q}$ be as in Lemma 1, constructed from the two independent copies of $(Y, X, Q)$. Applying Lemma 1 to the variables $\mathbf{Q} \to (X_+, X_-) \to (Y_+, Y_-)$, we obtain

$$2\mathsf{s}_\lambda(X, \varrho|Q) \ge \mathsf{s}_\lambda(X_+, \varrho|X_-, \mathbf{Q}) + \mathsf{s}_\lambda(X_-, \varrho|Y_+, \mathbf{Q}), \qquad (57)$$

and the symmetric inequality

$$2\mathsf{s}_\lambda(X, \varrho|Q) \ge \mathsf{s}_\lambda(X_+, \varrho|Y_-, \mathbf{Q}) + \mathsf{s}_\lambda(X_-, \varrho|X_+, \mathbf{Q}). \qquad (58)$$

By independence of $X_1$ and $X_2$ and the assumption that $\mathbb{E}X = 0$, we have

$$\mathbb{E}[X_+^2] = \mathbb{E}[X_-^2] = \frac{1}{2}\mathbb{E}[X_1^2] + \frac{1}{2}\mathbb{E}[X_2^2] = \mathbb{E}[X^2] \le 1.$$

Hence, it follows that the terms in the RHS of (57) and the RHS of (58) are each lower bounded by $\mathsf{V}_\lambda(\varrho)$. Combined with (55), we may conclude that

$$\frac{1}{2}\mathsf{s}_\lambda(X_+, \varrho|Y_-, \mathbf{Q}) + \frac{1}{2}\mathsf{s}_\lambda(X_-, \varrho|Y_+, \mathbf{Q}) \le \mathsf{V}_\lambda(\varrho) + 2\epsilon. \qquad (59)$$

Next, exactly as in Section IV-A, let $B \sim \text{Bernoulli}(1/2)$ be a Bernoulli random variable taking values on $\{+, -\}$, independent of $X_+, X_-, Y_+, Y_-, \mathbf{Q}$, and define $\bar{B}$ to be the complement of $B$ in $\{+, -\}$. Construct a new pair of random variables $(\tilde{X}, \tilde{Q})$ as follows: $\tilde{X} = X_B$ and $\tilde{Q} = (B, Y_{\bar{B}}, \mathbf{Q})$. Clearly, $\mathbb{E}\tilde{X}^2 = \mathbb{E}X^2 \le 1$. Also, by construction and (59),

$$\begin{aligned} \mathsf{s}_\lambda(\tilde{X}, \varrho|\tilde{Q}) &= \frac{1}{2}\mathsf{s}_\lambda(X_+, \varrho|Y_-, \mathbf{Q}) + \frac{1}{2}\mathsf{s}_\lambda(X_-, \varrho|Y_+, \mathbf{Q}) \\ &\le \mathsf{V}_\lambda(\varrho) + 2\epsilon. \end{aligned}$$

Moreover, using Markovity, we may establish

$$\begin{aligned} &h(Y|Q) - h(X|Q) \\ &\quad = \frac{1}{2}\left(h(Y_+, Y_-|\mathbf{Q}) - h(X_+, X_-|\mathbf{Q})\right) \\ &\quad = \frac{1}{2}\left(h(Y_-|Y_+, \mathbf{Q}) - h(X_-|Y_+, \mathbf{Q})\right) \\ &\qquad + \frac{1}{2}\left(h(Y_+|Y_-, \mathbf{Q}) - h(X_+|Y_-, \mathbf{Q})\right) \\ &\qquad + \frac{1}{2} I(X_+; X_-|Y_+, Y_-, \mathbf{Q}) \\ &\quad = h(\tilde{Y}|\tilde{Q}) - h(\tilde{X}|\tilde{Q}) + \frac{1}{2} I(X_+; X_-|Y_+, Y_-, \mathbf{Q}), \end{aligned}$$

where $\tilde{Y} = \sqrt{\varrho}\tilde{X} + Z$, with $Z \sim N(0, 1)$ independent of $(\tilde{X}, \tilde{Q})$. In particular, using (56), we have

$$h(\tilde{Y}|\tilde{Q}) - h(\tilde{X}|\tilde{Q}) + \frac{1}{2} I(X_+; X_-|Y_+, Y_-, \mathbf{Q}) \le h^\star(\varrho) + \epsilon.$$

Evidently, the pair $(\tilde{X}, \tilde{Q})$ satisfies nearly all of the stated properties of $(X', Q')$ in the claim to be proved. The only exception is that $\tilde{Q}$ takes values in the space $\{+, -\} \times \mathbb{R} \times \mathcal{Q}_3 \times \mathcal{Q}_3$, whereas we require that $Q'$ takes values in $\mathcal{Q}_3$. This is easily dealt with by applying the standard dimensionality reduction procedure described in Remark 2. Indeed, by the Fenchel-Caratheodory-Bunt [72, Theorem 18$(ii)$], there is a pair $(X', Q')$ with $Q'$ supported on at most three points that preserves the values $\mathbb{E}X'^2 = \mathbb{E}\tilde{X}^2$, $\mathsf{s}_\lambda(X', \varrho|Q') = \mathsf{s}_\lambda(\tilde{X}, \varrho|\tilde{Q})$ and $\big(h(Y'|Q') - h(X'|Q')\big) = (h(\tilde{Y}|\tilde{Q}) - h(\tilde{X}|\tilde{Q}))$. Thus, the claim is proved. □

Applying Lemma 5 to an appropriately chosen admissible sequence, we obtain the following asymptotic analog of property (43), which was a key milestone in the heuristic discussion of Section IV-A.

*Lemma 6: There exists an admissible sequence $\{X_n, Q_n\}$ and a distribution $P_{X_*Q_*}$ on $\mathbb{R} \times \mathcal{Q}_3$ such that $(X_n, Q_n) \xrightarrow{\mathcal{D}} (X_*, Q_*) \sim P_{X_*Q_*}$, with $\mathbb{E}[X_*^2] \le 1$ and*

$$\liminf_{n\to\infty} I(X_{1,n} + X_{2,n}; X_{1,n} - X_{2,n}|\mathbf{Y}_n, \mathbf{Q}_n = \mathbf{q}) = 0 \qquad (60)$$

*for $P_{Q_*} \times P_{Q_*}$-a.e. $\mathbf{q}$, where $\mathbf{Y}_n := (Y_{1,n}, Y_{2,n})$, $\mathbf{Q}_n := (Q_{1,n}, Q_{2,n})$, and $(Y_{1,n}, X_{1,n}, Q_{1,n})$ and $(Y_{2,n}, X_{2,n}, Q_{2,n})$ denote independent copies of $(Y_n, X_n, Q_n) \sim \mathbb{G}^\varrho_{Y_n|X_n} P_{X_nQ_n}$.*

*Proof:* Let $\{X_n, Q_n\}$ be an admissible sequence with the additional property that

$$\lim_{n\to\infty} \left(h(Y_n|Q_n) - h(X_n|Q_n)\right) = h^\star(\varrho).$$

By a diagonalization argument, such an admissible sequence exists. Since $\mathcal{Q}_3$ is finite and $\mathbb{E}[X_n^2] \le 1$,

the sequence $\{X_n, Q_n\}$ is tight. By Prokhorov's theorem [82], we may assume that there is some $(X_*, Q_*)$ for which $(X_n, Q_n) \xrightarrow{\mathcal{D}} (X_*, Q_*)$ by restricting our attention to a subsequence of $\{X_n, Q_n\}$ if necessary. Moreover, $\mathbb{E}[X_*^2] \leq \liminf_{n\to\infty} \mathbb{E}[X_n^2] \leq 1$ by Fatou's lemma.

Applying Lemma 5 along the sequence $\{X_n, Q_n\}$ for each $n$, we conclude the existence of another admissible sequence $\{X_n', Q_n'\}$ with the property that

$$\begin{aligned}\liminf_{n\to\infty} & I(X_{1,n} + X_{2,n}; X_{1,n} - X_{2,n}|\mathbf{Y}_n, \mathbf{Q}_n) \\ & + \liminf_{n\to\infty} \left(h(Y_n'|Q_n') - h(X_n'|Q_n')\right) \\ & \leq h^\star(\varrho).\end{aligned}$$

However, as $\{X_n', Q_n'\}$ is admissible, the definition of $h^\star(\varrho)$ combined with the fact that it is finite (Proposition 2) implies that we must have

$$\liminf_{n\to\infty} I(X_{1,n} + X_{2,n}; X_{1,n} - X_{2,n}|\mathbf{Y}_n, \mathbf{Q}_n) = 0. \quad (61)$$

Since $\mathcal{Q}_3$ is finite and $Q_n \xrightarrow{\mathcal{D}} Q_*$, the desired result follows from (61). □

As suggested in the discussion at the end of Section IV-A, to finish the proof of Theorem 9 we will require (i) a generalization of Bernstein's theorem with hypothesis compatible with (60); and (ii) a local semicontinuity property of $\mathsf{s}_\lambda(\cdot, \varrho)$. Stated precisely, the two required ingredients are the following:

*Lemma 7: Suppose* $(X_{1,n}, X_{2,n}) \xrightarrow{\mathcal{D}} (X_{1,*}, X_{2,*})$ *with* $\sup_n \mathbb{E}[X_{i,n}^2] < \infty$ *for* $i = 1, 2$. *Let* $(Z_1, Z_2) \sim N(0, \sigma^2\, I)$ *be pairwise independent of* $(X_{1,n}, X_{2,n})$ *and, for* $i = 1, 2$, *define* $Y_{i,n} = X_{i,n} + Z_i$. *If* $X_{1,n}, X_{2,n}$ *are independent and*

$$\liminf_{n\to\infty} I(X_{1,n} + X_{2,n}; X_{1,n} - X_{2,n}|Y_{1,n}, Y_{2,n}) = 0, \quad (62)$$

*then* $X_{1,*}, X_{2,*}$ *are independent Gaussian random variables with identical variances.*

*Lemma 8: If* $X_n \xrightarrow{\mathcal{D}} X_* \sim N(\mu, \sigma_X^2)$ *and* $\sup_n \mathbb{E}[X_n^2] < \infty$, *then*

$$\liminf_{n\to\infty} \mathsf{s}_\lambda(X_n, \varrho) \geq \mathsf{s}_\lambda(X_*, \varrho). \quad (63)$$

By assembling Lemmas 6, 7 and 8, Theorem 9 follows almost immediately. The argument is as follows.

*Proof of Theorem 9:* Let $\{X_n, Q_n\}$, $(Y_{1,n}, X_{1,n}, Q_{1,n})$, $(Y_{2,n}, X_{2,n}, Q_{2,n})$ and $(X_*, Q_*)$ be as in Lemma 6. Since $\mathcal{Q}_3$ is finite, we of course have $X_{i,n}|\{Q_{i,n} = q_i\} \xrightarrow{\mathcal{D}} X_*|\{Q_* = q_i\}$, for $i = 1, 2$ and $P_{Q_*}$-a.e. $q_i$. Now, (60) fulfills the hypothesis of Lemma 7, so we are able to conclude that for $P_{Q_*} \times P_{Q_*}$-a.e. $q_1, q_2$, the random variables $X_*|\{Q_* = q_1\}$ and $X_*|\{Q_* = q_2\}$ are Gaussian with identical variances. In other words, for $P_{Q_*}$-a.e. $q$, we have $X_*|\{Q_* = q\} \sim N(\mu_q, \sigma_X^2)$, where $\mu_q$ possibly depends on $q$, but $\sigma_X^2$ does not. Since $\mathbb{E}X_*^2 \leq 1$, we must have $\sigma_X^2 \leq 1$.

So, using the lower semicontinuity property of Lemma 7, for $P_{Q_*}$-a.e. $q$,

$$\begin{aligned}\liminf_{n\to\infty} \mathsf{s}_\lambda(X_n|\{Q_n = q\}, \varrho) &\geq \mathsf{s}_\lambda(N(\mu_q, \sigma_X^2), \varrho) \\ &= \mathsf{s}_\lambda(N(0, \sigma_X^2), \varrho),\end{aligned}$$

where the equality follows since $\mathsf{s}_\lambda(\cdot, \varrho)$ is invariant to translation of the mean. Thus,

$$\liminf_{n\to\infty} P_{Q_n}(q)\mathsf{s}_\lambda(X_n|\{Q_n = q\}, \varrho) \geq P_{Q_*}(q)\mathsf{s}_\lambda(N(0, \sigma_X^2), \varrho)$$

for each $q \in \mathcal{Q}_3$ satisfying $P_{Q_*}(q) > 0$.

We now turn our attention to the possible situation where $P_{Q_*}(q) = 0$ for some $q$. Following the same steps leading to (52), for $\lambda \geq 1$ we have the lower bound

$$\begin{aligned}P_{Q_n}(q)&\mathsf{s}_\lambda(X_n|\{Q_n = q\}, \varrho) \\ &\geq P_{Q_n}(q)(\lambda - 1)\left(\frac{1}{2}\log(2\pi e) - I(X_n; Y_n|Q_n = q)\right),\end{aligned} \quad (64)$$

valid for all $q \in \mathcal{Q}_3$. Since $\{X_n, Q_n\}$ is admissible, for each $n \geq 1$,

$$P_{Q_n}(q)\mathbb{E}[X_n^2|Q_n = q] \leq \mathbb{E}[X_n^2] \leq 1 \quad \forall q \in \mathcal{Q}_3.$$

Rearranging gives $\mathbb{E}[X_n^2|Q_n = q] \leq \left(P_{Q_n}(q)\right)^{-1}$. By the maximum-entropy property of Gaussians and definition of $(Y_n, X_n, Q_n)$,

$$\begin{aligned}I(X_n; Y_n|Q_n = q) &\leq \frac{1}{2}\log\left(1 + \varrho\,\mathbb{E}[X_n^2|Q_n = q]\right) \\ &\leq \frac{1}{2}\log\left(1 + \varrho\left(P_{Q_n}(q)\right)^{-1}\right),\end{aligned}$$

so that (64) yields

$$\begin{aligned}P_{Q_n}(q)&\mathsf{s}_\lambda(X_n|\{Q_n = q\}, \varrho) \\ &\geq \frac{1}{2}P_{Q_n}(q)(\lambda - 1)\left(\log(2\pi e) - \log\left(1 + \varrho\left(P_{Q_n}(q)\right)^{-1}\right)\right).\end{aligned}$$

Hence, if $q$ is such that $P_{Q_n}(q) \to P_{Q_*}(q) = 0$, then

$$\liminf_{n\to\infty} P_{Q_n}(q)\mathsf{s}_\lambda(X_n|\{Q_n = q\}, \varrho) \geq 0.$$

Combining the above establishes

$$\liminf_{n\to\infty} \mathsf{s}_\lambda(X_n, \varrho|Q_n) \geq \mathsf{s}_\lambda(N(0, \sigma_X^2), \varrho). \quad (65)$$

Since $\{X_n, Q_n\}$ was assumed to be an admissible sequence, (65) and (49) together ensure that

$$\mathsf{V}_\lambda(\varrho) \geq \mathsf{s}_\lambda(N(0, \sigma_X^2), \varrho),$$

which proves the claim. □

## Appendix A
## A Weak Form of Bernstein's Theorem

The aim of this appendix is to prove Lemma 7, which is largely a matter of assembling the needed ingredients. We begin by recalling two facts about random variables that are contaminated by Gaussian noise. Of particular interest to us are weakly convergent sequences of random variables, and corresponding continuity properties when densities are regularized via convolution with a Gaussian density.

*Lemma 9 [74, Lemma 5.1.3]: If* $X, Z$ *are independent random variables and* $Z$ *is normal, then* $X + Z$ *has a non-vanishing probability density function which has derivatives of all orders.*

*Lemma 10 [70, Propositions 16 and 18]: Let* $X_n \xrightarrow{\mathcal{D}} X_*$ *with* $\sup_n \mathbb{E}[\|X_n\|^2] < \infty$, *and let* $Z \sim N(0, \sigma^2 I)$ *be a non-degenerate Gaussian, independent of* $\{X_n\}$, $X_*$. *Let* $Y_n = X_n + Z$ *and* $Y_* = X_* + Z$. *Finally, let* $f_n$ *and* $f_*$ *denote the density of* $Y_n$ *and* $Y_*$, *respectively. Then*

1. $Y_n \xrightarrow{\mathcal{D}} Y_*$
2. $\|f_n - f_*\|_\infty \to 0$
3. $h(Y_n) \to h(Y_*)$.

Let us also make note of two characterizations of the normal distribution. First, we remind the reader of the result of Bernstein given previously as Lemma 2. Second, we will need the following observation:

*Lemma 11: Let* $Y = X + Z$, *where* $Z \sim N(0, \sigma^2)$ *is a non-degenerate Gaussian, independent of* $X$. *If* $X|\{Y = y\}$ *is normal for* $P_Y$*-a.e.* $y$, *with variance* $\sigma_X^2$ *not depending on* $y$, *then* $X$ *is normal with variance* $\frac{\sigma^2\sigma_X^2}{\sigma^2 - \sigma_X^2}$.

*Proof:* If $X|\{Y = y\}$ is normal for $P_Y$-a.e. $y$ with variance $\sigma_X^2$ not depending on $Y$, then $X = \mathbb{E}[X|Y] + W$ in distribution, where $W \sim N(0, \sigma_X^2)$ is independent of $Y$. In particular, $X$ has density $f_X$ by Lemma 9. Also by Lemma 9, $Y$ has density $f_Y$. The conditional density $f_{Y|X}$ exists and is Gaussian by definition, and $f_{X|Y}$ is a valid Gaussian density for $P_Y$-a.e. $y$, with corresponding variance $\sigma_X^2$ not depending on $y$. Thus, we have

$$\log f_X(x) = \log f_Y(y) + \log f_{Y|X}(y|x) - \log f_{X|Y}(x|y). \tag{66}$$

The key observation is that the RHS of (66) is a quadratic function in $x$. Since $f_X$ is a density and must integrate to unity, it must therefore be Gaussian. Direct computation reveals that $X$ has variance $\frac{\sigma^2\sigma_X^2}{\sigma^2 - \sigma_X^2}$. □

In final preparation for the proof of Lemma 7, we record the following consequence of Lemma 10 and lower semicontinuity of relative entropy:

*Lemma 12: Suppose* $(X_{1,n}, X_{2,n}) \xrightarrow{\mathcal{D}} (X_{1,*}, X_{2,*})$ *with* $\sup_n \mathbb{E}[X_{i,n}^2] < \infty$ *for* $i = 1, 2$. *Let* $(Z_1, Z_2) \sim N(0, \sigma^2 I)$ *be pairwise independent of* $(X_{1,n}, X_{2,n})$ *and* $(X_{1,*}, X_{2,*})$, *and, for* $i = 1, 2$, *define* $Y_{i,n} = X_{i,n} + Z_i$ *and* $Y_{i,*} = X_{i,*} + Z_i$. *Then* $(Y_{1,n}, Y_{2,n}) \xrightarrow{\mathcal{D}} (Y_{1,*}, Y_{2,*})$ *and*

$$\liminf_{n\to\infty} I(X_{1,n}; X_{2,n}|Y_{1,n}, Y_{2,n}) \geq I(X_{1,*}; X_{2,*}|Y_{1,*}, Y_{2,*}). \tag{67}$$

*Proof:* The fact that $(Y_{1,n}, Y_{2,n}) \xrightarrow{\mathcal{D}} (Y_{1,*}, Y_{2,*})$ follows from Lemma 10. Lemma 10 also establishes that

$$h(Y_{1,n}, Y_{2,n}) \to h(Y_{1,*}, Y_{2,*}). \tag{68}$$

On account of the Markov chains $(X_{2,n}, Y_{2,n}) \to X_{1,n} \to Y_{1,n}$ and $(X_{1,n}, Y_{1,n}) \to X_{2,n} \to Y_{2,n}$, we have the identity

$$\begin{aligned} &I(X_{1,n}; X_{2,n}|Y_{1,n}, Y_{2,n}) \\ &= I(X_{1,n}, Y_{2,n}; Y_{1,n}, X_{2,n}) - I(X_{1,n}, X_{2,n}; Y_{1,n}, Y_{2,n}), \end{aligned} \tag{69}$$

which is verified as follows:

$$\begin{aligned} &I(X_{1,n}; X_{2,n}|Y_{1,n}, Y_{2,n}) \\ &\quad = I(X_{1,n}; Y_{1,n}|Y_{2,n}) + I(X_{1,n}; X_{2,n}|Y_{1,n}, Y_{2,n}) \\ &\qquad - I(X_{1,n}; Y_{1,n}|Y_{2,n}) \\ &\quad = I(X_{1,n}; Y_{1,n}, X_{2,n}|Y_{2,n}) - I(X_{1,n}, X_{2,n}; Y_{1,n}|Y_{2,n}) \\ &\quad = I(Y_{2,n}; X_{2,n}) + I(X_{1,n}; Y_{1,n}, X_{2,n}|Y_{2,n}) \\ &\qquad - I(X_{1,n}, X_{2,n}; Y_{2,n}) - I(X_{1,n}, X_{2,n}; Y_{1,n}|Y_{2,n}) \\ &\quad = I(Y_{2,n}; Y_{1,n}, X_{2,n}) + I(X_{1,n}; Y_{1,n}, X_{2,n}|Y_{2,n}) \\ &\qquad - I(X_{1,n}, X_{2,n}; Y_{1,n}, Y_{2,n}) \\ &\quad = I(X_{1,n}, Y_{2,n}; Y_{1,n}, X_{2,n}) - I(X_{1,n}, X_{2,n}; Y_{1,n}, Y_{2,n}). \end{aligned}$$

Observe that $\liminf_{n\to\infty} I(X_{1,n}, Y_{2,n}; Y_{1,n}, X_{2,n}) \geq I(X_{1,*}, Y_{2,*}; Y_{1,*}, X_{2,*})$ due to lower semicontinuity of relative entropy, and $\lim_{n\to\infty} I(X_{1,n}, X_{2,n}; Y_{1,n}, Y_{2,n}) = I(X_{1,*}, X_{2,*}; Y_{1,*}, Y_{2,*})$ due to (68) and the fact that $h(Y_{1,*}, Y_{2,*}|X_{1,*}, X_{2,*}) = h(Y_{1,n}, Y_{2,n}|X_{1,n}, X_{2,n}) = h(Z_1, Z_2)$ is constant. Thus, (67) is proved by applying the identity (69) again for $(X_{1,*}, X_{2,*}, Y_{1,*}, Y_{2,*})$. □

Having assembled the required ingredients, we now prove Lemma 7.

*Proof of Lemma 7:* Let $Y_{i,*}$ be as in the statement of Lemma 12, and recall that the same lemma asserts $(Y_{1,n}, Y_{2,n}) \xrightarrow{\mathcal{D}} (Y_{1,*}, Y_{2,*})$. By definition of $Z_1, Z_2$, the random variables $(Z_1 + Z_2)$ and $(Z_1 - Z_2)$ are independent and Gaussian with respective variances $2\sigma^2$. Thus, noting that assumption (62) is equivalent to

$$\liminf_{n\to\infty} I(X_{1,n} + X_{2,n}; X_{1,n} - X_{2,n}|Y_{1,n} + Y_{2,n}, Y_{1,n} - Y_{2,n}) = 0,$$

we may apply Lemma 12 to the sequences $\{X_{1,n} + X_{2,n}, X_{1,n} - X_{2,n}\}$ and $\{Y_{1,n} + Y_{2,n}, Y_{1,n} - Y_{2,n}\}$ to obtain

$$\begin{aligned} &I(X_{1,*} + X_{2,*}; X_{1,*} - X_{2,*}|Y_{1,*}, Y_{2,*}) \\ &\quad = I(X_{1,*} + X_{2,*}; X_{1,*} - X_{2,*}|Y_{1,*} + Y_{2,*}, Y_{1,*} - Y_{2,*}) \\ &\quad = 0. \end{aligned} \tag{70}$$

Using independence of $X_{1,n}, X_{2,n}$, Lemma 12 applied directly yields

$$I(X_{1,*}; X_{2,*}|Y_{1,*}, Y_{2,*}) = 0. \tag{71}$$

In particular, for $P_{Y_{1,*}Y_{2,*}}$-a.e. $y_1, y_2$, the random variables $X_{1,*}|\{Y_{1,*}, Y_{2,*} = y_1, y_2\}$ and $X_{2,*}|\{Y_{1,*}, Y_{2,*} = y_1, y_2\}$ are independent by (71), and $(X_{1,*} + X_{2,*})|\{Y_{1,*}, Y_{2,*} = y_1, y_2\}$ and $(X_{1,*} - X_{2,*})|\{Y_{1,*}, Y_{2,*} = y_1, y_2\}$ are independent by (70). Therefore, Lemma 2 implies that $X_{1,*}|\{Y_{1,*}, Y_{2,*} = y_1, y_2\}$ and $X_{2,*}|\{Y_{1,*}, Y_{2,*} = y_1, y_2\}$ are each normal with identical variances for $P_{Y_{1,*}Y_{2,*}}$-a.e. $y_1, y_2$. Starting with the third claim of Lemma 10 and applying lower semicontinuity of relative entropy, we observe

$$\begin{aligned} I(X_{1,*}; Y_{1,*}) &= \lim_{n\to\infty} I(X_{1,n}; Y_{1,n}) \\ &= \lim_{n\to\infty} I(X_{1,n}; Y_{1,n}, Y_{2,n}) \\ &\geq I(X_{1,*}; Y_{1,*}, Y_{2,*}) \\ &= I(X_{1,*}; Y_{1,*}) + I(X_{1,*}; Y_{2,*}|Y_{1,*}), \end{aligned}$$

so it follows that $X_{1,*} \to Y_{1,*} \to Y_{2,*}$, and therefore $X_{1,*}|\{Y_{1,*}, Y_{2,*} = y_1, y_2\} \sim X_{1,*}|\{Y_{1,*} = y_1\}$ by conditional independence. Similarly, $X_{2,*}|\{Y_{1,*}, Y_{2,*} = y_1, y_2\} \sim X_{2,*}|\{Y_{2,*} = y_2\}$. So, we may conclude that the random variables $X_{1,*}|\{Y_{1,*} = y_1\}$ and $X_{2,*}|\{Y_{2,*} = y_2\}$ are normal, with identical variances not depending on $y_1, y_2$. Invoking Lemma 11, we find that both $X_{1,*}$ and $X_{2,*}$ are normal with identical variances, completing the proof. □

## Appendix B
## A Lower Semicontinuity Property of $\mathsf{s}_\lambda(\cdot, \varrho)$

The goal of this appendix is to establish the lower semicontinuity property of $\mathsf{s}_\lambda(\cdot, \varrho)$ as stated in Lemma 8. In particular, if $X_n \xrightarrow{\mathcal{D}} X_* \sim N(\mu, \sigma_X^2)$ and $\sup_n \mathbb{E}[X_n^2] < \infty$, then

$$\liminf_{n\to\infty} \mathsf{s}_\lambda(X_n, \varrho) \geq \mathsf{s}_\lambda(X_*, \varrho). \tag{72}$$

Recall that $\mathsf{s}_\lambda(X, \varrho)$ is defined in the context of the Gaussian channel $Y = \sqrt{\varrho}X + Z$. However, for the purposes of the proof, it will be convenient to omit the scaling factor $\varrho$, and instead parametrize the channel in terms of the noise variance. Toward this end, let $Z \sim N(0, \sigma^2)$. For $\lambda > 0$ and a random variable $X \sim P_X$, independent of $Z$, define $Y = X + Z$ and the functionals

$$\begin{aligned}\mathsf{F}_{\lambda,\sigma^2}(X) &= \inf_{V:X\to Y\to V} \Big(I(Y;V) - \lambda I(X;V)\Big)\\ \mathsf{G}_{\lambda,\sigma^2}(X) &= -h(X) + \lambda h(Y).\end{aligned}$$

The semicontinuity property (72) is an immediate corollary of weak lower semicontinuity of $\mathsf{G}_{\lambda,\sigma^2}(X)$ and $\mathsf{F}_{\lambda,\sigma^2}(X)$ at Gaussian $X$. These facts are established separately below in separate subsections. The former is straightforward, while the latter is a bit more delicate.

### A. Semicontinuity of $\mathsf{G}_{\lambda,\sigma^2}$

Here, we establish a semicontinuity property enjoyed by $\mathsf{G}_{\lambda,\sigma^2}$. Specifically,

*Lemma 13: If $X_n \xrightarrow{\mathcal{D}} X_* \sim N(\mu, \sigma_X^2)$ and $\sup_n \mathbb{E}[X_n^2] < \infty$, then*

$$\liminf_{n\to\infty} \mathsf{G}_{\lambda,\sigma^2}(X_n) \geq \mathsf{G}_{\lambda,\sigma^2}(X_*).$$

*Proof:* Fix $\delta > 0$ and define $N_\delta \sim N(0, \delta)$, pairwise independent of $\{X_n\}, X_*$. Observe that

$$\begin{aligned}\mathsf{G}_{\lambda,\sigma^2}(X_n) &= -h(X_n) + \lambda h(Y_n)\\ &\geq -h(X_n + N_\delta) + \lambda h(Y_n).\end{aligned}$$

By the third claim of Lemma 10, we have $-h(X_n + N_\delta) + \lambda h(Y_n) \to -h(X_* + N_\delta) + \lambda h(Y_*)$ as $n \to \infty$. Thus,

$$\liminf_{n\to\infty} \mathsf{G}_{\lambda,\sigma^2}(X_n) \geq -h(X_* + N_\delta) + \lambda h(Y_*).$$

Since $h(X_* + N_\delta) = \frac{1}{2}\log\big(2\pi e(\sigma_X^2 + \delta)\big)$ is continuous in $\delta$, we may take $\delta \downarrow 0$ to prove the claim. □

### B. Semicontinuity of $\mathsf{F}_{\lambda,\sigma^2}$

This section is devoted to proving the required semicontinuity property by $\mathsf{F}_{\lambda,\sigma^2}$. Specifically, we aim to show:

*Lemma 14: If $X_n \xrightarrow{\mathcal{D}} X_* \sim N(\mu, \sigma_X^2)$ and $\sup_n \mathbb{E}[X_n^2] < \infty$, then*

$$\liminf_{n\to\infty} \mathsf{F}_{\lambda,\sigma^2}(X_n) \geq \mathsf{F}_{\lambda,\sigma^2}(X_*). \tag{73}$$

The proof of Lemma 14 boils down to careful control over various error terms, which are ultimately shown to be negligible. To this end, we begin by establishing three technical estimates. These are then assembled together in the proof of Lemma 14.

*Lemma 15: Let $\{Y_n\}, Y_*$ be as in Lemma 10, and let $P_{V|Y}$ be given. Fix $b > 0$ and define $\{V_n\}, V_*$ according to $P_{V|Y} : Y_n \mapsto V_n$ and $P_{V|Y} : Y_* \mapsto V_*$. There exists a positive sequence $\{\epsilon_n\}$ depending on $b$ and $\{Y_n\}$, but not $P_{V|Y}$, which satisfies $\lim_{n\to\infty} \epsilon_n = 0$ and*

$$\begin{aligned}&I(V_n; Y_n \mid |Y_n| \leq b) \leq (1+\epsilon_n) I(V_*; Y_* \mid |Y_*| \leq b) + \epsilon_n,\\ &I(V_n; Y_n \mid |Y_n| \leq b) \geq (1-\epsilon_n) I(V_*; Y_* \mid |Y_*| \leq b) - \epsilon_n,\\ &\left|\frac{\mathbb{P}(|Y_n| \leq b)}{\mathbb{P}(|Y_*| \leq b)} - 1\right| \leq \epsilon_n.\end{aligned} \tag{74}$$

*Proof:* Let $f_n$ and $f_*$ denote the density of $Y_n$ and $Y_*$, respectively. By Lemma 9, the density $f_*$ is continuous and does not vanish, and is therefore bounded away from zero on the interval $B = [-b, b]$. By Lemma 10, $\|f_n - f_*\|_\infty \to 0$, so it follows that

$$\sup_{y\in B}\left|1 - \frac{f_n(y)}{f_*(y)}\right| \leq \epsilon_n \quad \text{and} \quad \sup_{y\in B}\left|1 - \frac{f_*(y)}{f_n(y)}\right| \leq \epsilon_n \tag{75}$$

for some positive $\epsilon_n \to 0$ as $n \to \infty$ (note that $\epsilon_n$ does not depend on $P_{V|Y}$). As a consequence,

$$\begin{aligned}\mathbb{P}(Y_* \in B) = \int_B f_*(y)dy &\leq (1+\epsilon_n)\int_B f_n(y)dy\\ &= (1+\epsilon_n)\mathbb{P}(Y_n \in B).\end{aligned}$$

By symmetry of (75), this establishes (74).

Now, for $y \in B$, the conditional densities of $Y_n|\{Y_n \in B\}$ and $Y_*|\{Y_* \in B\}$ satisfy

$$\frac{f_{Y_n|\{Y_n\in B\}}(y)}{f_{Y_*|\{Y_*\in B\}}(y)} = \frac{f_n(y)}{f_*(y)} \cdot \frac{\mathbb{P}(Y_* \in B)}{\mathbb{P}(Y_n \in B)} \leq (1+\epsilon_n)^2. \tag{76}$$

Therefore, for any Borel set[2] $A \subset \mathcal{V}$,

$$\begin{aligned}&\mathbb{P}(V_n \in A | Y_n \in B)\\ &\quad= \int_B f_{Y_n|\{Y_n\in B\}}(y) P_{V|Y=y}(A)dy\\ &\quad\leq (1+\epsilon_n)^2 \int_B f_{Y_*|\{Y_*\in B\}}(y) P_{V|Y=y}(A)dy\\ &\quad= (1+\epsilon_n)^2\, \mathbb{P}(V_* \in A | Y_* \in B).\end{aligned}$$

As a consequence,

$$\frac{dP_{V_n|Y_n\in B}}{dP_{V_*|Y_*\in B}}(v) \leq (1+\epsilon_n)^2, \quad v \in \mathcal{V}. \tag{77}$$

[2]We implicitly assume $P_{V|Y=y}$ is a Borel measure on a topological space $\mathcal{V}$ for each $y$.

By a symmetric argument, the following counterpart to (76) is obtained

$$f_{Y_n|\{Y_n\in B\}}(y) \geq (1+\epsilon_n)^{-2} f_{Y_*|\{Y_*\in B\}}(y) \quad \forall y \in B. \quad (78)$$

Using the fact that mutual information may be written as a relative entropy and $P_{V_n|Y_n} = P_{V|Y}$ for each $n$,

$$\begin{aligned} &I(V_n; Y_n|Y_n \in B) \\ &\quad = \int f_{Y_n|\{Y_n\in B\}}(y) D(P_{V_n|Y_n=y, Y_n\in B} \| P_{V_n|Y_n\in B}) dy \\ &\quad = \int f_{Y_n|\{Y_n\in B\}}(y) D(P_{V|Y=y} \| P_{V_n|Y_n\in B}) dy. \end{aligned}$$

Similarly, since $P_{V_*|Y_*} = P_{V|Y}$,

$$I(V_*; Y_*|Y_* \in B) = \int f_{Y_*|\{Y_*\in B\}}(y) D(P_{V|Y=y} \| P_{V_*|Y_*\in B}) dy.$$

Hence, if we can show that, for each $y$,

$$D(P_{V|Y=y} \| P_{V_n|Y_n\in B}) \geq D(P_{V|Y=y} \| P_{V_*|Y_*\in B}) - 3\epsilon_n,$$

then we may apply (78) and non-negativity of divergence to conclude

$$I(V_n; Y_n|Y_n \in B) \geq (1+\epsilon_n)^{-2} I(V_*; Y_*|Y_* \in B) - 3\epsilon_n. \quad (79)$$

Toward this end, let us recall the Donsker-Varadhan variational formula for relative entropy [83] between probability measures $P \ll Q$

$$D(P\|Q) = \sup_{\varphi} \left\{ \int \varphi dP - \log\left( \int e^{\varphi} dQ \right) \right\},$$

where the supremum is over all bounded $Q$-measurable functions $\varphi$. Applied to our situation, $P_{V|Y=y} \ll P_{V_*|Y_*\in B}$ for $y \in B$ (except possibly for $y$ in a null set, which we may neglect), so we are ensured the existence of a bounded function $\varphi_*$ for which

$$\int \varphi_* dP_{V|Y=y} - \log\left( \int e^{\varphi_*} dP_{V_*|Y_*\in B} \right) \geq D(P_{V|Y=y} \| P_{V_*|Y_*\in B}) - \epsilon_n.$$

Therefore, using the Donsker-Varadhan formula for $D(P_{V|Y=y} \| P_{V_n|Y_n\in B})$ and the estimate (77), we have

$$\begin{aligned} &D(P_{V|Y=y} \| P_{V_n|Y_n\in B}) \\ &\geq \int \varphi_* dP_{V|Y=y} - \log\left( \int e^{\varphi_*} dP_{V_n|Y_n\in B} \right) \\ &\geq \int \varphi_* dP_{V|Y=y} - \log\left( (1+\epsilon_n)^2 \int e^{\varphi_*} dP_{V_*|Y_*\in B} \right) \\ &= \int \varphi_* dP_{V|Y=y} - \log\left( \int e^{\varphi_*} dP_{V_*|Y_*\in B} \right) - \log(1+\epsilon_n)^2 \\ &\geq D(P_{V|Y=y} \| P_{V_*|Y_*\in B}) - 3\epsilon_n. \end{aligned}$$

The final step used the inequality $\log(1+x)^2 \leq 2x$ for purposes of simplifying the expression. Hence, (79) is proved. By symmetry of (75) and the above argument, we also have the complementary inequality

$$I(V_*; Y_*|Y_* \in B) \geq (1+\epsilon_n)^{-2} I(V_n; Y_n|Y_n \in B) - 3\epsilon_n.$$

It is now a simple matter to construct a new vanishing sequence $\{\epsilon_n'\}$ from $\{\epsilon_n\}$ which satisfies the claims of the lemma. □

*Lemma 16: Let $X \sim P_X$ and let $Z \sim N(0,\sigma^2)$ be a non-degenerate Gaussian, independent of $X$. It holds that*

$$\lim_{b\to\infty} \mathbb{P}(|X| > b) I(X; X+Z \mid |X| > b) = 0.$$

*Proof:* The proof follows that of [84, Theorem 6]. By lower semicontinuity of relative entropy, we have

$$\liminf_{b\to\infty} I(X; X+Z \mid |X| \leq b) \geq I(X; X+Z).$$

Also,

$$I(X; X+Z) \geq \mathbb{P}(|X| \leq b) I(X; X+Z \mid |X| \leq b),$$

so that

$$\begin{aligned} &\lim_{b\to\infty} I(X; X+Z \mid |X| \leq b) \\ &= \lim_{b\to\infty} \mathbb{P}(|X| \leq b) I(X; X+Z \mid |X| \leq b) = I(X; X+Z). \end{aligned}$$

By the chain rule for mutual information

$$\begin{aligned} &\mathbb{P}(|X| > b) I(X; X+Z \mid |X| > b) \\ &\quad = I(X; X+Z) - I(\mathbb{1}_{\{|X|\leq b\}}; X+Z) \\ &\qquad - \mathbb{P}(|X| \leq b) I(X; X+Z \mid |X| \leq b), \end{aligned}$$

so the claim is proved since $I(\mathbb{1}_{\{|X|\leq b\}}; X+Z)$ vanishes as $b \to \infty$. □

*Lemma 17: $\mathsf{F}_{\lambda,\sigma^2}(X)$ is continuous in $\lambda$. Furthermore, if $X \sim N(\mu, \sigma_X^2)$, then*

$$\mathsf{F}_{\lambda,\sigma^2}(X) = \frac{1}{2}\left[ \log\left( (\lambda-1)\frac{\sigma_X^2}{\sigma^2} \right) - \lambda \log\left( \frac{\lambda-1}{\lambda}\left( 1 + \frac{\sigma_X^2}{\sigma^2} \right) \right) \right]$$

*for $\lambda \geq 1 + \frac{\sigma^2}{\sigma_X^2}$, and $\mathsf{F}_{\lambda,\sigma^2}(X) = 0$ when $0 \leq \lambda \leq 1 + \frac{\sigma^2}{\sigma_X^2}$.*

*In particular, $\mathsf{F}_{\lambda,\sigma^2}(X)$ is continuous in the parameters $\sigma^2$, $\sigma_X^2$ and $\lambda$ for Gaussian $X$.*

*Proof:* The function $\mathsf{F}_{\lambda,\sigma^2}(X)$ is the pointwise infimum of linear functions in $\lambda$, and is therefore concave and continuous on the open interval $\lambda \in (0,\infty)$ for any distribution $P_X$. The explicit expression for $\mathsf{F}_{\lambda,\sigma^2}(X)$ follows by identifying $\gamma\ \varrho \leftarrow \frac{\sigma_X^2}{\sigma^2}$ in Lemma 3. □

We may now prove the desired result, which will conclude the proof of Lemma 8.

*Proof of Lemma 14:* Fix an interval $B = [-b, b]$, a conditional law $P_{V|Y}$, and $\delta$ satisfying $0 < \delta < \sigma^2/2$. Recalling the definition of $Z \sim N(0,\sigma^2)$, decompose $Z = N_1 + N_2 + N_3$, where $N_1 \sim N(0,\delta)$, $N_2 \sim N(0,\sigma^2 - 2\delta)$ and $N_3 \sim N(0,\delta)$ are mutually independent. Define $X_n^\delta = X_n + N_1$ and $Y_n^\delta = Y_n - N_3 = X_n + N_1 + N_2$. Note that we have $X_n \to X_n^\delta \to Y_n^\delta \to Y_n \to V_n$, where $V_n$ is defined by the stochastic transformation $P_{V|Y} : Y_n \mapsto V_n$. Using the notation of Lemma 10, we also have $X_* \to X_*^\delta \to Y_*^\delta \to Y_* \to V_*$, where $Y_* = X_* + Z$, $X_*^\delta = X_* + N_1$, $Y_*^\delta = X_* + N_1 + N_2$ and $V_*$ is defined via $P_{V|Y} : Y_* \mapsto V_*$. With these definitions in hand, we may apply Lemma 15 to the processes $\{X_n^\delta\}$, $\{Y_n^\delta\}$ to

conclude the existence of a sequence $\epsilon_n \to 0$, not depending on $P_{V|Y}$, that satisfies

$$I(V_n; X_n^\delta | X_n^\delta \in B) \le (1+\epsilon_n) I(V_*; X_*^\delta | X_*^\delta \in B) + \epsilon_n \quad (80)$$

$$I(V_n; Y_n^\delta | Y_n^\delta \in B) \ge (1-\epsilon_n) I(V_*; Y_*^\delta | Y_*^\delta \in B) - \epsilon_n \quad (81)$$

$$\mathbb{P}(X_n^\delta \in B) \le (1+\epsilon_n)\mathbb{P}(X_*^\delta \in B) \quad (82)$$

$$\mathbb{P}(Y_n^\delta \in B) \ge (1-\epsilon_n)\mathbb{P}(Y_*^\delta \in B). \quad (83)$$

Without loss of generality, we may assume that $\epsilon_n < 1$ for all $n$.

Having completed the setup, our goal at this point will be to obtain a lower bound on the quantity $I(Y_n; V_n) - \lambda I(X_n; V_n)$ which does not depend on the specific conditional law $P_{V|Y}$. The key idea will be to work with the perturbed random variables $X_n \to X_n^\delta$ and $Y_n \to Y_n^\delta$, and resist temptation to take any limits $n \to \infty$ or $b \to \infty$ until after dependence on $P_{V|Y}$ is eliminated. We begin with the following sequence of inequalities, each of whose steps are individually justified in the sequel:

$$\begin{aligned}
&I(Y_n; V_n) - \lambda I(X_n; V_n) \\
&\quad \ge I(Y_n^\delta; V_n) - \lambda I(X_n^\delta; V_n) && (84)\\
&\quad = I(Y_n^\delta, \mathbb{1}_{\{Y_n^\delta \in B\}}; V_n) - \lambda I(X_n^\delta, \mathbb{1}_{\{X_n^\delta \in B\}}; V_n) && (85)\\
&\quad = \mathbb{P}(Y_n^\delta \in B) I(Y_n^\delta; V_n | Y_n^\delta \in B) \\
&\qquad + \mathbb{P}(Y_n^\delta \notin B) I(Y_n^\delta; V_n | Y_n^\delta \notin B) + I(\mathbb{1}_{\{Y_n^\delta \in B\}}; V_n) \\
&\qquad - \lambda \Big( \mathbb{P}(X_n^\delta \in B) I(X_n^\delta; V_n | X_n^\delta \in B) \\
&\qquad + \mathbb{P}(X_n^\delta \notin B) I(X_n^\delta; V_n | X_n^\delta \notin B) + I(\mathbb{1}_{\{X_n^\delta \in B\}}; V_n) \Big) && (86)\\
&\quad \ge \mathbb{P}(Y_n^\delta \in B) I(Y_n^\delta; V_n | Y_n^\delta \in B) \\
&\qquad - \lambda \Big( \mathbb{P}(X_n^\delta \in B) I(X_n^\delta; V_n | X_n^\delta \in B) \\
&\qquad + \mathbb{P}(X_n^\delta \notin B) I(X_n^\delta; Y_n | X_n^\delta \notin B) + H(\mathbb{1}_{\{X_n^\delta \in B\}}) \Big) && (87)\\
&\quad \ge \mathbb{P}(Y_n^\delta \in B) \Big( (1-\epsilon_n) I(Y_*^\delta; V_* | Y_*^\delta \in B) - \epsilon_n \Big) \\
&\qquad - \lambda \mathbb{P}(X_n^\delta \in B) \Big( (1+\epsilon_n) I(X_*^\delta; V_* | X_*^\delta \in B) + \epsilon_n \Big) \\
&\qquad - \lambda \Big( P(X_n^\delta \notin B) I(X_n^\delta; Y_n | X_n^\delta \notin B) + H(\mathbb{1}_{\{X_n^\delta \in B\}}) \Big) && (88)\\
&\quad \ge \mathbb{P}(Y_n^\delta \in B)(1-\epsilon_n) I(Y_*^\delta; V_* | Y_*^\delta \in B) \\
&\qquad - \lambda \mathbb{P}(X_n^\delta \in B)(1+\epsilon_n) I(X_*^\delta; V_* | X_*^\delta \in B) \\
&\qquad - \lambda \Big( P(X_n^\delta \notin B) I(X_n^\delta; Y_n | X_n^\delta \notin B) + H(\mathbb{1}_{\{X_n^\delta \in B\}}) \Big) \\
&\qquad - (\lambda+1)\epsilon_n. && (89)
\end{aligned}$$

The above steps are justified as follows:

- (84) follows by the data processing inequality.
- (85) follows since $\mathbb{1}_{\{Y_n^\delta \in B\}}$ and $\mathbb{1}_{\{X_n^\delta \in B\}}$ are functions of $Y_n^\delta$ and $X_n^\delta$, respectively.
- (86) follows from the chain rule for mutual information.
- (87) follows from non-negativity of mutual information, the fact that $I(\mathbb{1}_{\{X_n^\delta \in B\}}; V_n) \le H(\mathbb{1}_{\{X_n^\delta \in B\}})$, and the data processing inequality which implies $I(X_n^\delta; V_n | X_n^\delta \notin B) \le I(X_n^\delta; Y_n | X_n^\delta \notin B)$.
- (88) follows from (80) and (81).
- (89) follows since $\mathbb{P}(Y_n^\delta \in B)$ and $\mathbb{P}(X_n^\delta \in B)$ are each at most one.

Now, let us separately bound the first two terms in (89). First, we note that the chain rule for mutual information implies

$$\begin{aligned}
\mathbb{P}(Y_*^\delta \in B) I(Y_*^\delta; V_* | Y_*^\delta \in B) &= I(Y_*^\delta; V_*) - I(\mathbb{1}_{\{Y_*^\delta \in B\}}; V_*) \\
&\quad - \mathbb{P}(Y_*^\delta \notin B) I(Y_*^\delta; V_* | Y_*^\delta \notin B).
\end{aligned}$$

So, the first term in (89) may be bounded as

$$\begin{aligned}
&\mathbb{P}(Y_n^\delta \in B)(1-\epsilon_n) I(Y_*^\delta; V_* | Y_*^\delta \in B) \\
&\quad = \frac{\mathbb{P}(Y_n^\delta \in B)}{\mathbb{P}(Y_*^\delta \in B)} (1-\epsilon_n) \Big( I(Y_*^\delta; V_*) - I(\mathbb{1}_{\{Y_*^\delta \in B\}}; V_*) \\
&\qquad - \mathbb{P}(Y_*^\delta \notin B) I(Y_*^\delta; V_* | Y_*^\delta \notin B) \Big) \\
&\quad \ge (1-\epsilon_n)^2 \Big( I(Y_*^\delta; V_*) - I(\mathbb{1}_{\{Y_*^\delta \in B\}}; V_*) \\
&\qquad - \mathbb{P}(Y_*^\delta \notin B) I(Y_*^\delta; V_* | Y_*^\delta \notin B) \Big) && (90)\\
&\quad \ge (1-\epsilon_n)^2 I(Y_*^\delta; V_*) \\
&\qquad - \Big( \mathbb{P}(Y_*^\delta \notin B) I(Y_*^\delta; Y_* | Y_*^\delta \notin B) + H(\mathbb{1}_{\{Y_*^\delta \in B\}}) \Big), && (91)
\end{aligned}$$

where

- (90) follows from (83) (note that the term inside the parentheses is proportional to $I(Y_*^\delta; V_* | Y_*^\delta \in B)$, and therefore nonnegative).
- (91) uses the simple facts that $I(\mathbb{1}_{\{Y_*^\delta \in B\}}; V_*) \le H(\mathbb{1}_{\{Y_*^\delta \in B\}})$, and $(1-\epsilon_n)^2 \le 1$, and

$$I(Y_*^\delta; V_* | Y_*^\delta \notin B) \le I(Y_*^\delta; Y_* | Y_*^\delta \notin B),$$

where the third is due to the data processing inequality.

Next, we bound the second term in (89). To start, note that the chain rule for mutual information combined with non-negativity of mutual information gives

$$I(X_*^\delta; V_* | X_*^\delta \in B) \le \frac{1}{\mathbb{P}(X_*^\delta \in B)} I(X_*^\delta; V_*).$$

Thus,

$$\begin{aligned}
&\lambda \mathbb{P}(X_n^\delta \in B)(1+\epsilon_n) I(X_*^\delta; V_* | X_*^\delta \in B) \\
&\quad \le \lambda \frac{\mathbb{P}(X_n^\delta \in B)}{\mathbb{P}(X_*^\delta \in B)} (1+\epsilon_n) I(X_*^\delta; V_*) \\
&\quad \le \lambda (1+\epsilon_n)^2 I(X_*^\delta; V_*),
\end{aligned}$$

where the last inequality is due to (82).

Combining the above two bounds with (89), we have established

$$\begin{aligned}
&I(Y_n; V_n) - \lambda I(X_n; V_n) \\
&\quad \ge (1-\epsilon_n)^2 I(Y_*^\delta; V_*) - \lambda (1+\epsilon_n)^2 I(X_*^\delta; V_*) \\
&\qquad - \Big( \mathbb{P}(Y_*^\delta \notin B) I(Y_*^\delta; Y_* | Y_*^\delta \notin B) + H(\mathbb{1}_{\{Y_*^\delta \in B\}}) \Big) \\
&\qquad - \lambda \Big( P(X_n^\delta \notin B) I(X_n^\delta; Y_n | X_n^\delta \notin B) + H(\mathbb{1}_{\{X_n^\delta \in B\}}) \Big) \\
&\qquad - (\lambda+1)\epsilon_n && (92)\\
&\quad = (1-\epsilon_n)^2 \Big( I(Y_*^\delta; V_*) - \lambda_n I(X_*^\delta; V_*) \Big) \\
&\qquad - \Big( \mathbb{P}(Y_*^\delta \notin B) I(Y_*^\delta; Y_* | Y_*^\delta \notin B) + H(\mathbb{1}_{\{Y_*^\delta \in B\}}) \Big)
\end{aligned}$$

$$- \lambda\Big(P(X_n^\delta \notin B) I(X_n^\delta; Y_n | X_n^\delta \notin B) + H(\mathbb{1}_{\{X_n^\delta \in B\}})\Big) - (\lambda+1)\epsilon_n, \tag{93}$$

$$\geq (1-\epsilon_n)^2 \mathsf{F}_{\lambda_n,(\sigma^2-2\delta)}(X_*^\delta) - \Big(\mathbb{P}(Y_*^\delta \notin B) I(Y_*^\delta; Y_* | Y_*^\delta \notin B) + H(\mathbb{1}_{\{Y_*^\delta \in B\}})\Big) - \lambda\Big(P(X_n^\delta \notin B) I(X_n^\delta; Y_n | X_n^\delta \notin B) + H(\mathbb{1}_{\{X_n^\delta \in B\}})\Big) - (\lambda+1)\epsilon_n. \tag{94}$$

Above, we make the definition $\lambda_n := \frac{(1+\epsilon_n)^2}{(1-\epsilon_n)^2}\lambda$ in (93) and used the definition of $\mathsf{F}_{\lambda_n,(\sigma^2-2\delta)}(X_*^\delta)$ in (94).

As desired, the RHS of (94) does not depend on $V_n$ (i.e., $P_{V|Y}$). Thus, taking the infimum over $V_n$ satisfying $X_n \to Y_n \to V_n$ and then letting $n \to \infty$, we arrive at

$$\liminf_{n\to\infty} \mathsf{F}_{\lambda,\sigma^2}(X_n) \geq \mathsf{F}_{\lambda,(\sigma^2-2\delta)}(X_*^\delta) - \Big(\mathbb{P}(Y_*^\delta \notin B) I(Y_*^\delta; Y_* | Y_*^\delta \notin B) + H(\mathbb{1}_{\{Y_*^\delta \in B\}})\Big) - \lambda\Big(\mathbb{P}(X_*^\delta \notin B) I(X_*^\delta; Y_* | X_*^\delta \notin B) + H(\mathbb{1}_{\{X_*^\delta \in B\}})\Big), \tag{95}$$

which follows due to $\epsilon_n \to 0$ and the following:

- $\mathsf{F}_{\lambda_n,(\sigma^2-2\delta)}(X_*^\delta) \to \mathsf{F}_{\lambda,(\sigma^2-2\delta)}(X_*^\delta)$ by continuity of the mapping $\lambda \mapsto F_{\lambda,\sigma^2}(X)$ (Lemma 17).
- $\mathbb{P}(X_n^\delta \notin B) \to \mathbb{P}(X_*^\delta \notin B)$ since $X_n^\delta \xrightarrow{\mathcal{D}} X_*^\delta$ by the first claim of Lemma 10. By the same token, $H(\mathbb{1}_{\{X_n^\delta \in B\}}) \to H(\mathbb{1}_{\{X_*^\delta \in B\}})$ by continuity of the binary entropy function.
- $I(X_n^\delta; Y_n | X_n^\delta \notin B) \to I(X_*^\delta; Y_* | X_*^\delta \notin B)$ by the third claim of Lemma 10 since $\limsup_n \mathbb{E}[\left(X_n^\delta\right)^2 | X_n^\delta \notin B] < \infty$ due to the fact that $\sup_n \mathbb{E}[X_n^2] < \infty$ and $\mathbb{P}(X_n^\delta \notin B) \to \mathbb{P}(X_*^\delta \notin B)$, a positive constant.

As we take $b \to \infty$, continuity of the binary entropy function and Lemma 16 together imply the latter two terms in the RHS of (95) vanish, yielding the inequality

$$\liminf_{n\to\infty} \mathsf{F}_\lambda(X_n) \geq \mathsf{F}_{\lambda,(\sigma^2-2\delta)}(X_*^\delta). \tag{96}$$

We finally arrive at the point where the hypothesis $X_* \sim N(0,\sigma^2)$ is needed. In particular, since $\delta$ was arbitrary and $\mathsf{F}_{\lambda,(\sigma^2-2\delta)}(X_*^\delta)$ is continuous in $\delta$ by Lemma 17, the proof is complete by letting $\delta \downarrow 0$. □

## ACKNOWLEDGMENT

The author acknowledges the Simons Institute for the Theory of Computing, where some of this research was completed

## REFERENCES


[1] T. A. Courtade, "Strengthening the entropy power inequality," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2016, pp. 2294–2298.

[2] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 4, pp. 623–656, Oct. 1948.

[3] A. J. Stam, "Some inequalities satisfied by the quantities of information of Fisher and Shannon," *Inf. Control*, vol. 2, no. 2, pp. 101–112, Jun. 1959.

[4] N. M. Blachman, "The convolution inequality for entropy powers," *IEEE Trans. Inf. Theory*, vol. IT-11, no. 2, pp. 267–271, Apr. 1965.

[5] E. Carlen and A. Soffer, "Entropy production by block variable summation and central limit theorems," *Commun. Math. Phys.*, vol. 140, no. 2, pp. 339–371, 1991.

[6] Y. Polyanskiy and Y. Wu, "Dissipation of information in channels with input constraints," *IEEE Trans. Inf. Theory*, vol. 62, no. 1, pp. 35–55, Jan. 2016.

[7] F. du Pin Calmon, Y. Polyanskiy, and Y. Wu, "Strong data processing inequalities in power-constrained Gaussian channels," in *Proc. ISIT*, Jun. 2015, pp. 2558–2562.

[8] G. Toscani, "A strengthened entropy power inequality for log-concave densities," *IEEE Trans. Inf. Theory*, vol. 61, no. 12, pp. 6550–6559, Dec. 2015.

[9] T. A. Courtade, M. Fathi, and A. Pananjady, "Wasserstein stability of the entropy power inequality for log-concave random vectors," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 659–663.

[10] M. Madiman and A. Barron, "Generalized entropy power inequalities and monotonicity properties of information," *IEEE Trans. Inf. Theory*, vol. 53, no. 7, pp. 2317–2329, Jul. 2007.

[11] S. Artstein, K. Ball, F. Barthe, and A. Naor, "Solution of Shannon's problem on the monotonicity of entropy," *J. Amer. Math. Soc.*, vol. 17, no. 4, pp. 975–982, 2004.

[12] M. Madiman, J. Melbourne, and P. Xu, "Forward and reverse entropy power inequalities in convex geometry," in *Convexity and Concentration* (The IMA Volumes in Mathematics and its Applications), vol. 161, E. Carlen, M. Madiman, and E. Werner, Eds. New York, NY, USA: Springer, 2017.

[13] M. H. M. Costa, "A new entropy power inequality," *IEEE Trans. Inf. Theory*, vol. IT-31, no. 6, pp. 751–760, Nov. 1985.

[14] R. Liu, T. Liu, H. V. Poor, and S. Shamai (Shitz), "A vector generalization of costa's entropy-power inequality with applications," *IEEE Trans. Inf. Theory*, vol. 56, no. 4, pp. 1865–1879, Apr. 2010.

[15] M. Costa, "On the Gaussian interference channel," *IEEE Trans. Inf. Theory*, vol. 31, no. 5, pp. 607–615, Sep. 1985.

[16] Y. Polyanskiy and Y. Wu, "Wasserstein continuity of entropy and outer bounds for interference channels," *IEEE Trans. Inf. Theory*, vol. 62, no. 7, pp. 3992–4002, Jul. 2016.

[17] G. Bagherikaram, A. S. Motahari, and A. K. Khandani, "The secrecy capacity region of the Gaussian MIMO broadcast channel," *IEEE Trans. Inf. Theory*, vol. 59, no. 5, pp. 2673–2682, May 2013.

[18] A. B. Wagner, S. Tavildar, and P. Viswanath, "Rate region of the quadratic Gaussian two-encoder source-coding problem," *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp. 1938–1961, May 2008.

[19] Y. Oohama, "Gaussian multiterminal source coding," *IEEE Trans. Inf. Theory*, vol. 43, no. 6, pp. 1912–1923, Nov. 1997.

[20] O. Rioul, "Information theoretic proofs of entropy power inequalities," *IEEE Trans. Inf. Theory*, vol. 57, no. 1, pp. 33–55, Jan. 2011.

[21] P. Bergmans, "A simple converse for broadcast channels with additive white Gaussian noise (corresp.)," *IEEE Trans. Inf. Theory*, vol. IT-20, no. 2, pp. 279–280, Mar. 1974.

[22] H. Weingarten, Y. Steinberg, and S. Shamai (Shitz), "The capacity region of the Gaussian multiple-input multiple-output broadcast channel," *IEEE Trans. Inf. Theory*, vol. 52, no. 9, pp. 3936–3964, Sep. 2006.

[23] M. Mohseni and J. M. Cioffi, "A proof of the converse for the capacity of Gaussian MIMO broadcast channels," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2006, pp. 881–885.

[24] S. Leung-Yan-Cheong and M. E. Hellman, "The Gaussian wire-tap channel," *IEEE Trans. Inf. Theory*, vol. IT-24, no. 4, pp. 451–456, Jul. 1978.

[25] E. Tekin and A. Yener, "The Gaussian multiple access wire-tap channel," *IEEE Trans. Inf. Theory*, vol. 54, no. 12, pp. 5747–5755, Dec. 2008.

[26] L. Ozarow, "On a source-coding problem with two channels and three receivers," *Bell Syst. Tech. J.*, vol. 59, no. 10, pp. 1909–1921, Dec. 1980.

[27] Y. Oohama, "Rate-distortion theory for Gaussian multiterminal source coding systems with several side informations at the decoder," *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2577–2593, Jul. 2005.

[28] V. Prabhakaran, D. Tse, and K. Ramachandran, "Rate region of the quadratic Gaussian CEO problem," in *Proc. Int. Symp. Inf. Theory (ISIT)*, Jun. 2004, p. 119.

[29] L. Gross, "Logarithmic Sobolev inequalities," *Amer. J. Math.*, vol. 97, no. 4, pp. 1061–1083, 1975.

[30] D. Bakry, I. Gentil, and M. Ledoux, *Analysis and Geometry of Markov Diffusion Operators*, vol. 348. Cham, Switzerland: Springer, 2013.

[31] M. Ledoux, *The Concentration of Measure Phenomenon* (Mathematical Surveys and Monographs). Providence, RI, USA: AMS, 2005, ch. 89.

[32] J. A. Thomas and T. M. Cover, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 2012.

[33] A. Dembo, "Simple proof of the concavity of the entropy power with respect to added Gaussian noise," *IEEE Trans. Inf. Theory*, vol. 35, no. 4, pp. 887–888, Jul. 1989.

[34] C. Villani, "A short proof of the 'concavity of entropy power,"' *IEEE Trans. Inf. Theory*, vol. 46, no. 4, pp. 1695–1696, Jul. 2000.
[35] T. A. Courtade, "Concavity of entropy power: Equivalent formulations and generalizations," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2017, pp. 56–60.
[36] T. A. Courtade, G. Han, and Y. Wu. (2017). "A counterexample to the vector generalization of costa's EPI, and partial resolution." [Online]. Available: https://arxiv.org/abs/1704.06164
[37] M. Madiman, "On the entropy of sums," in *Proc. IEEE Inf. Theory Workshop*, May 2008, pp. 303–307.
[38] M. Costa and T. Cover, "On the similarity of the entropy power inequality and the Brunn-Minkowski inequality (corresp.)," *IEEE Trans. Inf. Theory*, vol. IT-30, no. 6, pp. 837–839, Nov. 1984.
[39] S. Bobkov and M. Madiman, "Reverse Brunn-Minkowski and reverse entropy power inequalities for convex measures," *J. Funct. Anal.*, vol. 262, no. 7, pp. 3309–3339, 2012.
[40] V. D. Milman, "Inégalité de Brunn-Minkowski inverse et applicationsa la théorie locale des espaces normés," *CR Acad. Sci. Paris*, vol. 302, no. 1, pp. 25–28, 1986.
[41] K. Ball, F. Barthe, and A. Naor, "Entropy jumps in the presence of a spectral gap," *Duke Math. J.*, vol. 119, no. 1, pp. 41–63, 2003.
[42] O. Johnson and A. Barron, "Fisher information inequalities and the central limit theorem," *Probab. Theory Rel. Fields*, vol. 129, no. 3, pp. 391–409, Jul. 2004.
[43] O. Johnson, *Information Theory and the Central Limit Theorem*. Singapore: World Scientific, 2004.
[44] K. Ball and V. H. Nguyen, "Entropy jumps for isotropic log-concave random vectors and spectral gap," *Studia Math.*, vol. 213, no. 1, pp. 81–96, 2012.
[45] E. Nelson, "The free Markoff field," *J. Funct. Anal.*, vol. 12, no. 2, pp. 211–227, 1973.
[46] E. A. Carlen, "Superadditivity of Fisher's information and logarithmic Sobolev inequalities," *J. Funct. Anal.*, vol. 101, no. 1, pp. 194–211, 1991.
[47] M. Raginsky and I. Sason, "Concentration of measure inequalities in information theory, communications, and coding," *Found. Trends Commun. Inf. Theory*, vol. 10, nos. 1–2, pp. 1–247, 2013.
[48] S. G. Bobkov, N. Gozlan, C. Roberto, and P.-M. Samson, "Bounds on the deficit in the logarithmic Sobolev inequality," *J. Funct. Anal.*, vol. 267, no. 11, pp. 4110–4138, 2014.
[49] M. Fathi, E. Indrei, and M. Ledoux, "Quantitative logarithmic Sobolev inequalities and stability estimates," *Discrete Continuous Dyn. Syst., A*, vol. 36, no. 12, pp. 6835–6853, 2016.
[50] J. Dolbeault and G. Toscani, "Stability results for logarithmic Sobolev and Gagliardo–Nirenberg inequalities," *Int. Math. Res. Notices*, vol. 2016, no. 2, pp. 473–498, Jan. 2015.
[51] T. A. Courtade, M. Fathi, and A. Pananjady. (2017). "Existence of Stein kernels under a spectral gap, and discrepancy bounds." [Online]. Available: https://arxiv.org/abs/1703.07707
[52] A. El Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2012.
[53] T. Berger, "Multiterminal source coding," in *The Information Theory Approach to Communications*, G. Longo, Ed. New York, NY, USA: Springer-Verlag, 1977.
[54] S.-Y. Tung, "Multiterminal source coding," Ph.D. dissertation, Dept. Elect. Eng., Cornell Univ., Ithaca, NY, USA, 1978.
[55] J. Wang, J. Chen, and X. Wu, "On the sum rate of Gaussian multiterminal source coding: New proofs and results," *IEEE Trans. Inf. Theory*, vol. 56, no. 8, pp. 3946–3960, Aug. 2010.
[56] T. A. Courtade and J. Jiao, "An extremal inequality for long Markov chains," in *Proc. 52nd Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Sep. 2014, pp. 763–770.
[57] T. A. Courtade and T. Weissman, "Multiterminal source coding under logarithmic loss," *IEEE Trans. Inf. Theory*, vol. 60, no. 1, pp. 740–761, Jan. 2014.
[58] J. Jiao, T. A. Courtade, K. Venkat, and T. Weissman, "Justification of logarithmic loss via the benefit of side information," *IEEE Trans. Inf. Theory*, vol. 61, no. 10, pp. 5357–5365, Oct. 2015.
[59] H. Sato, "The capacity of the Gaussian interference channel under strong interference (corresp.)," *IEEE Trans. Inf. Theory*, vol. IT-27, no. 6, pp. 786–788, Nov. 1981.
[60] T. Han and K. Kobayashi, "A new achievable rate region for the interference channel," *IEEE Trans. Inf. Theory*, vol. IT-27, no. 1, pp. 49–60, Jan. 1981.
[61] M. H. M. Costa, "Noisebergs in Z-Gaussian interference channels," in *Proc. Inf. Theory Appl. Workshop (ITA)*, Feb. 2011, pp. 1–6.
[62] C. Nair and M. H. Costa, "Gaussian Z-interference channel: Around the corner," in *Proc. Inf. Theory Appl. Workshop (ITA)*, 2016, pp. 1–6.
[63] R. Ahlswede and P. Gács, "Spreading of sets in product spaces and hypercontraction of the Markov operator," *Ann. Probab.*, vol. 4, no. 6, pp. 925–939, Dec. 1976.
[64] C. Nair, "Equivalent formulations of hypercontractivity using information measures," in *Proc. Int. Zurich Seminar Commun.*, 2014, p. 42.
[65] V. Anantharam, A. A. Gohari, S. Kamath, and C. Nair, "On hypercontractivity and the mutual information between Boolean functions," in *Proc. Allerton*, 2013, pp. 13–19.
[66] M. Raginsky, "Strong data processing inequalities and Φ-Sobolev inequalities for discrete channels," *IEEE Trans. Inf. Theory*, vol. 62, no. 6, pp. 3355–3389, Jun. 2016.
[67] V. Anantharam, A. Gohari, S. Kamath, and C. Nair. (2013). "On maximal correlation, hypercontractivity, and the data processing inequality studied by Erkip and Cover." [Online]. Available: https://arxiv.org/abs/1304.6133
[68] T. A. Courtade, "Outer bounds for multiterminal source coding via a strong data processing inequality," in *Proc. IEEE Int. Symp. Inf. Theory Proc. (ISIT)*, Jul. 2013, pp. 559–563.
[69] C. Nair, "Upper concave envelopes and auxiliary random variables," *Int. J. Adv. Eng. Sci. Appl. Math.*, vol. 5, no. 1, pp. 12–20, 2013.
[70] Y. Geng and C. Nair, "The capacity region of the two-receiver Gaussian vector broadcast channel with private and common messages," *IEEE Trans. Inf. Theory*, vol. 60, no. 4, pp. 2087–2104, Apr. 2014.
[71] C. Nair, L. Xia, and M. Yazdanpanah, "Sub-optimality of Han-Kobayashi achievable region for interference channels," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2015, pp. 2416–2420.
[72] H. G. Eggleston, *Convexity* (Cambridge Tracts in Mathematics and Mathematical Physics), no. 47. Cambridge, U.K.: Cambridge Univ. Press, 1958.
[73] S. N. Bernstein, "On a property characteristic of the normal law," *Trudy Leningrad. Polytech. Inst.*, vol. 3, pp. 21–22, 1941.
[74] W. Bryc, *The Normal Distribution: Characterizations With Applications*, vol. 100. New York, NY, USA: Springer-Verlag, 2012.
[75] Y. Polyanskiy, "Channel coding: Non-asymptotic fundamental limits," Ph.D. dissertation, Dept. Elect. Eng., Princeton Univ., Princeton, NJ, USA: 2010.
[76] E. H. Lieb, "Gaussian kernels have only Gaussian maximizers," *Invent. Math.*, vol. 102, no. 1, pp. 179–208, 1990.
[77] F. Barthe, "Optimal Young's inequality and its converse: A simple proof," *Geometric Funct. Anal.*, vol. 8, no. 2, pp. 234–242, 1998.
[78] E. A. Carlen and D. Cordero-Erausquin, "Subadditivity of the entropy and its relation to Brascamp–Lieb type inequalities," *Geometric Funct. Anal.*, vol. 19, no. 2, pp. 373–405, 2009.
[79] 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.
[80] 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.
[81] S. Kamath, "Reverse hypercontractivity using information measures," in *Proc. 53rd Annu. Allerton Conf. Commun, Control, Comput. (Allerton)*, Sep. 2015, pp. 627–633.
[82] R. Durrett, *Probability: Theory and Examples*. Cambridge, U.K.: Cambridge Univ. Press, 2010.
[83] M. D. Donsker and S. R. S. Varadhan, "Asymptotic evaluation of certain Markov process expectations for large time, I," *Commun. Pure Appl. Math.*, vol. 28, no. 1, pp. 1–47, 1975.
[84] Y. Wu and S. Verdú, "Functional properties of minimum mean-square error and mutual information," *IEEE Trans. Inf. Theory*, vol. 58, no. 3, pp. 1289–1301, Mar. 2012.

**Thomas A. Courtade** (S'06–M'13) received the B.Sc. degree (summa cum laude) in electrical engineering from Michigan Technological University, Houghton, MI, USA, in 2007, and the M.S. and Ph.D. degrees from the University of California, Los Angeles (UCLA), CA, USA, in 2008 and 2012, respectively. He is an Assistant Professor with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA, USA. Prior to joining UC Berkeley in 2014, he was a Postdoctoral Fellow supported by the NSF Center for Science of Information.

Prof. Courtade's honors include a Distinguished Ph.D. Dissertation Award and an Excellence in Teaching Award from the UCLA Department of Electrical Engineering, and a Jack Keil Wolf Student Paper Award for the 2012 International Symposium on Information Theory. He is the recipient of a Hellman Fellowship and a NSF CAREER Award.