Wasserstein continuity of entropy and outer bounds for interference channels

Yury Polyanskiy and Yihong Wu*

January 31, 2016

Abstract

It is shown that under suitable regularity conditions, differential entropy is $O(\sqrt{n})$-Lipschitz as a function of probability distributions on $\mathbb{R}^n$ with respect to the quadratic Wasserstein distance. Under similar conditions, (discrete) Shannon entropy is shown to be $O(n)$-Lipschitz in distributions over the product space with respect to Ornstein's $\bar{d}$-distance (Wasserstein distance corresponding to the Hamming distance). These results together with Talagrand's and Marton's transportation-information inequalities allow one to replace the unknown multi-user interference with its i.i.d. approximations. As an application, a new outer bound for the two-user Gaussian interference channel is proved, which, in particular, settles the "missing corner point" problem of Costa (1985).

1 Introduction

Let X and $\tilde{X}$ be random vectors in $\mathbb{R}^n$. We ask the following question: If the distributions of X and $\tilde{X}$ are close in certain sense, can we guarantee that their differential entropies are close as well? For example, one can ask whether

$$D(P_X \| P_{\tilde{X}}) = o(n) \overset{?}{\Rightarrow} |h(X) - h(\tilde{X})| = o(n). \tag{1}$$

One motivation comes from multi-user information theory, where frequently one user causes interference to the other and in proving the converse one wants to replace the complicated non-i.i.d. interference by a simpler i.i.d. approximation. As a concrete example, we consider the so-called "missing corner point" problem in the capacity region of the two-user Gaussian interference channels (GIC) [Cos85a]. Perhaps due to the explosion in the number of interfering radio devices, this problem has attracted renewed attention recently [Cos11, BPS14, CR15, RC15]. For further information on capacity region of GIC and especially the problem of corner points, we refer to a comprehensive account just published by Igal Sason [Sas15].

Mathematically, the key question for settling "missing corner point" is the following: Given independent n-dimensional random vectors X_1, X_2, G_2, Z with the latter two being Gaussian, is it true that

$$D(P_{X_2+Z} \| P_{G_2+Z}) = o(n) \overset{?}{\Rightarrow} |h(X_1 + X_2 + Z) - h(X_1 + G_2 + Z)| = o(n). \tag{2}$$

*Y.P. is with the Department of EECS, MIT, Cambridge, MA, email: yp@mit.edu. Y.W. is with the Department of ECE and the Coordinated Science Lab, University of Illinois at Urbana-Champaign, Urbana, IL, email: yihongwu@illinois.edu.

To illustrate the nature of the problem, we first note that the answer to (1) is in fact negative as the counterexample of $X \sim \mathcal{N}(0, 2I_n)$ and $\tilde{X} \sim \frac{1}{2}\mathcal{N}(0, I_n) + \frac{1}{2}\mathcal{N}(0, 2I_n)$ demonstrates, in which case the divergence is $D(P_X \| P_{\tilde{X}}) \leq \log 2$ but the differential entropies differ by $\Theta(n)$. Therefore even for very smooth densities the difference in entropies is not controlled by the divergence. The situation for discrete alphabets is very similar, in the sense that the gap of Shannon entropies cannot be bounded by divergence in general (with essentially the same counterexample as that in the continuous case: X and $\tilde{X}$ being uniform on one and two Hamming spheres respectively).

The rationale of the above discussion is two-fold: a) Certain regularity conditions of the distributions must be imposed; b) Distances other than KL divergence might be more suited for bounding the entropy difference. Correspondingly, the main contribution of this paper is the following: Under suitable regularity conditions, the difference in entropy (in both continuous and discrete cases) can in fact be bounded by the *Wasserstein distance*, a notion originating from optimal transportation theory which turns out to be the main tool of this paper.

We start with the definition of the Wasserstein distance on the Euclidean space. Given probability measures P, Q on $\mathbb{R}^n$, define their p-Wasserstein distance ($p \geq 1$) as

$$W_p(P, Q) \triangleq \inf(\mathbb{E}[\|X - Y\|^p])^{1/p}, \tag{3}$$

where $\|\cdot\|$ denotes the Euclidean distance and the infimum is taken over all couplings of P and Q, i.e., joint distributions P_{XY} whose marginals satisfy $P_X = P$ and $P_Y = Q$. The following dual representation of the W_1 distance is useful:

$$W_1(P, Q) = \sup_{\mathrm{Lip}(f) \leq 1} \int f dP - \int f dQ. \tag{4}$$

Similar to (1), it is easy to see that in order to control $|h(X) - h(\tilde{X})|$ by means of $W_2(P_X, P_{\tilde{X}})$, one necessarily needs to assume some regularity properties of P_X and $P_{\tilde{X}}$; otherwise, choosing one to be a fine quantization of the other creates infinite gap between differential entropies, while keeping the W_2 distance arbitrarily small. Our main result in Section 2 shows that under moment constraints and certain conditions on the densities (which are in particular satisfied by convolutions with Gaussians), various information measures such as differential entropy and mutual information on $\mathbb{R}^n$ are in fact $\sqrt{n}$-Lipschitz continuous with respect to the W_2-distance. These results have natural counterparts in the discrete case where the Euclidean distance is replaced by Hamming distance (Section 4).

Furthermore, *transportation-information inequalities*, such as those due to Marton [Mar86] and Talagrand [Tal96], allow us to bound the Wasserstein distance by the KL divergence (see, e.g., [RS13] for a review). For example, Talagrand's inequality states that if $Q = \mathcal{N}(0, \Sigma)$, then

$$W_2^2(P, Q) \leq \frac{2\sigma_{\max}(\Sigma)}{\log e} D(P\|Q)\,, \tag{5}$$

where $\sigma_{\max}(\Sigma)$ denotes the maximal singular value of Σ. Invoking (5) in conjunction with the Wasserstein continuity of the differential entropy, we establish (2) and prove a new outer bound for the capacity region of the two-user GIC, finally settling the missing corner point in [Cos85a]. See Section 3 for details.

One interesting by-product is an estimate that goes in the reverse direction of (5). Namely, under regularity conditions on P and Q we have[1]

$$D(P\|Q) \lesssim \sqrt{\int_{\mathbb{R}^n} \|x\|^2 (dP + dQ)} \cdot W_2(P, Q) \tag{6}$$

[1] For positive a, b, denote $a \lesssim b$ if a/b is at most some universal constant.

See Proposition 1 and Corollary 4 in the next section. We want to emphasize that there are a number of estimates of the form $D(P_{X+Z}\|P_{\tilde X+Z})\lesssim W_2^2(P_X,P_{\tilde X})$ where $\tilde X, X$ are independent of a standard Gaussian vector Z, cf. [Vil03, Chapter 9, Remark 9.4]. The key difference of these estimates from (6) is that the W_2 distance is measured *after* convolving with P_Z.

Notations Throughout this paper log is with respect to an arbitrary base, which also specifies the units of differential entropy $h(\cdot)$, Shannon entropy $H(\cdot)$, mutual information $I(\cdot;\cdot)$ and divergence $D(\cdot\|\cdot)$. The natural logarithm is denoted by ln. The norm of $x\in\mathbb{R}^n$ is denoted by $\|x\|\triangleq(\sum_{j=1}^n x_j^2)^{1/2}$. For random variables X and Y, let $X\perp\!\!\!\perp Y$ denote their independence.

2 Wasserstein-continuity of information quantities

We say that a probability density function p on $\mathbb{R}^n$ is (c_1,c_2)-regular if $c_1>0, c_2\ge 0$ and

$$\|\nabla\log p(x)\|\le c_1\|x\|+c_2,\qquad \forall x\in\mathbb{R}^n\,.$$

Notice that in particular, regular density is never zero and furthermore

$$|\log p(x)-\log p(0)|\le\frac{c_1}{2}\|x\|^2+c_2\|x\|$$

Therefore, if X has a regular density and finite second moment then

$$|h(X)|\le|\log P_X(0)|+c_2\mathbb{E}[\|X\|]+\frac{c_1}{2}\mathbb{E}[\|X\|^2]<\infty\,.$$

Proposition 1. *Let U and V be random vectors with finite second moments. If V has a (c_1,c_2)-regular density p_V, then there exists a coupling P_{UV}, such that*

$$\mathbb{E}\left[\left|\log\frac{p_V(V)}{p_V(U)}\right|\right]\le\Delta\,,\tag{7}$$

where

$$\Delta=\left(\frac{c_1}{2}\sqrt{\mathbb{E}[\|U\|^2]}+\frac{c_1}{2}\sqrt{\mathbb{E}[\|V\|^2]}+c_2\right)W_2(P_U,P_V)\,.$$

Consequently,

$$h(U)-h(V)\le\Delta.\tag{8}$$

If both U and V are (c_1,c_2)-regular, then

$$|h(U)-h(V)|\le\Delta,\tag{9}$$

$$D(P_U\|P_V)+D(P_V\|P_U)\le 2\Delta.\tag{10}$$

Proof. First notice:

$$|\log p_V(v)-\log p_V(u)|=\left|\int_0^1 dt\,\langle\nabla\log p_V(tv+(1-t)u),u-v\rangle\right|\tag{11}$$

$$\le\int_0^1 dt(c_2+c_1t\|v\|+c_1(1-t)\|u\|)\|u-v\|\tag{12}$$

$$=(c_2+c_1\|v\|/2+c_1\|u\|/2)\|u-v\|,\tag{13}$$

where (12) follows from Cauchy-Schwartz inequality and the (c_1, c_2)-regularity of p_V. Taking expectation of (13) with respect to (u, v) distributed according to the optimal W_2-coupling of P_U and P_V and then applying Cauchy-Schwartz and triangle inequality for L_2-norm, we obtain (7).

To show (8) notice that by finiteness of second moment $h(U) < \infty$. If $h(U) = -\infty$ then there is nothing to prove. So assume otherwise, then in identity

$$h(U) - h(V) + D(P_U \| P_V) = \mathbb{E}\left[\log \frac{p_V(V)}{p_V(U)}\right] \tag{14}$$

all terms are finite and hence (8) follows. Clearly, (8) implies (9) (when applied with U and V interchanged).

Finally, for (10) just add the identity (14) to itself with U and V interchanged to obtain

$$D(P_U\|P_V) + D(P_V\|P_U) = \mathbb{E}\left[\log \frac{p_V(V)}{p_V(U)}\right] + \mathbb{E}\left[\log \frac{p_U(U)}{p_U(V)}\right]$$

and estimate both terms via (7). □

The key question now is what densities are regular. It turns out that convolution with sufficiently smooth density, such as Gaussians, produces a regular density.

Proposition 2. *Let $V = B + Z$ where $B \perp\!\!\!\perp Z \sim \mathcal{N}(0, \sigma^2 I_n)$ and $\mathbb{E}[\|B\|] < \infty$. Then the density of V is (c_1, c_2)-regular with $c_1 = \frac{3\log e}{\sigma^2}$ and $c_2 = \frac{4\log e}{\sigma^2}\mathbb{E}[\|B\|]$.*

Proof. First notice that whenever density p_Z of Z is differentiable and non-vanishing, we have:

$$\nabla \log p_V(v) = \frac{\mathbb{E}[\nabla p_Z(v - B)]}{p_V(v)} = \mathbb{E}[\nabla \log p_Z(v - B)|V = v]\,, \tag{15}$$

where $p_V(v) = \mathbb{E}[p_Z(v - B)]$ is the density of V. For $Z \sim \mathcal{N}(0, \sigma^2 I_n)$, we have

$$\nabla \log p_Z(v - B) = \frac{\log e}{\sigma^2}(B - v).$$

So the proof is completed by showing

$$\mathbb{E}[\|B - v\| \,|\, V = v] \le 3\|v\| + 4\mathbb{E}[\|B\|]\,. \tag{16}$$

For this, we mirror the proof in [WV12, Lemma 4]. Indeed, we have

$$\begin{aligned}\mathbb{E}[\|B - v\||V = v] &= \mathbb{E}\left[\|B - v\|\frac{p_Z(B - v)}{p_V(v)}\right] &(17)\\ &\le 2\mathbb{E}[\|B - v\|1\{a(B, v) \le 2\}] + \mathbb{E}[\|B - v\|a(B, v)1\{a(B, v) > 2\}]\,, &(18)\end{aligned}$$

where we denoted

$$a(B, v) \triangleq \frac{p_Z(B - v)}{p_V(v)}\,.$$

Next, notice that

$$\{a(B, v) > 2\} = \{\|B - v\|^2 \le -2\sigma^2 \ln((2\pi\sigma^2)^{n/2} 2p_V(v))\}\,.$$

Thus since $\mathbb{E}[p_Z(B - v)] = p_V(v)$ we have an upper bound for the second term in (18) as follows

$$\mathbb{E}[\|B - v\|a(B, v)1\{a(B, v) > 2\}] \le \sqrt{2}\sigma\sqrt{\ln^+ \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}} 2p_V(v)}}\,, \tag{19}$$

where $\ln^+ x \triangleq \max\{0, \ln x\}$. From Markov inequality we have $\mathbb{P}[\|B\| \le 2\mathbb{E}[\|B\|]] \ge 1/2$ and therefore

$$p_V(v) \ge \frac{1}{2(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{(\|v\|+2\mathbb{E}[\|B\|])^2}{2\sigma^2}} .$$

Using this estimate in (19) we get

$$\mathbb{E}[\|B - v\| a(B, v) 1\{a(B, v) > 2\}] \le \|v\| + 2\mathbb{E}[\|B\|] . \tag{20}$$

Upper-bounding the first term in (18) by $2\mathbb{E}[\|B\|] + 2\|v\|$ we finish the proof of (16). □

Another useful criterion for regularity is the following:

Proposition 3. *If W has (c_1, c_2)-regular density and $B \perp\!\!\!\perp W$ satisfies*

$$\|B\| \le \sqrt{nP} \qquad a.s. \tag{21}$$

then $V = B + W$ has $(c_1, c_2 + c_1\sqrt{nP})$-regular density.

Proof. Apply (15) and the estimate:

$$\mathbb{E}[\|\nabla \log p_W(v - B)\| \,|\, V = v] \le c_1(\|v\| + \sqrt{nP}) + c_2.$$ □

As a consequence of regularity, we show that when smoothed by Gaussian noise, mutual information, differential entropy and divergence are Lipschitz with respect to the W_2-distance under average power constraints:

Corollary 4. *Assume that $X, \tilde{X} \perp\!\!\!\perp Z$, with $\mathbb{E}[\|X\|^2], \mathbb{E}[\|\tilde{X}\|^2] \le nP$ and $Z \sim \mathcal{N}(0, \sigma^2 I_n)$. Then*

$$|I(X; X + Z) - I(\tilde{X}; \tilde{X} + Z)| = |h(X + Z) - h(\tilde{X} + Z)| \le \Delta, \tag{22}$$

$$D(P_{\tilde{X}+Z} \| P_{X+Z}) + D(P_{X+Z} \| P_{\tilde{X}+Z}) \le 2\Delta, \tag{23}$$

where $\Delta = \frac{\log e}{\sigma^2}(3\sqrt{n(\sigma^2 + P)} + 4\sqrt{nP}) W_2(P_{X+Z}, P_{\tilde{X}+Z})$.

Proof. Since $\mathbb{E}[\|X\|] \le \sqrt{nP}$, by Proposition 2, the densities of $X + Z$ and $\tilde{X} + Z$ are both $(\frac{3\log e}{\sigma^2}, \frac{4\sqrt{nP}\log e}{\sigma^2})$-regular. The desired statement then follows from applying (9)-(10) to $V = X + Z$ and $U = \tilde{X} + Z$. □

Remark 1. The Lipschitz constant $\sqrt{n}$ is order-optimal as the example of Gaussian X and $\tilde{X}$ with different variances (one of them could be zero) demonstrates. The linear dependence on W_2 is also optimal. To see this, consider $X \sim \mathcal{N}(0, 1)$ and $\tilde{X} \sim \mathcal{N}(0, 1 + t)$ in one dimension. Then $|h(X + Z) - h(\tilde{X} + Z)| = 1/2 \log(1 + t/2) = \Theta(t)$ and $W_2^2(X + Z, \tilde{X} + Z) = (\sqrt{2 + t} - \sqrt{2})^2 = \Theta(t^2)$, as $t \to 0$.

In fact, to get the best constants for applications to interference channels it is best to forgo the notion of regular density and deal directly with (15). Indeed, when the inputs has bounded norms, the next result gives a sharpened version of what can be obtained by combining Proposition 1 with 2.

Proposition 5. *Let B satisfying (21) and $G \sim \mathcal{N}(0, \sigma_G^2 I_n)$ be independent. Let $V = B + G$. Then for any U,*

$$h(U) - h(V) \le \frac{\log e}{2\sigma_G^2} \left(\mathbb{E}[\|U\|^2] - \mathbb{E}[\|V\|^2] + 2\sqrt{nP} W_1(P_U, P_V) \right) . \tag{24}$$

Proof. Plugging Gaussian density $p_G(z) = \frac{1}{\sqrt{2\pi}\sigma_G} e^{-z^2/(2\sigma_G^2)}$ into (15) we get

$$\nabla \log p_V(v) = \frac{\log e}{\sigma_G^2} (\hat{B}(v) - v)\,, \tag{25}$$

where $\hat{B}(v) \triangleq \mathbb{E}[B|V=v] = \frac{\mathbb{E}[Bp_G(v-B)]}{\mathbb{E}[p_G(v-B)]}$ satisfies

$$\|\hat{B}(v)\| \le \sqrt{nP},$$

since $\|B\| \le \sqrt{nP}$ almost surely. Next we use

$$\begin{aligned} \log \frac{p_V(v)}{p_V(u)} &= \int_0^1 dt\, \langle \nabla \log p_V(tv + (1-t)u), v-u \rangle & (26)\\ &= \frac{\log e}{\sigma_G^2} \int_0^1 dt \langle \hat{B}(tv+(1-t)u), v-u\rangle - \frac{\log e}{2\sigma_G^2}(\|v\|^2 - \|u\|^2) & (27)\\ &\le \frac{\sqrt{nP}\log e}{\sigma_G^2} \|v-u\| - \frac{\log e}{2\sigma_G^2}(\|v\|^2 - \|u\|^2)\,. & (28) \end{aligned}$$

Taking expectation of the last equation under the W_1-optimal coupling and in view of (14), we obtain (24). □

To get slightly better constants in one-sided version of (22) we apply Proposition 5:

Corollary 6. *Let A, B, G, Z be independent, with $G \sim \mathcal{N}(0, \sigma_G^2 I_n)$, $Z \sim \mathcal{N}(0, \sigma_Z^2 I_n)$ and B satisfying (21). Then for every $c \in [0,1]$ we have:*

$$\begin{aligned} &h(B+A+Z) - h(B+G+Z) \\ &\le \frac{\log e}{2(\sigma_G^2 + \sigma_Z^2)} \left(\mathbb{E}[\|A\|^2] + 2\langle \mathbb{E}[A], \mathbb{E}[B]\rangle - \mathbb{E}[\|G\|^2]\right) + \frac{\sqrt{2nP(\sigma_G^2 + c^2\sigma_Z^2)\log e}}{\sigma_G^2 + \sigma_Z^2} \sqrt{D(P_{A+cZ} \| P_{G+cZ})} \end{aligned} \tag{29}$$

Proof. First, notice that by definition Wasserstein distance is non-increasing under convolutions, i.e., $W_2(P_1 * Q, P_2 * Q) \le W_2(P_1, P_2)$. Since $c \le 1$ and Gaussian distribution is stable, we have

$$W_2(P_{B+A+Z}, P_{B+G+Z}) \le W_2(P_{A+Z}, P_{G+Z}) \le W_2(P_{A+cZ}, P_{G+cZ}),$$

which, in turn, can be bounded via Talagrand's inequality (5) by

$$W_2(P_{A+cZ}, P_{G+cZ}) \le \sqrt{\frac{2(\sigma_G^2 + c^2\sigma_Z^2)}{\log e} D(P_{A+cZ}\|P_{G+cZ})}\,.$$

From here we apply Proposition 5 with G replaced by $G+Z$ (and σ_G^2 by $\sigma_Z^2 + \sigma_G^2$). □

3 Applications to Gaussian interference channels

3.1 New outer bound

Consider the two-user Gaussian interference channel (GIC):

$$\begin{aligned} Y_1 &= X_1 + bX_2 + Z_1 \\ Y_2 &= aX_1 + X_2 + Z_2\,, \end{aligned} \tag{30}$$

with $a, b \geq 0$, $Z_i \sim \mathcal{N}(0, I_n)$ and a power constraint on the n-letter codebooks: either

$$\|X_1\| \leq \sqrt{nP_1}, \quad \|X_2\| \leq \sqrt{nP_2} \quad \text{a.s.} \tag{31}$$

or

$$\mathbb{E}[\|X_1\|^2] \leq nP_1, \quad \mathbb{E}[\|X_2\|^2] \leq nP_2. \tag{32}$$

Denote by $\mathcal{R}(a, b)$ the capacity region of the GIC (30). As an application of the results developed in Section 2, we prove an outer bound for the capacity region.

Theorem 7. *Let $0 < a \leq 1$. Let $C_2 = \frac{1}{2}\log(1+P_2)$ and $\tilde{C}_2 = \frac{1}{2}\log(1+\frac{P_2}{1+a^2P_1})$. Assume the almost sure power constraint (31). Then for any $b \geq 0$ and $\tilde{C}_2 \leq R_2 \leq C_2$, any rate pair $(R_1, R_2) \in \mathcal{R}(a, b)$ satisfies*

$$R_1 \leq \frac{1}{2}\log\min\left\{A - \frac{1}{a^2} + 1, A\frac{(1+P_2)(1-(1-a^2)\exp(-2\delta)) - a^2}{P_2}\right\} \tag{33}$$

where

$$A = (P_1 + a^{-2}(1+P_2))\exp(-2R_2), \tag{34}$$

$$\delta = C_2 - R_2 + a\sqrt{\frac{2P_1(C_2 - R_2)\log e}{1+P_2}}. \tag{35}$$

Assume the average power constraint (32). Then (33) holds with δ replaced by

$$\delta' = C_2 - R_2 + \sqrt{\frac{2(C_2 - R_2)\log e}{1+P_2}}(3\sqrt{1+a^2P_1+P_2} + 4a\sqrt{P_1}). \tag{36}$$

Consequently, in both cases, $R_2 \geq C_2 - \epsilon$ implies that $R_1 \leq \frac{1}{2}\log(1+\frac{a^2P_1}{1+P_2}) - \epsilon'$ where $\epsilon' = O(\sqrt{\epsilon})$ as $\epsilon \to 0$.

Proof. Without loss of generality, assume that all random variables have zero mean. First of all, setting $b = 0$ (which is equivalent to granting the first user access to X_2) will not shrink the capacity region of the interference channel (30). Therefore to prove the desired outer bound it suffices to focus on the following Z-interference channel henceforth:

$$\begin{aligned} Y_1 &= X_1 + Z_1 \\ Y_2 &= aX_1 + X_2 + Z_2\,. \end{aligned} \tag{37}$$

Let (X_1, X_2) be n-dimensional random variables corresponding to the encoder output of the first and second user, which are uniformly distributed on the respective codebook. For $i = 1, 2$ define

$$R_i \triangleq \frac{1}{n}I(X_i; Y_i)\,.$$

By Fano's inequality there is no difference asymptotically between this definition of rate and the operational one. Define the entropy-power function of the X_1-codebook:

$$N_1(t) \triangleq \exp\left\{\frac{2}{n}h(X_1 + \sqrt{t}Z)\right\}, \qquad Z \sim \mathcal{N}(0, I_n)\,.$$

We know the following general properties of $N_1(t)$:

- N_1 is monotonically increasing.
- $N_1(0) = 0$ (since X_1 is uniform over the codebook).
- $N_1'(t) \geq 2\pi e$ (since $N_1(t+\delta) \geq N_1(t) + 2\pi e\delta$ by entropy power inequality).
- $N_1(t)$ is concave (Costa's entropy power inequality [Cos85a]).
- $N_1(t) \leq 2\pi e(P_1 + t)$ (Gaussian maximizes differential entropy).

We can then express R_1 in terms of the entropy power function as

$$R_1 = \frac{1}{2}\log\frac{N_1(1)}{2\pi e}\,. \tag{38}$$

It remains to upper bound $N_1(1)$. Note that

$$nR_2 = I(X_2;Y_2) = h(X_2 + aX_1 + Z) - h(aX_1 + Z) \leq \frac{n}{2}\log 2\pi e(1 + P_2 + a^2P_1) - h(aX_1 + Z)\,,$$

and therefore

$$N_1\left(\frac{1}{a^2}\right) \leq 2\pi e A\,, \tag{39}$$

where A is defined in (34). This in conjunction with the slope property $N_1'(t) \geq 2\pi e$ yields

$$N_1(1) \leq N_1\left(\frac{1}{a^2}\right) - 2\pi e(a^{-2} - 1) \leq 2\pi e(A - a^{-2} + 1)\,, \tag{40}$$

which, in view of (38), yields the first part of the bound (33).

To obtain the second bound, let $G_2 \sim \mathcal{N}(0, P_2 I_n)$. Using $\mathbb{E}[\|X_2\|^2] \leq nP_2$ and $X_1 \perp\!\!\!\perp X_2$, we obtain

$$\begin{aligned} nR_2 &= I(X_2;Y_2) \leq I(X_2;Y_2|X_1) = I(X_2;X_2+Z_2) \\ &= nC_2 - h(G_2 + Z_2) + h(X_2 + Z_2) \leq nC_2 - D(P_{X_2+Z_2}\|P_{G_2+Z_2}), \end{aligned}$$

that is,

$$D(P_{X_2+Z_2}\|P_{G_2+Z_2}) \leq h(G_2 + Z_2) - h(X_2 + Z_2) \leq n(C_2 - R_2). \tag{41}$$

Furthermore,

$$\begin{aligned} nR_2 = I(X_2;Y_2) &= h(aX_1 + X_2 + Z_2) - h(aX_1 + G_2 + Z_2) \qquad (42)\\ &\quad + h(aX_1 + G_2 + Z_2) - h(aX_1 + Z_2)\,. \qquad (43) \end{aligned}$$

Note that the second term (43) is precisely $\frac{n}{2}\log\frac{N_1(\frac{1}{a^2})}{N_1(\frac{1+P_2}{a^2})}$. The first term (42) can be bounded by applying Corollary 6 and (41) with $B = aX_1$, $A = X_2$, $G = G_2$ and $c = 1$:

$$h(aX_1 + X_2 + Z_2) - h(aX_1 + G_2 + Z_2) \leq n\sqrt{\frac{2a^2P_1(C_2 - R_2)\log e}{1 + P_2}}. \tag{44}$$

Combining (42) – (44) yields

$$N_1\left(\frac{1}{a^2}\right) \leq \frac{\exp(2\delta)}{1 + P_2}N_1\left(\frac{1 + P_2}{a^2}\right)\,. \tag{45}$$

where δ is defined in (35). From the concavity of $N_1(t)$ and (45)

$$N_1(1) \le \gamma N_1\left(\frac{1}{a^2}\right) - (\gamma-1)N_1\left(\frac{1+P_2}{a^2}\right) \tag{46}$$

$$\le N_1\left(\frac{1}{a^2}\right)\left(\gamma - (\gamma-1)\frac{1+P_2}{\exp(2\delta)}\right), \tag{47}$$

where $\gamma = 1 + \frac{1-a^2}{P_2} > 1$. In view of (38), upper bounding $N_1\left(1/a^2\right)$ in (47) via (39) we get after some simplifications the second part of (33).

The outer bound for average power constraint (32) follows analogously with (44) replaced by (48) below: By Proposition 2, the density of $aX_1 + G_2 + Z_2$ is $(\frac{3\log e}{1+P2}, \frac{4a\log e\sqrt{nP_1}}{1+P2})$-regular. Applying Proposition 1 to (44), we have $h(aX_1 + X_2 + Z_2) - h(aX_1 + G_2 + Z_2) \le \Delta$, where

$$\Delta = (3\sqrt{1+a^2P_1+P_2} + 4a\sqrt{P_1})\frac{\log e}{1+P_2}\sqrt{n}W_2(P_{aX_1+X_2+Z_2}, P_{aX_1+G_2+Z_2}).$$

Again using the fact that W_2 distance is non-decreasing under convolutions and invoking Talagrand's inequality, we have

$$W_2(P_{aX_1+X_2+Z_2}, P_{aX_1+G_2+Z_2}) \le W_2(P_{X_2+Z_2}, P_{G_2+Z_2}) \le \sqrt{\frac{2(1+P_2)}{\log e}D(P_{X_2+Z_2}\|P_{G_2+Z_2})},$$

which yields

$$h(aX_1 + X_2 + Z_2) - h(aX_1 + G_2 + Z_2) \le n\sqrt{\frac{2(C_2 - R_2)\log e}{1+P_2}}(3\sqrt{1+a^2P_1+P_2} + 4a\sqrt{P_1}). \tag{48}$$

This yields the outer bound with δ' defined in (36).

Finally, in both cases, when $R_2 \to C_2$, we have $\delta \to 0$ and $A \to \frac{1}{a^2} + \frac{P_2}{1+P_1}$ and hence from (33) $R_1 \le C_1'$. □

Remark 2. The first part of the bound (33) coincides with Sato's outer bound [Sat78] and [Kra04, Theorem 2] by Kramer, which [Kra04, Theorem 2] was obtained by reducing the Z-interference channel to the degraded broadcast channel; the second part of (33) is new, which settles the missing corner point of the capacity region (see Section 3.2 for discussions). Note that our estimates on $N_1(1)$ in the proof of Theorem 7 are tight in the sense that there exists a concave function $N_1(t)$ satisfying the listed general properties, estimates (45) and (39) as well as attaining the minimum of (40) and (47) at $N_1(1)$. Hence, tightening the bound via this method would require inferring more information about $N_1(t)$.

Remark 3. The outer bound (33) relies on Costa's EPI. To establish the second statement about corner point, it is sufficient to invoke the concavity of $\gamma \mapsto I(X_2; \sqrt{\gamma}X_2 + Z_2)$ [GSSV05, Corollary 1], which is strictly weaker than Costa's EPI.

The outer bound (33) is evaluated on Fig. 1 for the case of $b = 0$ (Z-interference), where we also plot (just for reference) the simple Han-Kobayashi inner bound for the Z-GIC (37) attained by choosing $X_1 = U + V$ with $U \perp\!\!\!\perp V$ jointly Gaussian. This achieves rates:

$$\begin{cases} R_1 = \frac{1}{2}\log(1+P_1-s) + \frac{1}{2}\log\left(1+\frac{a^2s}{1+a^2(P_1-s)+P_2}\right) \\ R_2 = \frac{1}{2}\log\left(1+\frac{P_2}{1+a^2(P_1-s)}\right) \end{cases} \quad 0 \le s \le P_1. \tag{49}$$

For more sophisticated Han-Kobayashi bounds see [Sas04, Cos11].

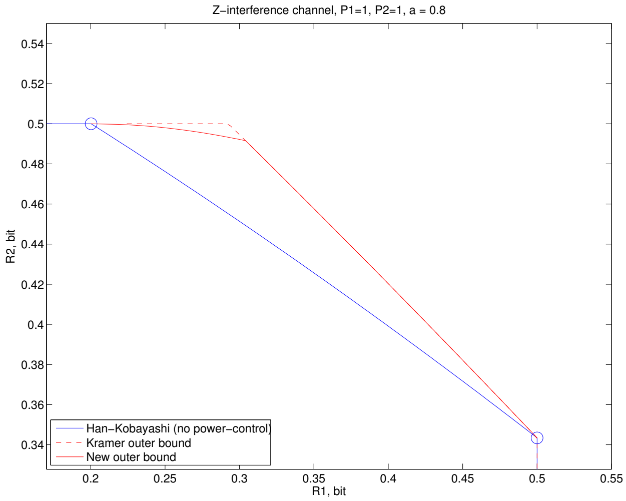

Figure 1: Illustration of the "missing corner point": The bound in Theorem 7 establishes the location of the upper corner point, as conjectured by Costa [Cos85b]. The bottom corner point has been established by Sato [Sat78].

3.2 Corner points of the capacity region

The two corner points of the capacity region are defined as follows:

$$C_1'(a,b) \triangleq \max\{R_1 : (R_1, C_2) \in \mathcal{R}(a,b)\}\,, \tag{50}$$

$$C_2'(a,b) \triangleq \max\{R_2 : (C_1, R_2) \in \mathcal{R}(a,b)\}\,, \tag{51}$$

where $C_i = \frac{1}{2}\log(1+P_i)$. As a corollary, Theorem 7 completes the picture of the corner points for the capacity region of GIC for all values of $a, b \in \mathbb{R}_+$ under the average power constraint (32). We note that the new result here is the proof of $C_1'(a,b) = \frac{1}{2}\log\left(1 + \frac{a^2 P_1}{1+P_2}\right)$ for $0 < a \le 1$ and $b \ge 0$. The interpretation is that if one user desires to achieve its own interference-free capacity, then the other user must guarantee that its message is decodable at both receivers. The achievability of this corner point was previously known, while the converse was previously considered by Costa [Cos85b] but with a flawed proof, as pointed out in [Sas04]. The high-level difference between our proof and that of [Cos85b] is the replacement of Pinsker's inequality by Talagrand's and the use of a coupling argument.[2]

Below we present a brief account of the corner points in various cases; for an extensive discussion see [Sas15]. We start with a few simple observations about the capacity region $\mathcal{R}(a,b)$:

- Any rate pair satisfying the following belongs to $\mathcal{R}(a,b)$:

$$\begin{aligned} R_1 &\le \frac{1}{2}\log(1 + P_1\min(1,a^2)) \\ R_2 &\le \frac{1}{2}\log(1 + P_2\min(1,b^2)) \\ R_1 + R_2 &\le \frac{1}{2}\log(1 + \min(P_1 + b^2 P_2, P_2 + a^2 P_1)), \end{aligned} \tag{52}$$

[2] After circulating our initial draft, we were informed that authors of [BPS14] posted an updated manuscript [BPS15a] that also proves Costa's conjecture. Their method is based on the analysis of the minimum mean-square error (MMSE) properties of good channel codes, but we were not able to verify all the details. A further update is in [BPS15b].

which corresponds to the intersection of two Gaussian multiple-access (MAC) capacity regions, namely, $(X_1, X_2) \to Y_1$ and $(X_1, X_2) \to Y_2$. These rate pairs correspond to the case when each receiver decodes both messages.

- For $a > 1$ and $b > 1$ (strong interference) the capacity region is known to coincide with (52) [Car75, Sat81]. So, without loss of generality we assume $a \le 1$ henceforth.

- Replacing either a or b with zero can only enlarge the region (genie argument).

- If $b \ge 1$ then for any $(R_1, R_2) \in \mathcal{R}(a, b)$ we have [Sat81]

$$R_1 + R_2 \le \frac{1}{2} \log\left(1 + b^2 P_2 + P_1\right) . \tag{53}$$

  This follows from the observation that in this case $I(X_1, X_2; Y_1) = H(X_1, X_2) - o(n)$, since conditioned on X_1, Y_2 is a noisier observation of X_2 than Y_1.

For the top corner, we have the following:

$$C_1'(a,b) = \begin{cases} \frac{1}{2}\log\left(1 + \frac{a^2 P_1}{1+P_2}\right), & 0 < a \le 1, b \ge 0 \\ C_1, & a = 0, b = 0 \text{ or } b \ge \sqrt{1+P_1} \\ \frac{1}{2}\log\left(1 + \frac{P_1 + (b^2-1)P_2}{1+P_2}\right), & a = 0, 1 < b < \sqrt{1+P_1} \\ \frac{1}{2}\log\left(1 + \frac{P_1}{1+b^2 P_2}\right), & a = 0, 0 < b \le 1. \end{cases} \tag{54}$$

Note that for any $b \ge 0$, $a \mapsto C_1'(a, b)$ is discontinuous as $a \downarrow 0$. To verify (54) we consider each case separately:

1. For $a > 0$ the converse bound follows from Theorem 7. For achievability, we consider two cases. When $b \le 1$, we have $\frac{a^2 P_1}{1+P_2} \le \frac{P_1}{1+b^2 P_2}$ and therefore treating interference X_2 as noise at the first receiver and using a Gaussian MAC-code for $(X_1, X_2) \to Y_2$ works. For $b > 1$, the achievability follows from the MAC inner bound (52). Note that since $\frac{1}{2}\log\left(1 + P_1 + b^2 P_2\right) \ge \frac{1}{2}\log\left(1 + P_2 + a^2 P_1\right)$, a Gaussian MAC-code that works for $(X_1, X_2) \to Y_2$ will also work for $(X_1, X_2) \to Y_1$. Alternatively, the achievability also follows from Han-Kobayashi inner bound (see, e.g., [EGK11, Theorem 6.4] with $(U_1, U_2) = (X_1, X_2)$ for $b \ge 1$ and $(U_1, U_2) = (X_1, 0)$ for $b \le 1$).

2. For $a = 0$ and $b \ge \sqrt{1 + P_1}$ the converse is obvious, while for achievability we have that $\frac{b^2 P_2}{1+P_1} \le P_2$ and therefore X_2 is decodable at Y_1.

3. For $a = 0$ and $1 < b < \sqrt{1 + P_1}$ the converse is (53) and the achievability is just the MAC code $(X_1, X_2) \to Y_1$ with rate $R_2 = C_2$.

4. For $a = 0$ and $0 < b \le 1$ the result follows from the treatment of $C_2'(a, b)$ below by interchanging $a \leftrightarrow b$ and $P_1 \leftrightarrow P_2$.

The bottom corner point is given by the following:

$$C_2'(a,b) = \begin{cases} \frac{1}{2}\log\left(1 + \frac{P_2}{1+a^2 P_1}\right), & 0 \le a \le 1, b = 0 \text{ or } b \ge \sqrt{\frac{1+P_1}{1+a^2 P_1}} \\ \frac{1}{2}\log\left(1 + \frac{b^2 P_2}{1+P_1}\right), & 0 \le a \le 1, 1 < b < \sqrt{\frac{1+P_1}{1+a^2 P_1}} \\ \frac{1}{2}\log\left(1 + \frac{b^2 P_2}{1+P_1}\right), & 0 \le a \le 1, 0 < b \le 1 \end{cases} \tag{55}$$

which is discontinuous as $b \downarrow 0$ for any fixed $a \in [0, 1]$. We treat each case separately:

1. The case of $C_2'(a,0)$ is due to Sato [Sat78] (see also [Kra04, Theorem 2]). The converse part also follows from Theorem 7 (for $a = 0$ there is nothing to prove). For the achievability, we notice that under $b \geq \sqrt{\frac{1+P_1}{1+a^2P_1}}$ we have $\frac{b^2P_2}{1+P_1} > \frac{P_2}{1+a^2P_1}$ and thus X_2 at rate $C_2'(a,0)$ can be decoded and canceled from Y_1 by simply treating X_1 as Gaussian noise (as usual, we assume Gaussian random codebooks). Thus the problem reduces to that of $b = 0$. For $b = 0$, the Gaussian random coding achieves the claimed result if the second receiver treats X_1 as Gaussian noise.

2. The converse follows from (53) and for the achievability we use the Gaussian MAC-code $(X_1, X_2) \to Y_1$ and treat X_1 as Gaussian interference at Y_2.

3. If $b \in (0,1]$, we apply results on $C_1'(a,b)$ in (54) by interchanging $a \leftrightarrow b$ and $P_1 \leftrightarrow P_2$.

4 Discrete version

4.1 Bounding entropy and information via Ornstein's distance

Fix a finite alphabet $\mathcal{X}$ and an integer n. On the product space $\mathcal{X}^n$ we define the Hamming distance

$$d_H(x,y) = \sum_{j=1}^{n} \mathbf{1}_{\{x_j \neq y_j\}}\,,$$

and consider the corresponding Wasserstein distance W_1. In fact, $\frac{1}{n}W_1(P,Q)$ is known as Ornstein's $\bar{d}$-distance [GNS75, Mar86], namely,

$$\bar{d}(P,Q) = \frac{1}{n}\inf \mathbb{E}[d_H(X,Y)], \tag{56}$$

where the infimum is taken over all couplings P_{XY} of P and Q. For $n = 1$, this coincides with the total variation, which is also expressible as $d_{\mathrm{TV}}(P,Q) = \frac{1}{2}\int |dP - dQ|$ for P, Q on $\mathcal{X}$.

For a pair of distributions P, Q on $\mathcal{X}^n$ we may ask the following questions:

1. Does $D(P\|Q)$ control the entropy difference $H(P) - H(Q)$?

2. Does $\bar{d}(P,Q)$ control the entropy difference $H(P) - H(Q)$?

Recall that in the Euclidean space the answer to both questions was negative unless the distributions satisfy certain regularity conditions. For discrete alphabets the answer to the first question is still negative in general (see Section 1 for a counterexample); nevertheless, the answer to the second one turns out to be positive:

Proposition 8. *Let P and Q be distributions on $\mathcal{X}^n$ and let*

$$F_{\mathcal{X}}(x) \triangleq x\log(|\mathcal{X}|-1) + x\log\frac{1}{x} + (1-x)\log\frac{1}{1-x}\,, \quad 0 \leq x \leq 1.$$

Then

$$|H(P) - H(Q)| \leq nF_{\mathcal{X}}(\bar{d}(P,Q))\,. \tag{57}$$

Proof. In fact, the statement holds for any translation-invariant distance $d(\cdot,\cdot)$ on $\mathcal{X}$ extended additively to $\mathcal{X}^n$, i.e., $d(x,x') = \sum_{i=1}^n d(x_i,x'_i)$ for any $x,x' \in \mathcal{X}^n$. Indeed, define

$$f_n(s) \triangleq \max_{P_X} \left\{ \frac{1}{n} H(X) : \mathbb{E}[d(X,x_0)] \le ns \right\},$$

where $x_0 \in \mathcal{X}^n$ is an arbitrary fixed string. It is easy to see that $s \mapsto f_n(s)$ is concave since $P \mapsto H(P)$ is. Furthermore, writing $X = (X_1,\ldots,X_n)$ and applying chain-rule for entropy we get

$$f_n(s) = f_1(s).$$

Thus, letting X,Y be distributed according to the $\bar{d}$-optimal coupling of P and Q, we get

$$\begin{aligned} H(X) - H(Y) &\le H(X,Y) - H(Y) = H(X|Y) && (58)\\ &\le n\mathbb{E}\left[f_n(\mathbb{E}[d(X,Y)|Y])\right] && (59)\\ &\le nf_n(\bar{d}(P,Q)), && (60) \end{aligned}$$

where (59) is by definition of $f_n(\cdot)$ and (60) is by Jensen's inequality. Finally, for the Hamming distance we have $f_1(s) = F_{\mathcal{X}}(s)$ by Fano's inequality. □

Notice that the right-hand side of (57) behaves like $n\bar{d}\log\frac{1}{\bar{d}}$ when $\bar{d}(P,Q)$ is small. This superlinear dependence is in fact sharp.[3] Nevertheless, if certain regularity of distributions is assumed, the estimate (57) can be improved to be linear in $\bar{d}(P,Q)$. The next result is the analog of Proposition 1 in the discrete space. We formulate it in a form convenient for applications in multi-user information theory.

Proposition 9. *Let $P_{Y|X,A}$ be a two-input blocklength-n memoryless channel, namely*

$$P_{Y|X,A}(y|x,a) = \prod_{j=1}^n W(y_j|x_j,a_j),$$

where $W(\cdot|\cdot)$ is a stochastic matrix and $y \in \mathcal{Y}^n, x \in \mathcal{X}^n, a \in \mathcal{A}^n$. Let $X, A, \tilde{A}$ be independent n-dimensional discrete random vectors. Let Y and $\tilde{Y}$ be the outputs generated by (X,A) and $(X,\tilde{A})$, respectively. Then

$$\begin{aligned} |H(Y) - H(\tilde{Y})| &\le cn\bar{d}(P_Y,P_{\tilde{Y}}) && (61)\\ D(P_Y\|P_{\tilde{Y}}) + D(P_{\tilde{Y}}\|P_Y) &\le 2cn\bar{d}(P_Y,P_{\tilde{Y}}) && (62)\\ |I(X;Y) - I(X;\tilde{Y})| &\le 2cn\mathbb{E}[\bar{d}(P_{Y|X},P_{\tilde{Y}|X})] && (63) \end{aligned}$$

where

$$\begin{aligned} c &\triangleq \max_{x,a,y,y'} \log\frac{W(y|x,a)}{W(y'|x,a)}, && (64)\\ \mathbb{E}[\bar{d}(P_{Y|X},P_{\tilde{Y}|X})] &\triangleq \sum_{x\in\mathcal{X}^n} P_X(x)\bar{d}(P_{Y|X=x},P_{\tilde{Y}|X=x}). && (65) \end{aligned}$$

[3] To see this, consider $Q = \text{Bern}(p)^{\otimes n}$ and choose P to be the output distribution of the optimal lossy compressor for Q at average distortion δn. By definition, $\bar{d}(P,Q) \le \delta$. On the other hand, $H(P) = n(h(p) - h(\delta) + o(1))$ as $n\to\infty$ and hence $|H(P) - H(Q)| = n(h(\delta) + o(1))$, which asymptotically meets the upper bound (57) with equality.

Proof. Given any stochastic matrix G, define $L(G) \triangleq \max_{u,u',v} \log \frac{G(u|v)}{G(u'|v)}$. Recall the following fact from [PV14, Eqn. (58)] about mixtures of product distributions: Let U and V be n-dimensional discrete random vector connected by a product channel, that is, $P_{V|U} = \prod_{i=1}^n P_{V_i|U_i}$. Then the mapping $v \mapsto \log P_V(v)$ is L-Lipschitz with respect to the Hamming distance, where $L = \max_{j\in[n]} L(P_{V_i|U_i})$. Consider another pair $(\tilde{U}, \tilde{V})$ connected by the same channel, i.e., $P_{\tilde{V}|\tilde{U}} = P_{V|U}$. Then Lipschitz continuity implies that $\mathbb{E}\left[\left|\log \frac{P_V(V)}{P_V(\tilde{V})}\right|\right] \le L\mathbb{E}[d_H(V,\tilde{V})]$ for any coupling $P_{V\tilde{V}}$. Optimizing over the coupling and in view of (56), we obtain

$$\mathbb{E}\left[\left|\log \frac{P_V(V)}{P_V(\tilde{V})}\right|\right] \le Ln\bar{d}(P_V, P_{\tilde{V}})\,.$$

Repeating the proof of (8)–(10), we have

$$|H(V) - H(\tilde{V})| \le Ln\bar{d}(P_V, P_{\tilde{V}}) \tag{66}$$

$$D(P_V\|P_{\tilde{V}}) + D(P_{\tilde{V}}\|P_V) \le 2cn\bar{d}(P_V, P_{\tilde{V}}) \tag{67}$$

Applying (66) and (67) to Y and (X, A) gives (61) and (62) with $L = c$ defined in (64).

To bound the mutual information, we first notice

$$|I(X;Y) - I(X;\tilde{Y})| \le |H(Y) - H(\tilde{Y})| + |H(Y|X) - H(\tilde{Y}|X)|\,.$$

Applying (66) conditioned on $X = x$ we get

$$|H(Y|X = x) - H(\tilde{Y}|X = x)| \le c_x n\bar{d}(P_{Y|X=x}, P_{\tilde{Y}|X=x})\,,$$

where $c_x = \max_{j\in[n]} \max_{y,y',a} \log \frac{W(y|x_j,a)}{W(y'|x_j,a)}$. Note that $c_x \le c$ for any x, averaging over P_X gives

$$|H(Y|X) - H(\tilde{Y}|X)| \le cn\mathbb{E}[\bar{d}(P_{Y|X}, P_{\tilde{Y}|X})]\,. \tag{68}$$

From the convexity of $(P, Q) \mapsto \bar{d}(P, Q)$, which holds for any Wasserstein distance, we have $\bar{d}(P_Y, P_{\tilde{Y}}) \le \mathbb{E}[\bar{d}(P_{Y|X}, P_{\tilde{Y}|X})]$ and so the left-hand side of (68) also bounds $|H(Y) - H(\tilde{Y})|$ from above. □

4.2 Marton's transportation inequality

In this section we discuss how previous bounds (Proposition 8 and 9) in terms of the $\bar{d}$-distance can be converted to bounds in terms of KL divergence. This is possible when Q is a product distribution, thanks to Marton's transportation inequality [Mar86, Lemma 1]. We formulate this together with a few other properties of the $\bar{d}$-distance in the following lemma proved in Appendix A.

Lemma 10.

1. *(Marton's transportation inequality [Mar86]): For any pair of distributions P and $Q = \prod_{i=1}^n Q_i$ on $\mathcal{X}^n$,*

$$\bar{d}\,(P, Q) \le \sqrt{\frac{D(P\|Q)}{2n\log e}}. \tag{69}$$

2. *(Tensorization)* $\bar{d}(\prod_{i=1}^n P_i, \prod_{i=1}^n Q_i) \le \frac{1}{n}\sum_{i=1}^n d_{\mathrm{TV}}(P_i, Q_i)$.

3. *(Contraction) For P_{XY} and Q_{XY} such that $P_{Y|X} = Q_{Y|X} = \prod_{i=1}^n P_{Y_i|X_i}$,*
$$\bar{d}(P_Y, Q_Y) \le \max_{i\in[n]} \eta_{\mathrm{TV}}(P_{Y_i|X_i})\bar{d}(P_X, Q_X). \tag{70}$$
where $\eta_{\mathrm{TV}}(W)$ is Dobrushin's contraction coefficient of a Markov kernel W defined as $\eta_{\mathrm{TV}}(W) = \sup_{x,x'} d_{\mathrm{TV}}(W(\cdot|x), W(\cdot|x'))$.

If we assume that $D(P\|Q) = \epsilon n$ for some small ϵ, then combining (57) and (69) gives
$$|H(P) - H(Q)| \le nF_{\mathcal{X}}\left(\sqrt{\frac{D(P\|Q)}{2n\log e}}\right),$$
where the right-hand side behaves as $n\sqrt{\epsilon}\log\frac{1}{\epsilon}$ when $\epsilon \to 0$. This estimate has a one-sided improvement (here again Q must be a product distribution):
$$H(P) - H(Q) \le \sqrt{\frac{2nD(P\|Q)}{\log e}}\log|\mathcal{X}| \tag{71}$$
(see [CS07] for $n = 1$ and [WV10, Appendix H] for the general case).

Switching to the setting in Proposition 9, let us consider the case where $\tilde{A}$ has i.i.d. components, i.e., $P_{\tilde{A}} = P_0^{\otimes n}$. Define
$$\eta_{\mathrm{TV}} \triangleq \max_{x,a,a'} d_{\mathrm{TV}}(W(\cdot|x,a), W(\cdot|x,a')), \tag{72}$$
which is the maximal Dobrushin contraction coefficients among all channels $W(\cdot|\cdot,x)$ indexed by $x \in \mathcal{X}$. Then
$$\bar{d}(P_Y, P_{\tilde{Y}}) \le \mathbb{E}[\bar{d}(P_{Y|X}, P_{\tilde{Y}|X})] \le \eta_{\mathrm{TV}}\bar{d}(P_A, P_{\tilde{A}}) \le \eta_{\mathrm{TV}}\sqrt{\frac{D(P_A\|P_{\tilde{A}})}{2n\log e}}, \tag{73}$$
where the left inequality is by convexity of the $\bar{d}$-distance as a Wasserstein distance, the middle inequality is by Lemma 10, and the right inequality is via (69). An alternative to the estimate (73) is the following:
$$\bar{d}(P_Y, P_{\tilde{Y}}) \le \mathbb{E}[\bar{d}(P_{Y|X}, P_{\tilde{Y}|X})] \tag{74}$$
$$\le \mathbb{E}\left[\sqrt{\frac{1}{2n\log e}D(P_{Y|X}\|P_{\tilde{Y}|X})}\right] \tag{75}$$
$$\le \sqrt{\frac{1}{2n\log e}D(P_{Y|X}\|P_{\tilde{Y}|X}|P_X)} \tag{76}$$
$$\le \sqrt{\frac{1}{2n\log e}\eta_{\mathrm{KL}}D(P_A\|P_{\tilde{A}})} \tag{77}$$
where (75) is by (69) since $P_{\tilde{Y}|X=x}$ is a product distribution as $\tilde{A}$ has a product distribution, (76) is by Jensen's inequality, and (77) is by the tensorization property of the strong data-processing constant for divergence [AG76]:
$$\eta_{\mathrm{KL}} \triangleq \max_{x,Q_0} \frac{D(\sum_a Q_0(a)W(\cdot|x,a)\|\sum_a P_0(a)W(\cdot|x,a))}{D(P_0\|Q_0)}.$$
To conclude, in the regime of $D(P_A\|P_{\tilde{A}}) \le \epsilon n$ for some small ϵ our main Proposition 9 yields
$$|H(Y) - H(\tilde{Y})| \lesssim n\sqrt{\epsilon} \tag{78}$$
matching the behavior of (71). However, the estimate (78) is stronger, because a) it is two-sided and b) $P_{\tilde{Y}}$ can be a mixture of product distributions (since X in Proposition 9 may be arbitrary).

4.3 Application: corner points for discrete interference channels

In order to apply Proposition 9 to determine corner points of capacity regions of discrete memoryless interference channels (DMIC) we will need an auxiliary tensorization result. This result appears to be a rather standard exercise for degraded channels and so we defer the proof to Appendix B.

Proposition 11. *Given channels $P_{A|X}$ and $P_{B|A}$ on finite alphabets, define*

$$F_c(t) \triangleq \max\{H(X|A,U)\colon H(X|B,U) \le t, U \to X \to A \to B\}\,. \tag{79}$$

Then the following hold:

1. *(Property of F_c) The function $F_c : \mathbb{R}_+ \to \mathbb{R}_+$ is concave, non-decreasing and $F_c(t) \le t$. Furthermore, $F_c(t) < t$ for all $t > 0$, provided that $P_{B|A}$ and $P_{A|X}$ satisfy*

$$P_{B|A=a} \not\perp P_{B|A=a'}, \quad \forall a \neq a' \tag{80}$$

and

$$P_{A|X=x} \neq P_{A|X=x'}, \quad \forall x \neq x', \tag{81}$$

respectively.

2. *(Tensorization) For any blocklength-n Markov chain $X^n \to A^n \to B^n$, where $P_{A^n|X^n} = P_{A|X}^{\otimes n}$ and $P_{B^n|A^n} = P_{B|A}^{\otimes n}$ are n-letter memoryless channels, we have*

$$H(X^n|A^n) \le nF_c\left(\frac{1}{n}H(X^n|B^n)\right)\,. \tag{82}$$

Remark 4. Neither of the sufficient condition (80) and (81) for strict inequality is superfluous, as can be seen from the example $B = A$ and $A \perp\!\!\!\perp X$, respectively; in both cases $F_c(t) = t$.

The important consequence of Proposition 11 is the following implication:[4]

Corollary 12. *Let $X^n \to A^n \to B^n$, where the memoryless channels $P_{A|X}$ and $P_{B|A}$ of blocklength n satisfy the conditions (80) and (81). Then there exists a continuous function $g : \mathbb{R}_+ \to \mathbb{R}_+$ satisfying $g(0) = 0$, such that for all n*

$$I(X^n; A^n) \le I(X^n; B^n) + \epsilon n \quad \Longrightarrow \quad H(X^n) \le I(X^n; B^n) + g(\epsilon)n\,, \tag{83}$$

Proof. By Proposition 11, we have $F_c(t) < t$ for all $t > 0$. This together with the concavity of F_c implies that $t \mapsto t - F_c(t)$ is convex, strictly increasing and strictly positive on $(0, \infty)$. Define g as the inverse of $t \mapsto t - F_c(t)$, which is increasing and concave and satisfies $g(0) = 0$. Since $I(X^n; A^n) \le I(X^n; B^n) + \epsilon n$, the tensorization result (79) yields

$$H(X^n|B^n) \le H(X^n|A^n) + \epsilon n \le nF_c\left(\frac{1}{n}H(X^n|B^n)\right) + \epsilon n,$$

i.e., $t \le F_c(t) + \epsilon$, where $t \triangleq \frac{1}{n}H(X^n|B^n)$. Then $t \le g(\epsilon)$ by definition, completing the proof. □

[4] This is the analog of the following property of Gaussian channels, exploited in Theorem 7 in the form of Costa's EPI: For i.i.d. Gaussian Z and $t_1 < t_2 < t_3$ we have

$$I(X; X + t_2 Z) = I(X; X + t_3 Z) + o(n) \implies I(X; X + t_1 Z) = I(X; X + t_3 Z) + o(n)\,.$$

This also follows from the concavity of $\gamma \mapsto I(X; \sqrt{\gamma}X + Z)$.

We are now ready to state a non-trivial example of corner points for the capacity region of DMIC. The proof strategy mirrors that of Theorem 7, with Corollary 6 and Costa's EPI replaced by Proposition 9 and Corollary 12, respectively.

Theorem 13. *Consider the two-user DMIC:*

$$Y_1 = X_1\,, \tag{84}$$
$$Y_2 = X_2 + X_1 + Z_2 \mod 3\,, \tag{85}$$

where $X_1 \in \{0,1,2\}^n$, $X_2 \in \{0,1\}^n, Z_2 \in \{0,1,2\}^n$ are independent and $Z_2 \sim P_2^{\otimes n}$ is i.i.d. for some non-uniform P_2 containing no zeros. The maximal rate achievable by user 2 is

$$C_2 = \max_{\mathrm{supp}(Q)\subset\{0,1\}} H(Q * P_2) - H(P_2). \tag{86}$$

At this rate the maximal rate of user 1 is

$$C_1' = \log 3 - \max_{\mathrm{supp}(Q)\subset\{0,1\}} H(Q * P_2). \tag{87}$$

Remark 5. As an example, consider $P_2 = \left[1-\delta, \frac{\delta}{2}, \frac{\delta}{2}\right]$ where $\delta \neq 0, 1, \frac{1}{3}$. Then the maximum in (86) is achieved by $Q = [\frac{1}{2}, \frac{1}{2}]$. Therefore $C_2 = H(P_3) - H(P_2)$ and $C_1' = \log 3 - H(P_3)$, where $P_3 = \left[\frac{2-\delta}{4}, \frac{2-\delta}{4}, \frac{\delta}{2}\right]$. Note that in the case of $\delta = \frac{1}{3}$, where Theorem 13 is not applicable, we simply have $C_2 = 0$ and $C_1' = \log 2$ since $X_2 \perp\!\!\!\perp Y_2$. Therefore the corner point is discontinuous in δ.

Remark 6. Theorem 13 continues to hold even if cost constraints are imposed. Indeed, if $X_2 \in \{0,1,2\}^n$ is required to satisfy

$$\sum_{i=1}^n \mathsf{b}(X_{2,i}) \leq nB$$

for some cost function $\mathsf{b} : \{0,1,2\} \to \mathbb{R}$. Then the maximum in (86) and (87) is taken over all Q such that $\mathbb{E}_Q[\mathsf{b}(U)] \leq B$. Note that taking $B = \infty$ is equivalent to dropping the constraint $X_2 \in \{0,1\}^n$ in (86). In this case, $C_1' = 0$ which can be shown by a simpler argument not involving Proposition 9.

Proof. We start with the converse. Given a sequence of codes with vanishing probability of error and rate pairs (R_1, R_2), where $R_2 = C_2 - \epsilon$, we show that $R_1 \leq C_1' - \epsilon'$, where $\epsilon' \to 0$ as $\epsilon \to 0$. Let Q_2 be the maximizer of (86), i.e., the capacity-achieving distribution of the channel $X_2 \mapsto X_2 + Z_2$. Let $\tilde{X}_2 \in \{0,1\}^n$ be distributed according to Q_2^n. Then $\tilde{X}_2 + Z_2 \sim P_3^{\otimes n}$, where $P_3 = Q_2 * P_2$. By Fano's inequality,

$$\begin{aligned} n(C_2 - \epsilon + o(1)) = n(R_2 + o(1)) &= I(X_2; Y_2) \\ &\leq I(X_2; Y_2|X_1) = I(X_2; X_2 + Z_2) \qquad (88) \\ &= nC_2 - D(P_{X_2+Z_2} \| P_{\tilde{X}_2+Z_2}), \qquad (89) \end{aligned}$$

that is,

$$D(P_{X_2+Z_2} \| P_{\tilde{X}_2+Z_2}) \leq n\epsilon + o(n).$$

Since $P_{\tilde{X}_2+Z_2} = P_3^{\otimes n}$ is a product distribution, Marton's inequality (69) yields

$$\bar{d}(P_{X_1+X_2+Z_2}, P_{X_1+\tilde{X}_2+Z_2}) \leq \bar{d}(P_{X_2+Z_2}, P_{\tilde{X}_2+Z_2}) \leq \sqrt{\frac{\epsilon}{2\log e}} + o(1).$$

Applying (63) in Proposition 9 and in view of the translation invariance of the $\bar{d}$-distance, we obtain

$$\begin{aligned}|I(X_1;Y_2)-I(X_1;X_1+\tilde{X}_2+Z_2)| &= |I(X_1;X_1+X_2+Z_2)-I(X_1;X_1+\tilde{X}_2+Z_2)| \\ &\le 2cn\mathbb{E}[\bar{d}(P_{X_1+X_2+Z_2|X_1},P_{X_1+\tilde{X}_2+Z_2|X_1})] \\ &= 2cn\bar{d}(P_{X_2+Z_2},P_{\tilde{X}_2+Z_2}) \\ &\le (\alpha\sqrt{\epsilon}+o(1))n,\end{aligned}$$

where $c=\max_{z,z'\in\{0,1,2\}}\log\frac{P_2(z)}{P_2(z')}$ and $\alpha=\frac{2c}{\sqrt{2\log e}}$ are finite since P_2 contains no zeros by assumption. On the other hand,

$$I(X_1;X_1+Z_2)=I(X_1;Y_2|X_2)=I(X_1;Y_2)+I(X_1;X_2|Y_2)=I(X_1;Y_2)+o(n),$$

where $I(X_1;X_2|Y_2)\le H(X_2|Y_2)=o(n)$ by Fano's inequality. Combining the last two displays, we have

$$I(X_1;X_1+\tilde{X}_2+Z_2)\le I(X_1;X_1+Z_2)+(\alpha\sqrt{\epsilon}+o(1))n.$$

Next we apply Corollary 12, with $X=X_1\to A=X_1+Z_2\to B=A+\tilde{X}_2$. To verify the conditions, note that the channel $P_{A|X}$ is memoryless and additive with non-uniform noise distribution P_2, which satisfies the condition (81). Similar, the channel $P_{B|A}$ is memoryless and additive with noise distribution Q_2 , which is the maximizer of (86). Since P_2 is not uniform, Q_2 is not a point mass. Therefore $P_{B|A}$ satisfies (80). Then Corollary 12 yields

$$nR_1=H(X_1)\le I(X_1;X_1+\tilde{X}_2+Z_2)+g(\alpha\sqrt{\epsilon})n\le nC_1'+o(n),$$

where the last inequality follows from the fact that $\max_{X_1}I(X_1;X_1+\tilde{X}_2+Z_2)=nC_1'$ attained by X_1 uniform on $\{0,1,2\}^n$.

Finally, note that the rate pair (C_1',C_2) is achievable by a random MAC-code for $(X_1,X_2)\to Y_2$, with X_1 uniform on $\{0,1,2\}^n$ and $X_2\sim Q_2^{\otimes n}$. □

Acknowledgment

Explaining that the "missing corner point" requires proving of (2), as well as the majority of our knowledge on interference channels were provided by Prof. Chandra Nair. We acknowledge his scholarship and patience deeply.

The research of Y.P. has been supported in part by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-09-39370 and by the NSF CAREER award under grant agreement CCF-12-53205. The research of Y.W. has been supported in part by NSF grants IIS-14-47879, CCF-14-23088 and CCF-15-27105. This work would not be possible without the generous support of the Simons Institute for the Theory of Computing and California SB-420.

A Proof of Lemma 10

Proof. To prove the tensorization inequality, let $(X,Y)=(X_i,Y_i)_{i=1}^n$ be independent and individually distributed as the optimal coupling of (P_i,Q_i). Then $\mathbb{E}[d_H(X,Y)]=\sum_{i=1}^n\mathbb{P}\left[X_i\neq Y_i\right]=\sum_{i=1}^n d_{\mathrm{TV}}(P_i,Q_i)$.

To show (70), let $\pi_{X,Y,\tilde{X},\tilde{Y}}$ be an arbitrary coupling of P_{XY} and Q_{XY} so that $(X, \tilde{X})$ is distributed according to the optimal coupling of $\bar{d}(P_X, Q_X)$, that is, $\mathbb{E}_\pi[d_H(X, \tilde{X})] = n\bar{d}(P_X, Q_X)$. By the first inequality we just proved, for any $x, x' \in \mathcal{X}^n$,

$$\bar{d}(P_{Y|X=x}, P_{Y|X=\tilde{x}}) \le \frac{1}{n}\sum_{i=1}^{n} d_{\mathrm{TV}}(P_{Y_i|X_i=x_i}, P_{Y_i|X_i=\tilde{x}_i}) \le \frac{1}{n}\sum_{i=1}^{n} \eta_{\mathrm{TV}}(P_{Y_i|X_i})\mathbf{1}_{\{x_i \neq \tilde{x}_i\}} \le \frac{\eta d_H(x, \tilde{x})}{n}.$$

where $\eta = \max_{i\in[n]} \eta_{\mathrm{TV}}(P_{Y_i|X_i})$ and the middle inequality follows from Dobrushin's contraction coefficient. Applying Dobrushin's contractoin [Dob70] (see [PW16, Proposition 18], with $\rho = \frac{1}{n}d_H$ and $r = \eta\rho$), there exists a coupling $\pi'_{X,Y,\tilde{X},\tilde{Y}}$ of P_{XY} and Q_{XY}, so that $\pi'_{X\tilde{X}} = \pi_{X\tilde{X}}$ and $\mathbb{E}_{\pi'}[d_H(Y, \tilde{Y})] \le \eta\mathbb{E}_\pi[d_H(X, \tilde{X})] = n\eta\bar{d}(P_X, Q_X)$, concluding the proof. □

B Proof of Proposition 11

Proof. Basic properties of F_c follow from standard arguments. To show the strict inequality $F_c(t) < t$ under the conditions (80) and (81), we first notice that F_c is simply the concave envelope of the set of achievable pairs $(H(X|A), H(X|B))$ obtained by iterating over all P_X. By Caratheodory's theorem, it is sufficient to consider a ternary-valued U in the optimization defining $F_c(t)$. Then the set of achievable pairs $(H(X|A,U), H(X|B,U))$ is convex and compact (as the continuous image of the compact set of distributions $P_{U,X}$). Consequently, to have $F_c(t) = t$ there must exist a distribution $P_{U,X}$, such that

$$H(X|A,U) = H(X|B,U) = t\,. \tag{90}$$

We next show that under the extra conditions on $P_{B|A}$ and $P_{A|X}$ we must have $t = 0$. Indeed, (80) guarantees the channel $P_{B|A}$ satisfies the strong data processing inequality (see, e.g., [CK11, Exercise 15.12 (b)] and [PW16, Section 1.2] for a survey) that there exists $\eta < 1$ such that

$$I(X;B|U) \le \eta I(X;A|U). \tag{91}$$

From (90) and (91) we infer that $I(X;A|U) = 0$, or equivalently

$$D(P_{A|X} \| P_{A|U} | P_{U,X}) = 0\,.$$

On the other hand, the condition (81) ensures that then we must have $H(X|U) = 0$. Clearly, this implies $t = 0$ in (90).

To show the single-letterization statement (82), we only consider the case of $n = 2$ since the generalization is straightforward by induction. Let $X^2 \to A^2 \to B^2$ be a Markov chain with blocklength-2 memoryless channel in between. We have

$$\begin{aligned} H(X^2|B^2) &= H(X_1|B^2) + H(X_2|B^2, X_1) && (92)\\ &= H(X_1|B^2) + H(X_2|B_2, X_1) && (93)\\ &\ge H(X_1|B_1, A_2) + H(X_2|B_2, X_1) && (94) \end{aligned}$$

where (93) is because $B_2 \to X_2 \to X_1 \to B_1$ and hence $I(X_2; B_1|X_2B_2) = 0$, and (94) is because

$B_1 \to X_1 \to A_2 \to B_2$. Next consider the chain

$$\begin{aligned} H(X|A^2) &= H(X_1|A^2) + H(X_2|A^2, X_1) && (95)\\ &= H(X_1|A^2) + H(X_2|A_2, X_1) && (96)\\ &\le F_c(H(X_1|B_1, A_2)) + F_c(H(X_2|B_2, X_1)) && (97)\\ &\le 2F_c\left(\frac{1}{2}H(X_1|B_1, A_2) + \frac{1}{2}H(X_2|B_2, X_1)\right) && (98)\\ &\le 2F_c\left(\frac{1}{2}H(X^2|B^2)\right) && (99) \end{aligned}$$

where (96) is by $A_2 \to X_2 \to X_1 \to A_1$ and hence $I(X_2; A_1|X_1, A_2) = 0$, (97) is by the definition of F_c and since we have both $A_2 \to X_1 \to A_1 \to B_1$ and $X_1 \to X_2 \to A_2 \to B_2$, (98) is by the concavity of F_c, and finally (99) is by the monotonicity of F_c and (94). □

References

[AG76] R. Ahlswede and P. Gács. Spreading of sets in product spaces and hypercontraction of the Markov operator. *Ann. Probab.*, pages 925–939, 1976.

[BPS14] Ronit Bustin, H Vincent Poor, and Shlomo Shamai. The effect of maximal rate codes on the interfering message rate. In *Proc. 2014 IEEE Int. Symp. Inf. Theory (ISIT)*, pages 91–95, Honolulu, HI, USA, July 2014.

[BPS15a] Ronit Bustin, H Vincent Poor, and Shlomo Shamai. The effect of maximal rate codes on the interfering message rate. *arXiv preprint arXiv:1404.6690v4*, Apr 2015.

[BPS15b] Ronit Bustin, H Vincent Poor, and Shlomo Shamai. Optimal point-to-point codes in interference channels: An incremental I-MMSE approach. *arXiv preprint arXiv:1510.08213*, Oct 2015.

[Car75] A. Carleial. A case where interference does not reduce capacity (corresp.). *IEEE Trans. Inf. Theory*, 21(5):569–570, Sep 1975.

[CK11] Imre Csiszar and János Körner. *Information theory: coding theorems for discrete memoryless systems.* Cambridge University Press, 2nd edition, 2011.

[Cos85a] Max H.M. Costa. A new entropy power inequality. *IEEE Trans. Inf. Theory*, 31(6):751–760, 1985.

[Cos85b] Max H.M. Costa. On the Gaussian interference channel. *IEEE Trans. Inf. Theory*, 31(5):607–615, 1985.

[Cos11] Max H. M. Costa. Noisebergs in Z-Gaussian interference channels. In *Proc. Information Theory and Applications Workshop (ITA)*, San Diego, CA, February 2011.

[CR15] Max H.M. Costa and Olivier Rioul. From almost Gaussian to Gaussian: Bounding differences of differential entropies. In *Proc. Information Theory and Applications Workshop (ITA)*, San Diego, CA, February 2015.

[CS07] C. Chang and A. Sahai. Universal quadratic lower bounds on source coding error exponents. In *41st Annual Conference on Information Sciences and Systems*, pages 714–719, 2007.

[Dob70] R. L. Dobrushin. Definition of random variables by conditional distributions. *Theor. Probability Appl.*, 15(3):469–497, 1970.

[EGK11] Abbas El Gamal and Young-Han Kim. *Network information theory.* Cambridge University Press, 2011.

[GNS75] Robert M. Gray, David L. Neuhoff, and Paul C. Shields. A generalization of Ornstein's $\bar{d}$ distance with applications to information theory. *The Annals of Probability*, pages 315–328, 1975.

[GSSV05] D. Guo, S. Shamai (Shitz), and S. Verdú. Mutual Information and Minimum Mean-Square Error in Gaussian Channels. *IEEE Trans. Inf. Theory*, 51(4):1261 – 1283, Apr. 2005.

[Kra04] Gerhard Kramer. Outer bounds on the capacity of Gaussian interference channels. *IEEE Trans. Inf. Theory*, 50(3):581–586, 2004.

[Mar86] Katalin Marton. A simple proof of the blowing-up lemma (corresp.). *IEEE Trans. Inf. Theory*, 32(3):445–446, 1986.

[PV14] Y. Polyanskiy and S. Verdú. Empirical distribution of good channel codes with non-vanishing error probability. *IEEE Trans. Inf. Theory*, 60(1):5–21, January 2014.

[PW16] Yury Polyanskiy and Yihong Wu. Dissipation of information in channels with input constraints. *IEEE Trans. Inf. Theory*, 62(1):35–55, January 2016. also arXiv:1405.3629.

[RC15] Olivier Rioul and Max H.M. Costa. Almost there – corner points of Gaussian interference channels. In *Proc. Information Theory and Applications Workshop (ITA)*, San Diego, CA, February 2015.

[RS13] Maxim Raginsky and Igal Sason. Concentration of measure inequalities in information theory, communications, and coding. *Found. and Trends in Comm. and Inform Theory*, 10(1-2):1–247, 2013.

[Sas04] Igal Sason. On achievable rate regions for the Gaussian interference channel. *IEEE Trans. Inf. Theory*, 50(6):1345–1356, 2004.

[Sas15] Igal Sason. On the corner points of the capacity region of a two-user Gaussian interference channel. *IEEE Trans. Inf. Theory*, 61(7):3682–3697, July 2015.

[Sat78] Hiroshi Sato. On degraded Gaussian two-user channels (corresp.). *IEEE Trans. Inf. Theory*, 24(5):637–640, 1978.

[Sat81] Hiroshi Sato. The capacity of the Gaussian interference channel under strong interference (corresp.). *IEEE Trans. Inf. Theory*, 27(6):786–788, 1981.

[Tal96] M. Talagrand. Transportation cost for Gaussian and other product measures. *Geometric and Functional Analysis*, 6(3):587–600, 1996.

[Vil03] C. Villani. *Topics in optimal transportation.* American Mathematical Society, Providence, RI, 2003.

[WV10] Yihong Wu and Sergio Verdú. Rényi information dimension: Fundamental limits of almost lossless analog compression. *IEEE Trans. Inf. Theory*, 56(8):3721–3748, Aug. 2010.

[WV12] Yihong Wu and Sergio Verdú. Functional properties of MMSE and mutual information. *IEEE Trans. Inf. Theory*, 58(3):1289 – 1301, Mar. 2012.