On Some Almost Properties

Olivier Rioul
LTCI CNRS, Télécom ParisTech
Université Paris-Saclay, 75013 Paris, France
Email: olivier.rioul@telecom-paristech.fr

Max H. M. Costa
School of Electrical and Computer Engineering
University of Campinas – Unicamp, Brazil
max@fee.unicamp.br

***Abstract*—Previous works have shown that regular distributions with differential entropy or mean-squared error behavior close to that of the Gaussian are also close to the Gaussian with respect to some distances like Kolmogorov-Smirnov or Wasserstein distances, or vice versa. In keeping with these results, we show that under the assumption of a functional dependence on the Gaussian, any regular distribution that is almost Gaussian in differential entropy has a mean-squared error behavior of an almost linear estimator. A partial converse result is established under the addition of an arbitrary independent quantity: a small mean-squared error yields a small entropy difference. The proofs use basic properties of Shannon's information measures and can be employed in an alternative solution to the missing corner point problem of Gaussian interference channels.**

I. Introduction

Throughout the paper we consider random vectors of dimension n and let $|\cdot|$ denote the Euclidean norm in $\mathbb{R}^n$. Let $X \in \mathbb{R}^n$ be an random vector with finite second moments and differentiable density p_X, and X^G be its Gaussian counterpart of the same covariance matrix as X. Loosely speaking, we say that X is *almost Gaussian* if X is close to X^G in distribution. The precise definition depends on the criterion used to evaluate the "distance" between the two distributions P_X, P_{X^G}.

There are two well-known informational distances. The differential entropy difference $h(X^G) - h(X)$ where $h(X) = h(p_X) = -\mathbb{E}\log p_X(X)$ coincides with the Kullback-Leibler (KL) "distance" or divergence $D_{\text{KL}}(X\|X^G)$ where

$$D_{\text{KL}}(X\|Y) = D_{\text{KL}}(P_X\|P_Y) = \mathbb{E}\big\{\log \frac{dP_X}{dP_Y}(X)\big\}. \quad (1)$$

Similarly the Fisher information difference $J(X) - J(X^G)$ where $J(X) = J(p_X) = \mathbb{E}\big\{|\nabla \log p_X(X)|^2\big\}$ coincides with the Fisher information distance $D_{\text{F}}(X\|X^G)$ where

$$D_{\text{F}}(X\|Y) = D_{\text{F}}(P_X\|P_Y) = \mathbb{E}\big\{\big|\nabla \log \frac{dP_X}{dP_Y}(X)\big|^2\big\}. \quad (2)$$

Many other distances can be also considered, such as total variation $D_{\text{TV}}(X,Y) = \|p_X - p_Y\|_1$ (the L^1 norm of the difference of pdfs), Kolmogorov-Smirnov $D_{\text{KS}}(X,Y) = \|P_X - P_Y\|_\infty$ (the L^∞ norm of the difference of cdfs), and (L^2) Wasserstein distance $D_W^2(X,Y) = \inf \mathbb{E}\{|X-Y|^2\}$ where the infimum is taken over all joint distributions of (X,Y) with the given marginals $X \sim P_X$ and $Y \sim P_Y$.

That two distributions P_X, P_Y are "close" in distance D—i.e., $D(X,Y)$ is "small"—can be more precisely defined depending on the application. A useful definition for solving multi-user information-theoretic problems is an upper bound of the form $D(X,Y) \le n\epsilon(n)$ where $\epsilon(n) > 0$ is a sequence that tends to 0 as the dimension n increases.

A general problem is the determination of all situations in which X being close to X^G with respect to some distance(s) carries over under the addition of another random variable Y (or a Gaussian random variable Z) independent of (X, X^G), possibly with respect to another distance, or vice versa.

We shall not attempt to develop a general theory here but will focus on some simple derivations. We first review some known results in the area in Section II. As a preliminary, Section III then shows that there exists an appropriate function F such that if X is almost Gaussian, then making the corresponding change of variable $F(X)$ is truly Gaussian. In Section IV we show that if X is almost Gaussian in differential entropy, then in a certain sense X and X^G are almost linearly dependent and F is almost linear, the corresponding mean-squared error being necessarily small. Section V establishes a partial converse under the addition of an arbitrary independent random quantity. Finally, Section VI applies these results to the two-user Z-interference channel by deriving a new simple solution to the missing corner point problem.

II. Previous Results

A. Well-Known Results

We first review three related results that come from well-known inequalities in Shannon's information theory.

If X is almost Gaussian, that is, close to X^G (in KL distance), then $X+Y$ is close to X^G+Y (in KL distance). This is an immediate consequence of the *Data Processing Inequality* (DPI)

$$D_{\text{KL}}(X+Y\|X^G+Y) \le D_{\text{KL}}(X\|X^G) \quad (3)$$

which holds similarly for Fisher and Wasserstein distances.

Another related result is that if X, Y are independent and both almost Gaussian (in KL distance), then their sum $X+Y$ is also almost Gaussian (in KL distance). This was shown in [1] to be an immediate consequence of the *Entropy Power Inequality* (EPI):

$$e^{(2/n)h(X+Y)} \ge e^{(2/n)h(X)} + e^{(2/n)h(Y)}. \quad (4)$$

Notice that when $Y = Z$ is Gaussian, the stated result reduces to the fact that if X is almost Gaussian, then so is $X+Z$—the same result as the one above for the DPI. In this case, the DPI appears to be a (strictly) weaker form of the EPI.

Interestingly, in the determination of Sato's corner point of the two-user Gaussian weak Z-interference channel as was made in [1], the stronger form (EPI) is in fact not necessary as the same conclusion easily follows from the DPI applied to the Kullback-Leibler divergence.

Finally, if X and Y are close in KL distance, then they are also close in squared total variation distance by *Pinsker's inequality*

$$D_{\mathrm{TV}}^2(X,Y) \le 2D_{\mathrm{KL}}(X\|Y). \tag{5}$$

This was used by Costa [2] along with the concavity of the entropy power in his determination of the other corner point of the two-user Gaussian weak Z-interference channel which considers the maximal rate that an interfering signal may have when the interfered link operates at maximal rate. As detected by Sason [3], it turns out that Pinsker's inequality was not strong enough to settle the problem, which has since been known as the "missing" corner point problem [4].

B. Lesser Known Results

The following results come from transportation-information inequalities known in optimal transport theory (see e.g., [5], [6] for reviews).

The (Gaussian) *logarithmic Sobolev inequality*[1]

$$D_{\mathrm{KL}}(X\|X^G) \le c \cdot D_{\mathrm{F}}(X\|X^G) \tag{6}$$

implies that if X is almost Gaussian in Fisher distance, than it is also almost Gaussian in KL distance.

Talagrand's inequality

$$D_{\mathrm{W}}^2(X,X^G) \le c \cdot D_{\mathrm{KL}}(X\|X^G), \tag{7}$$

implies that if X is almost Gaussian in KL distance, then it is also almost Gaussian in squared Wasserstein distance. This is one of the ingredients used by Polyanskiy and Wu in [7].

Conversely, if X is close to Y in squared Wasserstein distance, then under the addition an independent Gaussian Z, $X+Z$ is also close to $Y+Z$ in KL distance. This is a consequence of the inequality [8]

$$D_{\mathrm{KL}}(X+Z\|Y+Z) \le \frac{1}{2} D_{\mathrm{W}}^2(X,Y) \tag{8}$$

which can be used to show the following *HWI inequality*[2]:

$$D_{\mathrm{KL}}(X\|X^G) \le \sqrt{D_{\mathrm{W}}^2(X,X^G) D_F(X\|X^G)}, \tag{9}$$

which in turn implies that if X is almost Gaussian both in squared Wasserstein distance and in Fisher information distance, then it is also almost Gaussian in KL distance.

[1]The notation c stands for some universal constant which depends only on the covariance matrix of X^G.

[2]H is for (relative) entropy (i.e., KL divergence), W is for Wasserstein and I is for (Fisher) Information.

C. Recent Results

We mention two recent results.

Calmon *et al.* [9] have shown that if X is almost Gaussian in the sense that the linear minimum mean-squared error (LMMSE) is close to the (non linear) MMSE when estimating X from a noisy observation $X+Z$ (the output of an AWGN channel), then X is also almost Gaussian in Kolmogorov-Smirnov (KS) distance. It is perhaps worthwhile to note that this result can be interpreted as follows. Define

$$\begin{aligned} \mathsf{mmse}(X|X+Z) &= \min_f \mathbb{E}\{|X-f(X+Z)|^2\} \\ \mathsf{lmmse}(X|X+Z) &= \min_{f \text{ linear}} \mathbb{E}\{|X-f(X+Z)|^2\} \end{aligned} \tag{10}$$

As noticed e.g. by Rioul [10], [11] one has the following identity which states that the MMSE and Fisher information are complementary quantities:

$$\mathsf{mmse}(X|X+Z) + J(X+Z) = n. \tag{11}$$

Since X and X^G have identical covariance matrices, it follows that $\mathsf{lmmse}(X|X+Z) = \mathsf{lmmse}(X^G|X^G+Z) = \mathsf{mmse}(X^G|X^G+Z) = n - J(X^G+Z)$. Hence the MMSE difference:

$$\mathsf{lmmse}(X|X+Z) - \mathsf{mmse}(X|X+Z) = J(X+Z) - J(X^G+Z) \tag{12}$$

is identical to the Fisher distance $D_{\mathrm{F}}(X+Z\|X^G+Z)$. Thus the result of Calmon *et al.* can be rewritten as follows: if $X+Z$ is almost Gaussian in Fisher distance, then X is almost Gaussian in KS distance. In this statement the Fisher distance can be replaced by the KL distance by integration using the I-MMSE method [9] or de Bruijn's identity.

Another recent result of Polyanskiy and Wu [7] is that under some regularity conditions on p_X and p_Y, if X and Y are close in squared Wasserstein distance, then they are also close in KL distance:

$$D_{\mathrm{KL}}(X\|Y) \le c \cdot \sqrt{n \cdot D_{\mathrm{W}}^2(X,Y)}. \tag{13}$$

The entropy difference $h(Y)-h(X)$ is also shown to be small in this case. This is the key result used by Polyanskiy and Wu [7] to settle the missing corner point problem.

In the sequel we shall not use the above inequalities and results but only derive a few simple related facts.

III. From Not Gaussian to Gaussian

To simplify the following derivations we assume without loss of generality that the considered random vectors have zero mean. We also assume that the covariance matrix X^G is proportional to the identity matrix: $X^G \sim \mathcal{N}(0, P\mathbf{I}_n)$ where P is the average power of X.

In order to compare the distributions of X and X^G, it is immaterial whether or not X and X^G are independent of each other. We find it convenient to assume a functional dependence of the form

$$X = F(X^G). \tag{14}$$

This is made possible by the following "not Gaussian to Gaussian" lemma, by means of an invertible transformation.

Lemma 1 (Not Gaussian to Gaussian)**.** *There exists a diffeomorphism F of the form*

$$\begin{cases} y_1 = F_1(x_1) \\ y_2 = F_2(x_1, x_2) \\ \quad \vdots \\ y_n = F_n(x_1, x_2, \ldots, x_n) \end{cases} \tag{15}$$

where for all $1 \le k \le n$ and any fixed value of $x^{k-1} = (x_1, \ldots, x_{k-1})$,

$$x_k \mapsto F(x_1, x_2, \ldots, x_k) \text{ is nondecreasing,} \tag{16}$$

such that $F(X^G)$ has the same distribution as X.

Proof: We first prove that we can choose F satisfying (15)–(16) such that $F(X)$ is uniformly distributed in $[0,1]^n$. We proceed by induction on n. For $n = 1$ this is a well-known result (which is at the basis of the inverse transform sampling method): Take

$$F(x) = \mathbb{P}(X \le x) \tag{17}$$

be the cdf of X. Clearly F is nondecreasing differentiable and for any $u \in [0,1]$,

$$\mathbb{P}(F(X) \le u) = u \tag{18}$$

so that $F(X)$ is uniformly distributed in $[0,1]$. Now suppose that $F^{n-1}(X^{n-1})$ is uniformly distributed in $[0,1]^{n-1}$ where $F_1, F_2, \ldots, F_{n-1}$ satisfy conditions (15)–(16) and take

$$F_n(x_1, x_2, \ldots, x_n) = \mathbb{P}(X_n \le x_n \mid X^{n-1} = x^{n-1}) \tag{19}$$

As above $x_n \mapsto F_n(x_1, x_2, \ldots, x_n)$ is nondecreasing differentiable and for any $u_n \in [0,1]$,

$$\mathbb{P}(F_n(X_1, X_2, \ldots, X_n) \le u_n \mid X^{n-1} = x^{n-1}) = u_n \tag{20}$$

which show that $F_n(X_1, X_2, \ldots, X_n)$ is uniformly distributed in $[0,1]$ independently of $F^{n-1}(X^{n-1})$, hence the resulting n-dimensional transformation $F(X)$ is uniformly distributed in $[0,1]^n$.

In particular, the diagonal transformation $\Phi(X^G) = (\Phi(X_1^G), \Phi(X_2^G), \ldots, \Phi(X_n^G))$ is uniformly distributed in $[0,1]^n$, where Φ denotes the c.d.f. of the standard Gaussian. Thus $F^{-1}(\Phi(X^G))$ is identically distributed as X where the diffeomorphism $F^{-1}\Phi$ has the prescribed form. ∎

Lemma 2. *Let X be a random vector with density and F be any diffeomorphism $\mathbb{R}^n \to \mathbb{R}^n$. Then*

$$h(F(X)) = h(X) + \mathbb{E} \log J_F(X) \tag{21}$$

where $J_F(x)$ is the Jacobian of the transformation $y = F(x)$.

Proof: Well known and easily checked by making the change of variable. ∎

Remark 1. Let F be such that $U = F(X)$ is uniformly distributed in $[0,1]^n$ as in the proof of Lemma 1. Then $h(U) = 0$ so that

$$h(X) = -\mathbb{E} \log J_F(X) \tag{22}$$

(an interesting formula, which is non trivial for dependent vector components).

IV. From Almost Gaussian to Gaussian

Hereafter we assume that (14) holds with conditions (15)–(16).

Proposition 1. *The component-wise correlation coefficients defined as*

$$\rho(X_i^G, X_i) = \frac{\mathbb{E}\big(X_i^G \cdot X_i\big)}{P} \in [0,1] \tag{23}$$

are such that

$$-\sum_{i=1}^{n} \log \rho(X_i^G, X_i) \le h(X^G) - h(X). \tag{24}$$

Thus if X is close to X^G in differential entropy (or KL distance), this inequality forces all correlation coefficients to be close to 1. The condition $\rho(X_i^G, X_i) = 1$ means that X_i^G and X_i are linearly dependent. In other words, the components of an almost Gaussian vector are *almost linearly dependent* on the Gaussian components.

Proof: One has

$$\begin{aligned} h(X^G) - h(X) &\overset{(a)}{=} h(X^G) - h(F(X^G)) && (25)\\ &\overset{(b)}{=} -\mathbb{E} \log J_F(X^G) && (26)\\ &\overset{(c)}{=} -\sum_{i=1}^{n} \mathbb{E} \log \frac{\partial F_i}{\partial y_i}(X^G) && (27)\\ &\overset{(d)}{\ge} -\sum_{i=1}^{n} \log \mathbb{E} \frac{\partial F_i}{\partial y_i}(X^G) && (28)\\ &\overset{(e)}{=} -\sum_{i=1}^{n} \log \frac{\mathbb{E}\big\{X_i^G \cdot F_i(X^G)\big\}}{P} && (29) \end{aligned}$$

where

(a) follows from Lemma 1;
(b) follows from Lemma 2;
(c) follows from the fact that by Lemma 1, the Jacobian matrix of the transformation $x = F(y)$ is triangular with nonnegative diagonal elements;
(d) is Jensen's inequality;
(e) is Stein's Lemma for the normal $X^G \sim \mathcal{N}(0, P\mathbf{I}_n)$. ∎

Lemma 3 (MSE lemma)**.** *Under the above assumptions, we have*

$$\frac{\mathbb{E}\big\{|X - X^G|^2\big\}}{P} \le h(X^G) - h(X). \tag{30}$$

Proof: Expanding $\mathbb{E}\big\{|X - X^G|^2\big\} = \mathbb{E}\{|X|^2\} + \mathbb{E}\{|X^G|^2\} - 2\,\mathbb{E}\big\{X^G \cdot X\big\}$ we get

$$\begin{aligned} \mathbb{E}\big\{|X - X^G|^2\big\} &\le nP + nP - 2\,\mathbb{E}\big\{X^G \cdot X\big\} && (31)\\ &= 2P \cdot \sum_{i=1}^{n} \Big(1 - \frac{\mathbb{E}\big\{X_i^G \cdot X_i\big\}}{P}\Big) && (32)\\ &\le -2P \sum_{i=1}^{n} \log \frac{\mathbb{E}\big\{X_i^G \cdot X_i\big\}}{P} && (33) \end{aligned}$$

The conclusion follows from Proposition 1. ∎

Remark 2. This result appears similar to the one of Calmon et al. [9] mentioned above, yet in the opposite direction: if X is

close to Gaussian in differential entropy, then the mean-squared error $\mathbb{E}\{|X - X^G|^2\}$ must be small, where $X = F(X^G)$ so that F is "almost linear" (close to the identity). This is looking much like a mean-squared error behavior of an "almost linear" estimator although it is not clear to us what would be the estimation problem at stake here.

V. A Converse Result under Addition of Another Random Quantity

Let V be an arbitrary (not necessarily Gaussian) random vector, independent of (X, X^G) and consider

$$\begin{aligned} Y &= X + V \\ \tilde{Y} &= X^G + V \end{aligned} \tag{34}$$

Since of course $|Y - \tilde{Y}|^2 = |X - X^G|^2$ it follows from the MSE Lemma 3 that the MSE $\mathbb{E}\{|Y - \tilde{Y}|^2\}$ will be small if X is close to Gaussian in differential entropy. In this section, we establish a partial converse.

Although the entropy difference $h(X^G) - h(X)$ is always non negative, the addition of V could make the difference $h(X^G + V) - h(X + V) = h(\tilde{Y}) - h(Y)$ negative, as was observed e.g., in [12], [13]. The following proposition shows how negative it can be. The negative part turns out to be small if the MSE is small:

Proposition 2. *Suppose V has zero mean and satisfies the constraint $|V|^2 \le nQ$. If $\mathbb{E}\{|Y - \tilde{Y}|^2\} \le n\epsilon(n)$ then*

$$h(Y) - h(\tilde{Y}) \le n\epsilon'(n)$$

where $\epsilon(n), \epsilon'(n) \to 0$ as $n \to +\infty$.

In other words, normalizing by dimension, the negative part of $\big(h(\tilde{Y}) - h(Y)\big)/n$ will be small if $\frac{1}{n}\,\mathbb{E}\{|Y - \tilde{Y}|^2\} = \frac{1}{n}\,\mathbb{E}\{|X - X^G|^2\}$ is small.

Proof. Let the densities be $Y \sim p$ and $\tilde{Y} \sim q$. Since $D_{\mathrm{KL}}(Y\|\tilde{Y}) \ge 0$, we have

$$h(Y) - h(\tilde{Y}) = \mathbb{E}\log\frac{q(\tilde{Y})}{p(Y)} \le \mathbb{E}\log\frac{q(\tilde{Y})}{q(Y)}. \tag{35}$$

The p.d.f. of $\tilde{Y} = X^G + V$ takes the form

$$q(\tilde{y}) = \mathbb{E}\{q(\tilde{y}|Z)\} = \frac{\mathbb{E}\exp\Big(-\dfrac{|\tilde{y} - V|^2}{2P}\Big)}{(2\pi)^{n/2}P^n}. \tag{36}$$

Since for any $y \in \mathbb{R}^n$,

$$|y - V|^2 \le |\tilde{y} - V|^2 + |y - \tilde{y}|^2 + 2|y - \tilde{y}| \cdot |\tilde{y} - V| \tag{37}$$

where $|\tilde{y} - V| \le |\tilde{y}| + \sqrt{nQ}$, we have

$$\log\frac{q(\tilde{y})}{q(y)} \le \frac{|y - \tilde{y}|^2}{2P} + \frac{|y - \tilde{y}|}{P} \cdot \big(|y| + \sqrt{nQ}\big). \tag{38}$$

Taking expectations, we have

$$h(Y) - h(\tilde{Y}) \le \frac{\mathbb{E}\{|Y - \tilde{Y}|^2\}}{2P} + \mathbb{E}\Big(\frac{|Y - \tilde{Y}|}{P} \cdot \big(|\tilde{Y}| + \sqrt{nQ}\big)\Big). \tag{39}$$

By Cauchy-Schwarz inequality, the second term is bounded by $\sqrt{\mathbb{E}\{|Y - \tilde{Y}|^2\}} \cdot \big(\sqrt{\mathbb{E}\{|\tilde{Y}|^2\}} + \sqrt{nN}\big)/P$ where $\sqrt{\mathbb{E}\{|\tilde{Y}|^2\}} \le \sqrt{nP} + \sqrt{nQ}$. Therefore, $\mathbb{E}\{|Y - \tilde{Y}|^2\} \le n\epsilon(n)$ implies $h(Y) - h(\tilde{Y}) \le \frac{1}{P}\Big(n\epsilon(n)/2 + (\sqrt{P} + 2\sqrt{Q})\sqrt{n^2\epsilon(n)}\Big) = n\epsilon'(n)$ where $\epsilon'(n) \to$ as $\epsilon(n) \to 0$. □

VI. Application to the Determination of the "Missing" Corner Point

As mentioned above, Polyanskiy and Wu [7] recently solved the missing corner point problem using optimal transport theory by showing Lipschtiz continuity of differential entropy with respect to the Wasserstein distance and Talagrand's transportation-information inequality. An independent solution using the I-MMSE approach was given by Bustin *et al.* [14], [15] for a restricted subset of inputs—and later more generally—by integration of the MMSE over a continuum of SNR values.

We provide yet another solution to the problem in continuation of our previous investigations [1], [12], [13] that relies only on basic properties of Shannon's information theory. Our proof is based on the MSE Lemma 3 and Proposition 2.

We follow the notations and definitions of [1], in particular:

Definition 1 (Asymptotic Almost Inequalities)**.** Let $\epsilon(n)$ denote *any* positive function of n which tends to 0^+ as $n \to +\infty$. Given real number sequences a_n, b_n, we write $a_n \lesssim b_n$ (a_n is *almost less* than b_n) if $a_n \le b_n + n\epsilon(n)$ or equivalently $b_n \ge a_n - n\epsilon(n)$. We also write $b_n \gtrsim a_n$ (b_n is *almost greater* than a_n).

Definition 2 (Almost Gaussianness)**.** X (with average power P) is *almost (white) Gaussian* (AG) if

$$h(X) \gtrsim h(X^G) = \frac{n}{2}\log(2\pi eP). \tag{40}$$

Definition 3 (Almost Losslessness)**.** Let Z and Z' be mutually independent (not necessarily Gaussian) vectors, independent of X. The addition of Z' in $X + Z + Z'$ is *almost lossless* (AL) with respect to X if their mutual information is almost nondecreasing:

$$I(X; X + Z + Z') \gtrsim I(X; X + Z). \tag{41}$$

We say that $X + Z + Z'$ is *almost lossless* compared to $X + Z$ with respect to X, or more briefly that $(X + Z) + Z'$ is AL (w.r.t. X).

Using the concavity of entropy power [16] we have shown the following [1, Cor. 2].

Proposition 3. *Let $a^2 \le 1$. If $aX_1 + X_2 + Z$ is almost lossless compared to $aX_1 + Z$, and if $X_2 + Z$ is white Gaussian, then $aX_1 + X_2 + Z$ is almost lossless compared to $X_1 + Z$ w.r.t. X_1.*

In this section we prove the following almost identical version of Proposition 3, which differs from it only by the addition of the word "almost":

Theorem 1. *Let $a^2 \le 1$. If $aX_1 + X_2 + Z$ is almost lossless compared to $aX_1 + Z$, and if $X_2 + Z$ is **almost** white Gaussian, then $aX_1 + X_2 + Z$ is almost lossless compared to $X_1 + Z$ w.r.t. X_1.*

As we have shown in [1, Prop. 6], this settles a long standing conjecture about the missing corner point for the Gaussian interference channel. We have also shown [1, Prop. 7] that it suffices to prove the following, which is essentially the problematic Appendix B in [2]:

Theorem 2. *If $X_2 + Z$ is AG, then $I(X_1; aX_1 + X_2 + Z) \lesssim I(X_1; aX_1 + X_2^G + Z)$, that is,*

$$h(aX_1 + X_2 + Z) \lesssim h(aX_1 + X_2^G + Z), \tag{42}$$

Proof: Let $X_2^G \sim \mathcal{N}(0, P_2\mathbf{I}_n)$. The condition that $X_2 + Z$ is AG is equivalent to $h(X_2^G + Z) - h(X_2 + Z) = n\epsilon(n)$ where $X_2^G + Z \sim \mathcal{N}(0, (P_2 + N)\mathbf{I}_n)$ and $X_2 + Z$ has a continuous density (for a proof see e.g., [11, Lemma 1]). Let F be as in Lemma 1 so that $X_2 + Z = F(X_2^G + Z)$. By the MSE Lemma 3,

$$\mathbb{E}\big\{|X_2 - X_2^G|^2\big\} = \mathbb{E}\big\{|(X_2 + Z) - (X_2^G + Z)|^2\big\} = n\epsilon(n) \tag{43}$$

Upon addition of aX_1 (which satisfies the constraint $|aX_1|^2 \leq na^2P_1$), it follows from Proposition 2 that

$$h(aX_1 + X_2 + Z) - h(aX_1 + X_2^G + Z) \leq n\epsilon'(n) \tag{44}$$

which ends the proof. ∎

ACKNOWLEDGMENTS

The authors would like to thank Chandra Nair, Shlomo Shamai, Flavio Calmon and Yihong Wu for their discussions. This work was partially supported by FAPESP.

REFERENCES

[1] O. Rioul and M. H. M. Costa, "Almost there: Corner points of Gaussian interference channels," in *Information Theory and Applications Workshop (ITA 2015)*, San Diego, Feb. 2–6 2015.

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

[3] I. Sason, "On achievable rate regions for the Gaussian interference channel," *IEEE Trans. Inf. Theory*, vol. 50, no. 6, pp. 1345–1356, June 2004.

[4] G. Kramer, "Review of rate regions for interference channels," in *Proc. IEEE Int. Zurich Seminar on Communications (IZS)*, Feb. 22–24, 2006, pp. 162–165.

[5] C. Villani, *Topics in optimal transportation*, ser. Graduate Studies in Mathematics. American Mathematical Society, 2003, vol. 58.

[6] M. Raginsky and I. Sason, *Concentration of Measure Inequalities in Information Theory, Communications, and Coding*, ser. Foundations and Trends in Communications and Information Theory. now Publishers, 2013, vol. 10, no. 1–2.

[7] Y. Polyanskiy and Y. Wu, "Wasserstein continuity of entropy and outer bounds for interference channels," 2015, draft at http://arxiv.org/abs/1504.04419.

[8] Y. Wu, "A simple proof of (a slightly improved) Gaussian HWI inequality," Dec. 2015, preprint (First version: Sept. 8, 2011). Private Communication.

[9] F. P. Calmon, Y. Polyanskiy, and Y. Wu, "Strong data processing inequalities in power constrained Gaussian channels," in *Proceedings of the 2015 IEEE International Symposium on Information Theory*, Hong Kong, June 2015.

[10] O. Rioul, "A simple proof of the entropy-power inequality via properties of mutual information," in *Proceedings of the IEEE International Symposium on Information Theory*, Nice, France, June 24-29th, 2007.

[11] ——, "Information theoretic proofs of entropy power inequalities," *IEEE Transactions on Information Theory*, vol. 57, no. 1, pp. 33–55, Jan. 2011.

[12] M. H. M. Costa and O. Rioul, "From almost Gaussian to Gaussian," in *AIP Proc. Int. Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering (MaxEnt)*, Amboise, France, Sept. 21–26, 2014.

[13] ——, "From almost Gaussian to Gaussian: Bounding differences of differential entropies," in *Information Theory and Applications Workshop (ITA 2015)*, San Diego, Feb. 2–6 2015.

[14] R. Bustin, H. V. Poor, and S. Shamai, "The effect of maximal rate codes on the interfering message rate," in *Proc. ISIT'14*, Honolulu, Hawaii, USA, July 2014, pp. 91–95, longer draft available at http://arxiv.org/abs/1404.6690.

[15] ——, "Optimal point-to-point codes in interference channels: An incremental approach," 2015, draft at http://arxiv.org/abs/1510.08213.

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