Strong Data Processing Inequalities in Power-Constrained Gaussian Channels

Flavio P. Calmon, Yury Polyanskiy, Yihong Wu

***Abstract*—This work presents strong data processing results for the power-constrained additive Gaussian channel. Explicit bounds on the amount of decrease of mutual information under convolution with Gaussian noise are shown. The analysis leverages the connection between information and estimation (I-MMSE) and the following estimation-theoretic result of independent interest. It is proved that any random variable for which there exists an almost optimal (in terms of the mean-squared error) linear estimator operating on the Gaussian-corrupted measurement must necessarily be almost Gaussian (in terms of the Kolmogorov-Smirnov distance).**

I. Introduction

Strong data-processing inequalities quantify the decrease of mutual information under the action of a noisy channel. Such inequalities have apparently been first discovered by Ahlswede and Gács in a landmark paper [1]. Among the work predating [1] and extending it we mention [1]–[5]. Notable connections include topics ranging from existence and uniqueness of Gibbs measures and log-Sobolev inequalities to performance limits of noisy circuits. We refer the reader to the introduction in [6] and the recent monographs [7], [8] for more detailed discussions of applications and extensions. Below we only review the necessary minimum to set the stage for our work.

For a fixed channel $P_{Y|X}: \mathcal{X} \to \mathcal{Y}$, let $P_{Y|X} \circ P$ be the distribution on $\mathcal{Y}$ induced by the push-forward of the distribution P. One approach to strong data processing seeks to find the contraction coefficients

$$\eta_f \triangleq \sup_{P,Q:P\neq Q} \frac{D_f\left(P_{Y|X}\circ P \| P_{Y|X}\circ Q\right)}{D_f(P\|Q)}, \qquad (1)$$

where the $D_f(P\|Q)$ is an arbitrary f-divergence of Csiszár [9]. When the divergence D_f is the KL-divergence and total variation[1], we denote the coefficient η_f as η_{KL} and η_{TV}, respectively.

For discrete channels, [1] showed equivalence of $\eta_{\mathrm{KL}} < 1$, $\eta_{\mathsf{TV}} < 1$ and connectedness of the bipartite graph describing the channel. Having $\eta_{\mathrm{KL}} < 1$ implies reduction in the usual data-processing inequality for mutual information [10, Exercise III.2.12], [11]:

$$\forall\, W \to X \to Y : I(W;Y) \le \eta_{\mathrm{KL}} \cdot I(W;X)\,.$$

When $P_{Y|X}$ is an additive white Gaussian noise channel, i.e. $Y = X+Z$ with $Z \sim \mathcal{N}(0,1)$, the authors showed [6] that restricting maximization in (1) to distributions with a bounded second moment (or any moment) still leads to no-contraction, giving $\eta_{\mathrm{KL}} = \eta_{\mathsf{TV}} = 1$ for AWGN. Nevertheless, the contraction does indeed take place, except not multiplicatively. Namely [6] found the region

$$\left\{(\mathsf{TV}(P,Q), \mathsf{TV}(P*P_Z, Q*P_Z)) : \mathbb{E}_{(P+Q)/2}[X^2] \le \gamma\right\},$$

where $*$ denotes convolution. The boundary of this region, deemed the *Dobrushin curve* of the channel, turned out to be strictly bounded away from the diagonal (identity). In other words, except for the trivial case where $\mathsf{TV}(P,Q) = 0$, total variation decreases by a non-trivial amount in Gaussian channels.

Unfortunately, the similar region for KL-divergence turns out to be trivial, so that no improvement in the inequality

$$D(P_X * P_Z \| Q_Z * P_Z) \le D(P_X\|Q_X)$$

is possible (given the knowledge of the right-hand side and moment constraints on P_X and Q_X). In [6], in order to study how mutual information dissipates on a chain of Gaussian links, this problem was resolved by a rather lengthy workaround which entails first reducing questions regarding the mutual information to those about the total variation and then converting back.

A more direct approach, in the spirit of the joint-range idea of Harremoës and Vajda [12], is to find (or bound) the *best possible data-processing function* F_I defined as follows.

Definition 1. Let $Y_\gamma = \sqrt{\gamma}X + Z$, where $Z \sim \mathcal{N}(0,1)$ is independent of X. We define

$$F_I(t,\gamma) \triangleq \sup\left\{I(W;Y_\gamma) \colon I(W;X) \le t, W \to X \to Y_\gamma\right\}, \qquad (2)$$

where the supremum is over all joint distributions $P_{W,X}$ such that $\mathbb{E}\left[X^2\right] \le 1$.

The significance of the function F_I is that it gives the optimal input-independent strong data processing inequality on Gaussian channel:

$$I(W;Y_\gamma) \le F_I(I(W;X),\gamma).$$

Before discussing properties of F_I, we mention two related quantities considered previously in the literature. Witsenhausen and Wyner [13] defined

$$F_H(P_{XY}, h) = \inf H(Y|W), \qquad (3)$$

with the infimum taken over all joint distributions satisfying

$$W \to X \to Y, H(X|W) = h, \mathbb{P}[X=x, Y=y] = P_{XY}(x,y)\,.$$

Clearly, by a simple reparametrization $h = H(X) - t$, this function would correspond to $H(Y) - F_I(t)$ if $F_I(t)$ were defined with restriction to a given input distribution P_X. The P_X-independent version of (3) has also been studied by

F. P. Calmon and Y. Polyanskiy are with the Department of EECS, MIT, Cambridge, MA, 02139, USA. E-mail:{flavio,yp}@mit.edu. Y. Wu is with the Department of ECE and the Coordinated Science Lab, University of Illinois at Urbana-Champaign, Urbana, IL, 61801, USA. E-mail: yihongwu@illinois.edu.

This work is supported in part by the National Science Foundation (NSF) CAREER award under Grant CCF-12-53205, the NSF Grant IIS-1447879 and CCF-1423088 and by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under Grant CCF-09-39370.

[1]The total variation between two distributions P and Q is $\mathsf{TV}(P,Q) \triangleq \sup_E |P[E] - Q[E]|$.

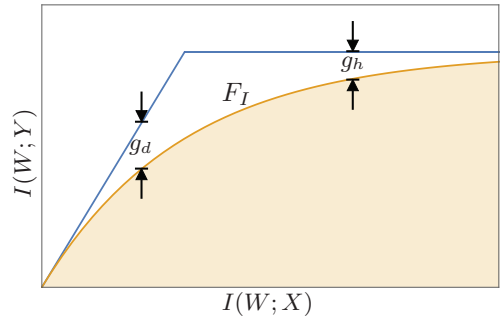

Fig. 1. The strong data processing function F_I and gaps g_d and g_h to the trivial data processing bound (4).

Witsenhausen [14]:

$$F_H(P_{Y|X}, h) = \inf H(Y|W),$$

with the infimum taken over all

$$W \to X \to Y, H(X|W) = h, \mathbb{P}[Y = y|X = x] = P_{Y|X}(y|x)\,.$$

This quantity plays a role in a generalization of Mrs. Gerber's lemma and satisfies a convenient tensorization property:

$$F_H((P_{Y|X})^n, nh) = nF(P_{Y|X}, h)\,.$$

There is no one-to-one correspondence between $F_H(P_{Y|X}, h)$ and $F_I(t)$ and in fact, alas, $F_I(t)$ does not satisfy any (known to us) tensorization property.

Apriori, the only bounds we can state on F_I are consequences of capacity and the data processing inequality:

$$F_I(t,\gamma) \le \min\{t, C(\gamma)\}\,, \tag{4}$$

where $C(\gamma) = \frac{1}{2}\ln(1+\gamma)$ is the Gaussian channel capacity. Recently, we were able to show (for any noise distribution P_Z subject to natural regularity conditions) the following two inequalities hold [15]

$$F_I(t,\gamma) \le t - g_d(t,\gamma), \tag{5}$$
$$F_I(t,\gamma) \le C(\gamma) - g_h(t,\gamma) \tag{6}$$

with strictly positive g_d and g_h. See Fig. 1 for an illustration. Extracting explicit expressions for g_d and g_h from [15] appears tedious, however.

In this work, we treat the special case of the Gaussian noise and derive explicit asymptotically sharp estimates showing that $F_I(t,\gamma)$ is strictly bounded away from the trivial (4). To that end, we leverage Fourier-analytic tools and methods specific to Gaussian distributions, namely, Talagrand's transportation inequality [16] and information-estimation connection [17].

Specifically, Theorem 1 provides a lower bound for the function $g_d(t,\gamma)$ defined in (5), which is asymptotically tight as $t \to 0$:

$$g_d(t,\gamma) = e^{-\frac{\gamma}{t}\ln\frac{1}{t} + \Theta(\ln\frac{1}{t})}. \tag{7}$$

A repeated application of (5) shows that the mutual information between the input X_0 and the output Y_n of the chain of n energy-constrained Gaussian relays converges to zero $I(X_0;Y_n) \to 0$. In fact, (7) recovers the convergence rate of $O(\frac{\log\log n}{\log n})$ first reported in [6, Theorem 1].

We also characterize the asymptotic behaviour of $F_I(t,\gamma)$ approaching $C(\gamma)$ as $t \to \infty$, which turns out to be *double-exponential*. In Theorem 2 and Remark 4, we prove that

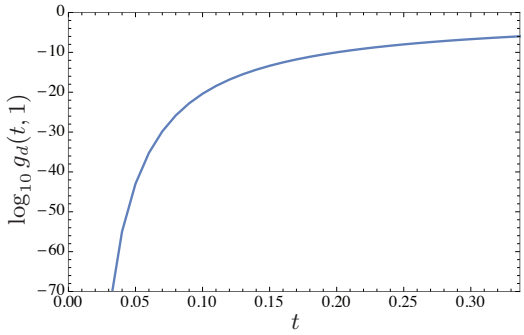

Fig. 2. Lower bound on $g_d(t,1)$ derived in Theorem 1.

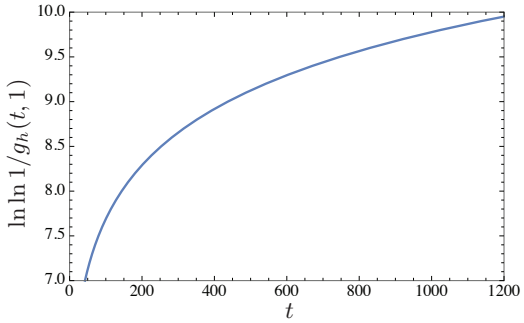

Fig. 3. Lower bound on $g_h(t,1)$ derived in Theorem 2.

$g_h(t,\gamma)$, defined in (6), satisfies

$$e^{-c_1(\gamma)e^{4t}} \le g_h(t,\gamma) \le e^{-c_2(\gamma)e^t + O(\ln\gamma)} \tag{8}$$

as $t \to \infty$, where $c_1(\gamma)$ and $c_2(\gamma)$ are strictly positive functions of γ. The lower bounds for g_d and g_h are illustrated in Fig. 2 and 3.

In order to bound $g_h(t,\gamma)$ from below, we obtain two ancillary results that are of independent interest. Lemma 1 shows that if the linear estimator of X given Y_γ is near-optimal in terms of the mean squared error, then X is almost Gaussian in terms of Kolmogorov-Smirnov (KS) distance. By applying the I-MMSE relationship, this result is then used to prove that if $I(X;Y_\gamma)$ is close to $C(\gamma)$, then X is also almost Gaussian in terms of the KS-distance (Lemma 2).

The rest of the paper is organized as follows. Section II presents a lower bound for $g_d(t,\gamma)$. Section III describes explicit upper bounds for the KS-distance between the distribution of X and $\mathcal{N}(0,1)$ when (i) the linear estimator of X given Y_γ performs almost as well as the mmse estimator, and (ii) $I(X;Y_\gamma)$ is close to $C(\gamma)$. These results are then used in Section IV to lower bound $g_h(t,\gamma)$. Finally, in Section V we consider the infinite-dimensional discrete Gaussian channel, and show that in this case there exists no non-trivial strong data processing inequality for mutual information.

II. Diagonal Bound

In this section we show that $F_I(t)$ is bounded away from t for all $t > 0$ (Theorem 1) and investigate the behaviour of $F_I(t)$ for small t (Corollary 1).

Theorem 1. *For $t \ge 0$, $F_I(t,\gamma) = t - g_d(t,\gamma)$, where*

$$g_d(t,\gamma) \ge \max_{x\in[0,1/2]} 2Q\left(\sqrt{\frac{\gamma}{x}}\right)\left(t - h(x) - \frac{x}{2}\ln\left(1+\frac{\gamma}{x}\right)\right), \tag{9}$$

$h(x) \triangleq x\ln\frac{1}{x} + (1-x)\ln\frac{1}{1-x}$ and $Q(x) \triangleq \frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-y^2/2}\mathrm{d}y$.

Proof. Let $E = \mathbf{1}_{\{|X|>A/\sqrt{\gamma}\}}$ and $\mathbb{E}[E] = p$. Observe that

$$\mathbb{E}\left[\gamma X^2 | E=1\right] \le \gamma/p \quad \text{and} \quad p \le \gamma/A^2. \tag{10}$$

Therefore, for $\bar{p} \triangleq 1-p$,

$$\begin{aligned} I(W;Y) \le & I(W;E) + pI(W;Y|E=1) + \bar{p}I(W;Y|E=0) \\ \le & I(W;E) + pI(W;Y|E=1) \\ & + \bar{p}\eta(A) I(W;X|E=0), \end{aligned} \tag{11}$$

where the last inequality follows from [6], and $\eta(t) = 1 - 2Q(t)$. Noting that

$$\bar{p}I(W;X|E=0) = I(W;X) - pI(W;X|E=1) - I(W;E),$$

we can further bound (11) by

$$\begin{aligned} I(W;Y) \le & (1-\eta(A))I(W;E) + \eta(A)I(W;X) \\ & + p\eta(A)\left(I(W;Y|E=1) - I(W;X|E=1)\right) \\ & + p(1-\eta(A))I(W;Y|E=1) \\ \le & (1-\eta(A))\left(I(W;E) + pI(W;Y|E=1)\right) \\ & + \eta(A)I(W;X) \qquad (12) \\ = & I(W;X) - (1-\eta(A))\left(I(W;X)\right. \\ & \left. - I(W;E) - pI(W;Y|E=1)\right), \qquad (13) \end{aligned}$$

where (12) follows from $I(W;Y|E=1) \le I(W;X|E=1)$. Now observe that, for $p = \gamma/A^2 \le 1/2$,

$$I(W;E) \le H(E) \le h\left(\gamma/A^2\right). \tag{14}$$

In addition,

$$\begin{aligned} pI(W;Y|E=1) \le & pI(X;Y|E=1) \\ \le & \frac{p}{2}\ln\left(1+\frac{\gamma}{p}\right) \qquad (15) \\ \le & \frac{\gamma}{2A^2}\ln(1+A^2). \qquad (16) \end{aligned}$$

Here (15) follows from the fact that mutual information is maximized when X is Gaussian under the power constraint (10), and (16) follows by noticing that $x \mapsto x\ln(1+a/x)$ is monotonically increasing for any $a > 0$. Combining (14) and (16), and for $A \ge \sqrt{2\gamma}$,

$$I(W;E) + pI(W;Y|E=1) \le h\left(\frac{\gamma}{A^2}\right) + \frac{\gamma}{2A^2}\ln\left(A^2+1\right). \tag{17}$$

Choosing $A = \sqrt{\gamma/x}$, where $0 \le x \le 1/2$, (17) becomes

$$I(W;E) + pI(W;Y|E=1) \le h(x) + \frac{x}{2}\ln\left(1+\frac{\gamma}{x}\right). \tag{18}$$

Substituting (18) in (13) yields the desired result. □

Remark 1. Note that $f_d(x,\gamma) \triangleq h(x) + \frac{x}{2}\ln\left(1+\frac{\gamma}{x}\right)$ is 0 when $x = 0$ and is continuous and strictly positive for $0 < x \le 1/2$. Then $g_d(t,\gamma)$ is strictly positive for $t > 0$. The next corollary characterizes the behaviour of $g_d(t,\gamma)$ for small t.

Corollary 1. *For fixed γ, $t = 1/u$ and u sufficiently large, there is a constant $c_3(\gamma) > 0$ dependent on γ such that*

$$g_d(1/u,\gamma) \ge \frac{c_3(\gamma)}{u\sqrt{u\gamma\ln u}} e^{-\gamma u \ln u}. \tag{19}$$

In particular, $g_d(1/u,\gamma) \ge e^{-\gamma u \ln u + O(\ln \gamma u^{3/2})}$.

Remark 2. Fix γ and define a binary random variable X with $\mathbb{P}[X=a] = 1/a^2$ and $\mathbb{P}[X=0] = 1 - 1/a^2$ for $a > 0$. Furthermore, let $\hat{X}$ denote the minimum distance estimate of X produced from Y_γ. Then the probability of error satisfies $P_e = \mathbb{P}[X \ne \hat{X}] \le Q(\sqrt{\gamma}a/2)$. In addition, $h\left(Q(\sqrt{\gamma}a/2)\right) = O(e^{-\gamma a^2/8}\sqrt{\gamma}a)$ and $H(X) = a^{-2}\ln a(2+o(1))$ as $a \to \infty$. Therefore,

$$h\left(Q(\sqrt{\gamma}a/2)\right) \le e^{-\frac{\gamma}{H(X)}\ln\frac{1}{H(X)} + O(\ln(\gamma/H(X)))}. \tag{20}$$

Using Fano's inequality, $I(X;Y_\gamma)$ can be bounded as

$$\begin{aligned} I(X;Y_\gamma) \ge & I(X;\hat{X}) \\ \ge & H(X) - h(P_e) \\ \ge & H(X) - h\left(Q(\sqrt{\gamma}a/2)\right) \\ = & H(X) - e^{-\frac{\gamma}{H(X)}\ln\frac{1}{H(X)} + O(\ln(\gamma/H(X)))}. \end{aligned}$$

Setting $W = X$, this result yields the sharp asymptotics (7).

III. MMSE

We now show that if the linear least-square error of estimating X from Y_γ is small (i.e. close to the minimum mean-squared error), then X must be almost Gaussian in terms of the KS-distance. With this result in hand, we use the I-MMSE relationship [17] to show that if $I(X;Y_\gamma)$ is close to $C(\gamma)$, then X is also almost Gaussian. This result, in turn, will be applied in the next section to bound $F_I(t,\gamma)$ aways from $C(\gamma)$.

Denote the linear least-square error estimator of X given Y_γ by $f_L(y) \triangleq \sqrt{\gamma}y/(1+\gamma)$, whose mean squared error is

$$\mathsf{lmmse}(X|Y_\gamma) \triangleq \mathbb{E}\left[(X - f_L(Y_\gamma))^2\right] = \frac{1}{1+\gamma}.$$

Assume that $\mathsf{lmmse}(X|Y_\gamma) - \mathsf{mmse}(X|Y_\gamma) \le \epsilon$. It is well known that $\epsilon = 0$ if and only if $X \sim \mathcal{N}(0,1)$ (e.g. [18]). To develop a finitary version of this result, we ask the following question: If ϵ is small, how close is P_X to Gaussian? The next lemma provides a quantitative answer.

Lemma 1. *If $\mathsf{lmmse}(X|Y_\gamma) - \mathsf{mmse}(X|Y_\gamma) \le \epsilon$, then there are absolute constants a_0 and a_1 such that*

$$\begin{aligned} d_{\mathrm{KS}}(F_X, \mathcal{N}(0,1)) \le & a_0\sqrt{\frac{1}{\gamma\log(1/\epsilon)}} \\ & + a_1(1+\gamma)\epsilon^{1/4}\sqrt{\gamma\log(1/\epsilon)}, \end{aligned} \tag{21}$$

where F_X is the CDF of X, and d_{KS} is the Kolmogorov-Smirnov distance, defined as $d_{\mathrm{KS}}(F_1,F_2) \triangleq \sup_{x\in\mathbb{R}}|F_1(x) - F_2(x)|$.

Remark 3. Note that the gap between the linear and non-linear MMSE can be expressed as the Fisher distance between the convolutions, i.e., $\mathsf{lmmse}(X|Y_\gamma) - \mathsf{mmse}(X|Y_\gamma) = I(P_{Y_\gamma}\|N(0,1+\gamma))$, where $I(P\|Q) = \int[(\log\frac{\mathrm{d}P}{\mathrm{d}Q})']^2\mathrm{d}P$ is the Fisher distance, which is very strong and dominates the KL divergence according to the log-Sobolev inequality. Therefore Lemma 1 can be interpreted as a deconvolution result, where bounds on a stronger (Fisher) distance of the convolutions lead to bounds on the distance between the original distributions under a weaker (KS) metric.

Proof. Denote $f_M(y) = \mathbb{E}[X|Y_\gamma = y]$. Then

$$\begin{aligned}&\mathsf{lmmse}(X|Y_\gamma) - \mathsf{mmse}(X|Y_\gamma)\\ =&\mathbb{E}\left[(X - f_L(Y_\gamma))^2\right] - \mathbb{E}\left[(X - f_M(Y_\gamma))^2\right]\\ =&\mathbb{E}\left[(f_M(Y_\gamma) - f_L(Y_\gamma))^2\right] \le \epsilon. \end{aligned} \tag{22}$$

Denote $\Delta(y) \triangleq f_M(y) - f_L(y)$. Then $\mathbb{E}[\Delta(Y_\gamma)] = 0$ and $\mathbb{E}\left[\Delta(Y_\gamma)^2\right] \le \epsilon$. From the orthogonality principle:

$$\mathbb{E}\left[e^{itY_\gamma}(X - f_M(Y_\gamma))\right] = 0. \tag{23}$$

Let φ_X denote the characteristic function of X. Then

$$\begin{aligned}&\mathbb{E}\left[e^{itY_\gamma}(X - f_M(Y_\gamma))\right] = \mathbb{E}\left[e^{itY_\gamma}(X - f_L(Y_\gamma) - \Delta(Y_\gamma))\right]\\ =&\frac{1}{1+\gamma}\left(e^{-t^2/2}\mathbb{E}\left[e^{i\sqrt{\gamma}tX}X\right] - \sqrt{\gamma}\varphi_X(\sqrt{\gamma}t)\mathbb{E}\left[Ze^{itZ}\right]\right)\\ &- \mathbb{E}\left[e^{itY_\gamma}\Delta(Y_\gamma)\right]\\ =&\frac{-ie^{-u^2/2\gamma}}{1+\gamma}\left(\varphi_X'(u) + u\varphi_X(u)\right) - \mathbb{E}\left[e^{itY_\gamma}\Delta(Y_\gamma)\right], \end{aligned} \tag{24}$$

where the last equality follows by changing variables $u = \sqrt{\gamma}t$. Consequently,

$$\frac{e^{-u^2/2\gamma}}{1+\gamma}\left|\varphi_X'(u) + u\varphi_X(u)\right| = \left|\mathbb{E}\left[e^{itY_\gamma}\Delta(Y_\gamma)\right]\right| \tag{25}$$

$$\le \mathbb{E}\left[|\Delta(Y_\gamma)|\right] \le \sqrt{\epsilon}. \tag{26}$$

Put $\phi_X(u) = e^{-u^2/2}(1 + z(u))$. Then

$$|\varphi_X'(u) + u\varphi_X(u)| = e^{-u^2/2}|z'(u)|,$$

and, from (26), $|z'(u)| \le (1+\gamma)\sqrt{\epsilon}e^{\frac{u^2(\gamma+1)}{2\gamma}}$. Since $z(0) = 0$,

$$|z(u)| \le \int_0^u |z'(x)|dx \le u(1+\gamma)\sqrt{\epsilon}e^{\frac{u^2(\gamma+1)}{2\gamma}}. \tag{27}$$

Observe that $|\varphi_X(u) - e^{-u^2/2}| = e^{-u^2/2}|z(u)|$. Then, from (27),

$$\left|\frac{\varphi_X(u) - e^{-u^2/2}}{u}\right| \le (1+\gamma)\sqrt{\epsilon}e^{\frac{u^2}{2\gamma}}. \tag{28}$$

Thus the Esseen inequality (cf. [19, Eq. (3.13), pg. 512]) yields

$$\begin{aligned}d_{\mathrm{KS}}(F_X, \mathcal{N}(0,1)) &\le \frac{1}{\pi}\int_{-T}^{T}(1+\gamma)\sqrt{\epsilon}e^{\frac{u^2}{2\gamma}}du + \frac{12\sqrt{2}}{\pi^{3/2}T}\\ &\le \frac{2T}{\pi}(1+\gamma)\sqrt{\epsilon}e^{\frac{T^2}{2\gamma}} + \frac{12\sqrt{2}}{\pi^{3/2}T}.\end{aligned}$$

Choosing $T = \sqrt{\frac{\gamma}{2}\ln(\frac{1}{\epsilon})}$, we find

$$\begin{aligned}d_{\mathrm{KS}}(F_X, \mathcal{N}(0,1)) \le& a_0\sqrt{\frac{1}{\gamma\ln(1/\epsilon)}}\\ &+ a_1(1+\gamma)\epsilon^{1/4}\sqrt{\gamma\ln(1/\epsilon)},\end{aligned}$$

where $a_0 = \frac{24}{\pi^{3/2}}$ and $a_1 = \frac{\sqrt{2}}{\pi}$. □

Through the I-MMSE relationship [17], the previous lemma can be extended to bound the KS-distance between the distribution of X and the Gaussian distribution when $I(X;Y_\gamma)$ is close to $C(\gamma)$.

Lemma 2. *Let $C(\gamma) - I(X;Y_\gamma) \le \epsilon$. Then, for $\gamma > 4\epsilon$,*

$$\begin{aligned}d_{\mathrm{KS}}(F_X, \mathcal{N}(0,1)) \le& a_0\sqrt{\frac{2}{\gamma\ln\left(\frac{\gamma}{4\epsilon}\right)}}\\ &+ a_1(1+\gamma)(\gamma\epsilon)^{1/4}\sqrt{2\ln\left(\frac{\gamma}{4\epsilon}\right)}.\end{aligned} \tag{29}$$

IV. Horizontal Bound

In this section we show that $F_I(t,\gamma)$ is bounded away from the capacity $C(\gamma)$ for all t. In particular, Theorem 2 proves that if $C(\gamma) - F_I(t,\gamma) \le \epsilon$, then $t = \Omega(\ln\ln 1/\epsilon)$ as $\epsilon \to 0$. We first give an auxiliary lemma.

Lemma 3. *If $D(\mathcal{N}(0,1)\|P_X * \mathcal{N}(0,1)) \le 2\epsilon$, then there exists an absolute constant $a_2 > 0$ such that*

$$\mathbb{P}[|X| > \epsilon^{1/8}] \le a_2\epsilon^{1/8}. \tag{30}$$

Theorem 2. *Let $C(\gamma) - F_I(t,\gamma) \le \epsilon$. Then*

$$t \ge \frac{1}{4}\ln\ln\frac{1}{\epsilon} - \ln c_1(\gamma), \tag{31}$$

where $c_1(\gamma)$ is some constant depending on γ. In particular,

$$F_I(t,\gamma) = C(\gamma) - g_h(t,\gamma), \tag{32}$$

where $g_h(t,\gamma) \ge e^{-c_1(\gamma)e^{4t}}$.

Proof. Let $C(\gamma) - I(W;Y_\gamma) \le \epsilon$. Observe that

$$\begin{aligned}I(W;Y_\gamma) =& C(\gamma) - D(P_{\sqrt{\gamma}X} * \mathcal{N}(0,1)\|\mathcal{N}(0,1+\gamma))\\ &- I(X;Y_\gamma|W).\end{aligned} \tag{33}$$

Therefore, if $I(W;Y_\gamma)$ is close to $C(\gamma)$, then (a) P_X needs to be Gaussian like, and (b) $P_{X|W}$ needs to be almost deterministic with high P_W-probability. Consequently, $P_{X|W}$ and P_X are close to being mutually singular and hence $I(W;X)$ will be large, since

$$I(W;X) = D(P_{X|W}\|P_X|P_W).$$

Let $\tilde{X} \triangleq \sqrt{\gamma}X$ and then $W \to \tilde{X} \to Y$. Define

$$\begin{aligned}d(x,w) &\triangleq D(P_{Y|\tilde{X}=x}\|P_{Y|W=w})\\ &= D(\mathcal{N}(x,1)\|P_{\tilde{X}|W=w} * \mathcal{N}(0,1)).\end{aligned}$$

Then $(x,w) \mapsto d(x,w)$ is jointly measurable[2] and $I(X;Y|W) = \mathbb{E}[d(\tilde{X},W)]$. Similarly, $w \mapsto \tau(w) \triangleq D(P_{X|W=w}\|P_X)$ is measurable and $I(X;W) = \mathbb{E}[\tau(W)]$. Since $\epsilon \ge I(X;Y|W)$ in view of (33), we have

$$\epsilon \ge \mathbb{E}[d(\tilde{X},W)] \ge 2\epsilon \cdot \mathbb{P}[d(\tilde{X},W) \ge 2\epsilon]. \tag{34}$$

Therefore

$$\mathbb{P}[d(\tilde{X},W) < 2\epsilon] > \frac{1}{2}. \tag{35}$$

Denote $B(x,\delta) \triangleq [x-\delta, x+\delta]$. In view of Lemma 3, if

[2]By definition of the Markov kernel, $x \mapsto P_{Y\in A|\tilde{X}=x}$ and $w \mapsto P_{Y\in A|W=w}$ are both measurable for any measurable subset A. Since Y is real-valued, by data processing and lower semicontinuity of divergence, we have $D(P_{[Y]_k|\tilde{X}=x}\|P_{[Y]_k|W=w}) \to D(P_{Y|\tilde{X}=x}\|P_{Y|W=w})$ as $k \to \infty$, where $[y]_k = \lfloor ky\rfloor/k$ denotes the uniform quantizer. Therefore the joint measurability of $(x,w) \mapsto D(P_{Y|\tilde{X}=x}\|P_{Y|W=w})$ follows from that of $(x,w) \mapsto D(P_{[Y]_k|\tilde{X}=x}\|P_{[Y]_k|W=w})$.

$d(x,w) < 2\epsilon$, then

$$\begin{aligned}&\mathbb{P}[\tilde{X} \in B(x,\epsilon^{1/8})|W=w]\\ =&\mathbb{P}\left[X \in B\left(\frac{x}{\sqrt{\gamma}},\frac{\epsilon^{1/8}}{\sqrt{\gamma}}\right)\middle|W=w\right] \geq 1-a_2\epsilon^{1/8}.\end{aligned}$$

Therefore, with probability at least $1/2$, $\tilde{X}$ and, consequently, X is concentrated on a small ball. Furthermore, Lemma 2 implies that there exist absolute constants a_3 and a_4 such that

$$\begin{aligned}&\mathbb{P}\left[X \in B\left(\frac{x}{\sqrt{\gamma}},\frac{\epsilon^{1/8}}{\sqrt{\gamma}}\right)\right]\\ &\leq \mathbb{P}\left[Z \in B\left(\frac{x}{\sqrt{\gamma}},\frac{\epsilon^{1/8}}{\sqrt{\gamma}}\right)\right] + 2d_{\mathrm{KS}}(F_X,\mathcal{N}(0,1))\\ &\leq \frac{\sqrt{2}\epsilon^{1/8}}{\sqrt{\pi\gamma}} + a_3\sqrt{\frac{1}{\gamma\ln\left(\frac{\gamma}{4\epsilon}\right)}} + a_4(1+\gamma)(\gamma\epsilon)^{1/4}\sqrt{\ln\left(\frac{\gamma}{4\epsilon}\right)}\\ &\leq \kappa(\gamma)\left(\ln\frac{1}{\epsilon}\right)^{-1/2},\end{aligned}$$

where $\kappa(\gamma)$ is some positive constant depending only on γ. Therefore, for any $w \in \mathcal{B}$ and ϵ sufficiently small, denoting $E = B(\frac{x}{\sqrt{\gamma}},\frac{\epsilon^{1/8}}{\sqrt{\gamma}})$, we have by data processing inequality:

$$\begin{aligned}\tau(w) = D(P_{X|W=w}\|P_X) \geq& P_{X|W=w}(E)\ln\frac{P_{X|W=w}(E)}{P_X(E)}\\ &+ P_{X|W=w}(E^c)\ln\frac{P_{X|W=w}(E^c)}{P_X(E^c)}\\ \geq& \frac{1}{2}\ln\ln\frac{1}{\epsilon} - \ln\kappa(\gamma) - a_5, \quad (36)\end{aligned}$$

where a_5 is an absolute positive constant. Combining (36) with (35) and letting $c_1^2(\gamma) \triangleq e^{a_5}\kappa(\gamma)$, we obtain

$$\mathbb{P}\left[\tau(W) \geq \frac{1}{2}\ln\ln\frac{1}{\epsilon} - 2\ln c_1(\gamma)\right] \geq \mathbb{P}[d(\tilde{X},W) < 2\epsilon] \geq \frac{1}{2},$$

which implies that $I(W;X) = \mathbb{E}[\tau(W)] \geq \frac{1}{4}\ln\ln\frac{1}{\epsilon} - \ln c_1(\gamma)$, proving the desired (31). □

Remark 4. The double-exponential convergence rate in Theorem 2 is in fact sharp. To see this, note that [20, Theorem 8] showed that there exists a sequence of zero-mean and unit-variance random variables X_m with m atoms, such that

$$C(\gamma) - I(X_m;\sqrt{\gamma}X_m + Z) \leq 4(1+\gamma)\left(\frac{\gamma}{1+\gamma}\right)^{2m}. \quad (37)$$

Consequently,

$$\begin{aligned}C(\gamma) - F_I(t,\gamma) &\leq C(\gamma) - F_I(\ln\lfloor e^t\rfloor,\gamma)\\ &\leq 4(1+\gamma)\left(\frac{\gamma}{1+\gamma}\right)^{2(e^t-1)}\\ &= e^{-2e^t\ln\frac{1+\gamma}{\gamma}+O(\ln\gamma)},\end{aligned}$$

proving the right-hand side of (8).

V. Dimension greater than 1

It is possible to reproduce the techniques above for the case when the channel $X \to Y$ is a d-dimensional Gaussian channel subject to a total-energy constraint $\mathbb{E}\left[\sum_i X_i^2\right] \leq 1$. Unfortunately, the resulting bound has strong dependence on dimension and in particular does not improve the trivial estimate (4) as $d \to \infty$. It turns out this dependence is unavoidable as we show next.

To that end we consider an infinite-dimension discrete-time Gaussian channel. Here the input $X = (X_1, X_2, \dots)$ and $Y = (Y_1, Y_2, \dots)$ are sequences, where $Y_i = X_i + Z_i$ and $Z_i \sim \mathcal{N}(0,1)$ are i.i.d. Similar to Definition 1, we denote

$$F_I^\infty(t,\gamma) = \sup\{I(W;Y)\colon I(W;X) \leq t, W \to X \to Y\}, \quad (38)$$

where the supremum is over all P_{WX} such that $\mathbb{E}\left[\|X\|_2^2\right] = \mathbb{E}\left[\sum X_i^2\right] \leq \gamma$. Note that, in this case, $F_I^\infty(t,\gamma) \leq \min\{t,\gamma/2\}$. The next theorem shows that unlike in the scalar case, there is no improvement over the trivial upper bound (38) in the infinite-dimensional case. This is in stark contrast with the strong data processing behavior of total variation in Gaussian noise which turns out to be dimension-independent [6, Corollary 6].

Theorem 3. *For $0 \leq t \leq \gamma/2$, $F_I^\infty(t,\gamma) = t$.*

References

[1] R. Ahlswede and P. Gács, "Spreading of sets in product spaces and hypercontraction of the Markov operator," *The Annals of Probability*, pp. 925–939, 1976.

[2] R. Dobrushin, "Central limit theorem for nonstationary Markov chains. I," *Theory of Probab Appl.*, vol. 1, no. 1, pp. 65–80, Jan. 1956.

[3] O. Sarmanov, "Maximum correlation coefficient (nonsymmetric case)," *Selected Translations in Mathematical Statistics and Probability*, vol. 2, pp. 207–210, 1962.

[4] J. Cohen, Y. Iwasa, G. Rautu, M. Ruskai, E. Seneta, and G. Zbaganu, "Relative entropy under mappings by stochastic matrices," *Linear algebra and its applications*, vol. 179, pp. 211–235, 1993.

[5] R. Subramanian, B. Vellambi, and I. Land, "An improved bound on information loss due to finite block length in a Gaussian line network," in *Proc. 2013 IEEE Int. Symp. on Inf. Theory*, Jul. 2013, pp. 1864–1868.

[6] Y. Polyanskiy and Y. Wu, "Dissipation of information in channels with input constraints," *arXiv:1405.3629 [cs, math]*, May 2014.

[7] M. Raginsky, "Strong data processing inequalities and ϕ-Sobolev inequalities for discrete channels," *arXiv:1411.3575 [cs, math]*, Nov. 2014.

[8] M. Raginsky and I. Sason, "Concentration of measure inequalities in information theory, communications, and coding," *Found. and Trends in Comm. and Inf. Theory*, vol. 10, no. 1-2, pp. 1–247, 2013.

[9] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observation," *Studia Sci. Math. Hungar.*, vol. 2, pp. 229–318, 1967.

[10] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1981.

[11] V. Anantharam, A. Gohari, S. Kamath, and C. Nair, "On maximal correlation, hypercontractivity, and the data processing inequality studied by Erkip and Cover," *arXiv preprint arXiv:1304.6133*, 2013.

[12] P. Harremoës and I. Vajda, "On pairs of-divergences and their joint range," *IEEE Trans. Inform. Theory*, vol. 57, no. 6, pp. 3230–3235, 2011.

[13] H. Witsenhausen and A. Wyner, "A conditional entropy bound for a pair of discrete random variables," *IEEE Trans. Inform. Theory*, vol. 21, no. 5, pp. 493–501, Sep. 1975.

[14] H. Witsenhausen, "Entropy inequalities for discrete channels," *IEEE Trans. Inform. Theory*, vol. 20, no. 5, pp. 610–616, Sep. 1974.

[15] Y. Polyanskiy and Y. Wu, "Strong data-processing of mutual information: beyond Ahlswede and Gács," in *Proc. Information Theory and Applications Workshop*, 2015.

[16] M. Talagrand, "Transportation cost for Gaussian and other product measures," *Geom. Funct. Anal.*, vol. 6, no. 3, pp. 587–600, May 1996.

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

[18] D. Guo, Y. Wu, S. Shamai, and S. Verdú, "Estimation in Gaussian Noise: Properties of the Minimum Mean-Square Error," *IEEE Trans. Inf. Theory*, vol. 57, no. 4, pp. 2371–2385, Apr. 2011.

[19] W. Feller, *An Introduction to Probability Theory and Its Applications, Vol. 2*, 1st ed. New York: Wiley, 1966.

[20] Y. Wu and S. Verdú, "The impact of constellation cardinality on Gaussian channel capacity," in *Proc. 48th Annual Allerton Conf. on Commun., Control, and Comput.*, Sep. 2010, pp. 620–628.