International Zurich Seminar on Communications (IZS), February 26 – 28, 2014

On an Extremal Data Processing Inequality for long Markov Chains

Thomas A. Courtade
University of California, Berkeley
Department of Electrical Engineering and Computer Science
Email: courtade@eecs.berkeley.edu

Jiantao Jiao and Tsachy Weissman
Stanford University
Department of Electrical Engineering
Email: {jiantao, tsachy}@stanford.edu

***Abstract*—We pose the following extremal conjecture: Let X, Y be jointly Gaussian random variables with linear correlation ρ. For any random variables U, V for which U, X, Y, V form a Markov chain, in that order, we conjecture that:**

$$2^{-2[I(X;V)+I(Y;U)]} \geq (1-\rho^2)2^{-2I(U;V)} + \rho^2 2^{-2[I(X;U)+I(Y;V)]}.$$

By letting V be constant, we see that this inequality generalizes a well-known extremal result proved by Oohama in his work on the quadratic Gaussian one-helper problem. If valid, the conjecture would have some interesting consequences. For example, the converse for the quadratic Gaussian two-encoder source coding problem would follow from the converse for multiterminal source coding under logarithmic loss, thus unifying the two results under a common framework.

Although the conjecture remains open, we discuss both analytical and numerical evidence supporting its validity.

I. Introduction

This paper is a brief exposition on the following conjecture, its potential applications, and evidence supporting its validity. To this end, we propose:

Conjecture 1. *Suppose X, Y are jointly Gaussian, each with unit variance and correlation ρ. Then, for any U, V satisfying $U - X - Y - V$, the following inequality holds:*

$$2^{-2[I(Y;U)+I(X;V|U)]} \geq (1-\rho^2) + \rho^2 2^{-2[I(X;U)+I(Y;V|U)]}. \tag{1}$$

In the statement of Conjecture 1, we employ the conventional notation $U - X - Y - V$ to denote that U, X, Y, V form a Markov chain, in that order. Throughout this paper, X, Y will have the distribution given in the statement of the conjecture.

Our interest in Conjecture 1 stems from previous work by two of the present authors on multiterminal source coding under logarithmic loss [1]. In order to illustrate the connection between these problems, define $\mathcal{R} \subset \mathbb{R}^2$ as follows. Let $(R, I) \in \mathcal{R}$ if and only if there exists Q independent of X, Y, and U, V satisfying

$$R \geq I(X, Y; U, V|Q) \tag{2}$$
$$I \leq I(X; U, V|Q) + I(Y; U, V|Q), \tag{3}$$

and, conditioned on Q, the Markov relation $U - X - Y - V$.

This work was supported by the NSF Center for Science of Information under grant agreement CCF-0939370.

Next, let P_{XY} denote the joint distribution of X, Y, and assume $(X^n, Y^n) \sim \prod_{i=1}^n P_{XY}(x_i, y_i)$. For functions

$$\mathsf{f}_x : X^n \mapsto \mathsf{f}_x(X^n) \in \{1, 2, \dots, 2^{nR_x}\} \tag{4}$$
$$\mathsf{f}_y : Y^n \mapsto \mathsf{f}_y(Y^n) \in \{1, 2, \dots, 2^{nR_y}\}, \tag{5}$$

define

$$I(n, \mathsf{f}_x, \mathsf{f}_y) \triangleq \frac{1}{n}\Big(I(X^n; \mathsf{f}_x(X^n), \mathsf{f}_y(Y^n)) + I(Y^n; \mathsf{f}_x(X^n), \mathsf{f}_y(Y^n))\Big),$$

$$\mathsf{mmse}(X^n|\mathsf{f}_x, \mathsf{f}_y) \triangleq \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\Big(X_i - \mathbb{E}[X_i|\mathsf{f}_x(X^n), \mathsf{f}_y(Y^n)]\Big)^2\right], \tag{6}$$

and $\mathsf{mmse}(Y^n|\mathsf{f}_x, \mathsf{f}_y)$ in an analogous manner. These quantities satisfy the inequality

$$-\frac{1}{2}\log\Big(\mathsf{mmse}(X^n|\mathsf{f}_x, \mathsf{f}_y)\Big) - \frac{1}{2}\log\Big(\mathsf{mmse}(Y^n|\mathsf{f}_x, \mathsf{f}_y)\Big) \leq I(n, \mathsf{f}_x, \mathsf{f}_y), \tag{7}$$

which easily follows by convexity, the maximum entropy property of Gaussian random variables, and the memoryless property of X^n, Y^n.

An immediate consequence of the converse for multiterminal source coding under logarithmic loss is that $(R_x + R_y, I(n, \mathsf{f}_x, \mathsf{f}_y)) \in \mathcal{R}$, which easily follows from [1] and the corresponding entropy characterization result [2, Theorem 2].

Now, to show an interesting application of Conjecture 1, assume (1) holds. Combined with the fact that $(R_x + R_y, I(n, \mathsf{f}_x, \mathsf{f}_y)) \in \mathcal{R}$, elementary manipulations on (1) and (7) would reveal that

$$R_1 + R_2 \geq \frac{1}{2}\log\left[\frac{(1-\rho^2)\beta(D)}{2D}\right], \tag{8}$$

where we have defined

$$D \triangleq \mathsf{mmse}(X^n|\mathsf{f}_x, \mathsf{f}_y) \times \mathsf{mmse}(Y^n|\mathsf{f}_x, \mathsf{f}_y), \text{ and} \tag{9}$$

$$\beta(\xi) \triangleq 1 + \sqrt{1 + \frac{4\rho^2\xi}{(1-\rho^2)^2}} \tag{10}$$

for notational convenience. We note that (8) is precisely the sum-rate constraint for the quadratic Gaussian two-encoder source coding problem first established in the seminal work [3] by Wagner et al.

Thus, while we have only sketched the argument here, we hope the reader is convinced that the sum-rate constraint for the quadratic Gaussian two-encoder source coding problem would follow in a relatively straightforward manner from known results on compression under logarithmic loss and the conjectured extremal inequality (1). In fact, the entire converse (not only the sum-rate constraint) for the quadratic Gaussian two-encoder source coding problem would follow from Conjecture 1 and the characterization of the rate-distortion region for compression under log loss. Details are omitted due to space constraints.

On the term "Data Processing"

As the title suggests, we refer to (1) as a *data processing inequality* since it gives the upper bound

$$I(Y;U)+I(X;V|U) \leq -\frac{1}{2}\log\left[1-\rho^2+\rho^2 2^{-2[I(X;U)+I(Y;V|U)]}\right]. \quad (11)$$

By straightforward calculus, a simple corollary is, for example, the upper bound

$$I(Y;U) \leq \rho^2 I(X;U), \quad (12)$$

which falls into the category of so-called *strong data processing inequalities* (cf. [4], [5]). Since (1) is met with equality when U,X,Y,V are jointly Gaussian, (11) would provide the best possible data processing inequality of the form

$$I(Y;U)+I(X;V|U) \leq \psi(I(X;U)+I(Y;V|U)), \quad (13)$$

under our assumptions on U,X,Y,V.

II. Observations on Conjecture 1

There are many equivalent forms of Conjecture 1. It seems particularly useful to consider dual forms of Conjecture 1. For instance, one such form is stated as follows:

Conjecture 2. *Let X,Y be jointly Gaussian, each with unit variance and correlation ρ. For $\lambda > 1/\rho^2$, the infimum of*

$$I(X;U)+I(Y;V|U)-\lambda\Big(I(Y;U)+I(X;V|U)\Big) \quad (14)$$

taken over all U,V satisfying $U-X-Y-V$ is attained when U,X,Y,V are jointly Gaussian.

Note that we only conjecture that the minimum of (14) is attained by U,V which are jointly Gaussian with X,Y. Clearly, since mutual information is invariant under one-to-one transformations, there are minimizers of (14) which are non-Gaussian.

Let $F^\star_\lambda$ be the infimum of the functional (14) for fixed $\lambda > 1/\rho^2$. If Conjecture 1 were to hold, then straightforward computations reveal that $F^\star_\lambda$ would be given by

$$F^\star_\lambda = \frac{1}{2}\left[\log\left(\frac{\rho^2(\lambda-1)}{1-\rho^2}\right)-\lambda\log\left(\frac{\lambda-1}{\lambda(1-\rho^2)}\right)\right]. \quad (15)$$

It is interesting to note that we also have[1]

$$\inf_{U:U-X-Y}\Big\{I(X;U)-\lambda I(Y;U)\Big\} = \frac{1}{2}\left[\log\left(\frac{\rho^2(\lambda-1)}{1-\rho^2}\right)-\lambda\log\left(\frac{\lambda-1}{\lambda(1-\rho^2)}\right)\right]. \quad (16)$$

Since (14) can be rewritten as

$$\Big(I(X;U)-\lambda I(Y;U)\Big)+\Big(I(Y;V)-\lambda I(X;V)\Big) + (\lambda-1)I(U;V) \quad (17)$$

by Markovity, the conjecture implies an unexpected conservation property: either U and V can be optimized jointly in minimizing (14), or we can set V to be constant and only optimize over U (or vice versa). Assuming the conjecture is valid, both approaches yield the same optimal value, which suggests one should eliminating one of the variables is a viable proof strategy. Unfortunately, this has proved difficult to do. In any case, (16) and (17) yield the lower bound

$$F^\star_\lambda \geq \left[\log\left(\frac{\rho^2(\lambda-1)}{1-\rho^2}\right)-\lambda\log\left(\frac{\lambda-1}{\lambda(1-\rho^2)}\right)\right], \quad (18)$$

which reveals why we need only consider $\lambda > 1/\rho^2$ in Conjecture 2: for $\lambda \leq 1/\rho^2$, the infimum of (14) is zero.

Moving on, if we were to assume the conjecture were true, and let optimizing $U^\star, V^\star$ be of the form

$$U^\star = \rho_u X + Z_u \quad (19)$$
$$V^\star = \rho_v X + Z_v, \quad (20)$$

where $Z_u \sim N(0,1-\rho_u^2)$ and $Z_v \sim N(0,1-\rho_v^2)$ are independent additive Gaussian noises, then the parameters ρ_u, ρ_v should satisfy the following equation, which gives an intuitive sense for the tension between the conjectured optimizers $U^\star$ and $V^\star$:

$$(1-\rho^2)(1-\rho^2\rho_u^2\rho_v^2) = \rho^2(\lambda-1)(1-\rho_u^2)(1-\rho_v^2). \quad (21)$$

In particular, for given ρ,λ, there is a continuously parametrized family of conjectured optimizers.

III. Analytical Evidence Supporting Conjecture 1

There are several partial results which suggest the validity of Conjecture 1. To this end, note that Conjecture 1 generalizes the following well-known consequence of the conditional entropy power inequality to a longer Markov chain.

Lemma 1 (From [6])**.** *Suppose X,Y are jointly Gaussian, each with unit variance and correlation ρ. For any U satisfying $U-X-Y$, the following inequality holds:*

$$2^{-2I(Y;U)} \geq 1-\rho^2+\rho^2 2^{-2I(X;U)}. \quad (26)$$

Proof: Consider any U satisfying $U-X-Y$. Let Y_u, X_u denote the random variables X,Y conditioned on $U=u$. By Markovity and definition of X,Y, we have that $Y_u = \rho X_u + Z$,

[1]This is a consequence of Lemma 1 in Section III.

Given $P^{(0)}_{U|X}, P^{(0)}_{V|Y}$, **initialize** $P_{UVXY} := P^{(0)}_{U|X}P^{(0)}_{V|Y}P_{XY}$
for $i = 1, 2, ...$ **do**

$$P^{(i)}_{U|X}(u|x) := \frac{\exp\left\{\lambda \int P_{Y|X}(y|x)\log\left(P_{U|Y}(u|y)\right)dy - (\lambda-1)\int P_{V|X}(v|x)\log\left(P_{U|V}(u|v)\right)dv\right\}}{\int \exp\left\{\lambda \int P_{Y|X}(y|x)\log\left(P_{U|Y}(s|y)\right)dy - (\lambda-1)\int P_{V|X}(v|x)\log\left(P_{U|V}(s|v)\right)dv\right\}ds} \tag{22}$$

$$P_{UVXY} \leftarrow P^{(i)}_{U|X}P_{VXY} \tag{23}$$

$$P^{(i)}_{V|Y}(v|y) := \frac{\exp\left\{\lambda \int P_{X|Y}(x|y)\log\left(P_{V|X}(v|x)\right)dx - (\lambda-1)\int P_{U|Y}(u|y)\log\left(P_{V|U}(v|u)\right)du\right\}}{\int \exp\left\{\lambda \int P_{X|Y}(x|y)\log\left(P_{V|X}(s|x)\right)dx - (\lambda-1)\int P_{U|Y}(u|y)\log\left(P_{V|U}(s|u)\right)du\right\}ds} \tag{24}$$

$$P_{UVXY} \leftarrow P^{(i)}_{V|Y}P_{UXY} \tag{25}$$

Algorithm 1: Iterative procedure for solving the Euler-Lagrange equations (35)-(36).

where $Z \sim N(0, 1-\rho^2)$ is independent of X_u. Hence, the conditional entropy power inequality implies that

$$2^{2h(Y|U)} \geq \rho^2 2^{2h(X|U)} + 2\pi e(1-\rho^2) \tag{27}$$

$$= 2\pi e\rho^2 2^{-2I(X;U)} + 2\pi e(1-\rho^2). \tag{28}$$

From here, the lemma easily follows. ■

Lemma 1 can be applied to prove the following special case of Conjecture 1. This result subsumes many special cases that could be analyzed.

Proposition 1. *Suppose* X, Y *are jointly Gaussian, each with unit variance and correlation* ρ*. Let* U *be a random variable for which* $X|\{U = u\} \sim N(\mathbb{E}[X|U = u], \sigma^2)$ *for all* u*. If* $U - X - Y - V$*, then* (1) *holds.*

Proof: Since $X|\{U = u\} \sim N(\mathbb{E}[X|U = u], \sigma^2)$, we have $h(X|U) = h(X|u) = \frac{1}{2}\log(2\pi e\sigma^2)$, and therefore

$$I(X;U) = -\frac{1}{2}\log\sigma^2. \tag{29}$$

By Markovity, it is easy to see that $\mathrm{Var}(Y|U = u) = \rho^2\sigma^2 + (1-\rho^2)$, and hence

$$I(Y;U) = -\frac{1}{2}\log\left(\rho^2\sigma^2 + (1-\rho^2)\right). \tag{30}$$

Let X_u, Y_u, V_u denote the random variables X, Y, V conditioned on $U = u$, respectively. Define $\rho_{XY|u}$ to be the correlation coefficient between X_u and Y_u. It is readily verified that

$$\rho_{XY|u} = \frac{\rho\sigma}{\sqrt{\rho^2\sigma^2 + (1-\rho^2)}}, \tag{31}$$

which does not depend on the particular value of u. By plugging (29)-(31) into (1), we see that (1) is equivalent to

$$2^{-2I(X;V|U)} \geq (1-\rho^2_{XY|u}) + \rho^2_{XY|u}2^{-2I(Y;V|U)}. \tag{32}$$

For every u, X_u, Y_u are jointly Gaussian with correlation coefficient $\rho_{XY|u}$ and $X_u - Y_u - V_u$ form a Markov chain, hence Lemma 1 implies

$$2^{-2I(X_u;V_u)} \geq (1-\rho^2_{XY|u}) + \rho^2_{XY|u}2^{-2I(Y_u;V_u)}. \tag{33}$$

The desired inequality (32) follows by convexity of

$$\log\left[(1-\rho^2_{XY|u}) + \rho^2_{XY|u}2^{-2z}\right] \tag{34}$$

as a function of z. ■

IV. Numerical Evidence Supporting Conjecture 1

Conjecture 2 is amenable to numerical experiments. Dispensing with technicalities in favor of a cleaner exposition, some insight can be gained by deriving the Euler-Lagrange equations corresponding to the functional (14) and attempting to solve them. To this end, the Euler-Lagrange equations are given by:

$$\begin{aligned}\log P_{U|X}(u|x) &= \lambda\int P_{Y|X}(y|x)\log\left(P_{U|Y}(u|y)\right)dy \\ &\quad - (\lambda-1)\int P_{V|X}(v|x)\log\left(P_{U|V}(u|v)\right)dv - g(x),\end{aligned} \tag{35}$$

$$\begin{aligned}\log P_{V|Y}(v|y) &= \lambda\int P_{X|Y}(x|y)\log\left(P_{V|X}(v|x)\right)dx \\ &\quad - (\lambda-1)\int P_{U|Y}(u|y)\log\left(P_{V|U}(v|u)\right)du - h(y),\end{aligned} \tag{36}$$

where the functions $g(x)$ and $h(y)$ serve for the purpose of normalization so that $\int P_{U|X}(u|x)du = 1$ and $\int P_{V|Y}(v|y)dv = 1$, for each x and y, respectively. Note that (35) should hold for all x, u, and (36) should hold for all y, v.

Though characterizing the family of solutions to the nonlinear system of equations given by (35) and (36) may be difficult, it may be possible to compute a particular solution satisfying (35) and (36). In this case, the iterative procedure given by Algorithm 1 is a natural candidate for computing a stationary point. In fact, Algorithm 1 has the desirable property of monotone convergence, which we discuss in the next subsection.

A. Monotone Convergence of Algorithm 1

Let $I^{(i)}(X;U)$, $I^{(i)}(Y;V)$, *etc.* be mutual informations evaluated for the joint distribution $P^{(i)}_{UVXY} = P^{(i)}_{V|Y}P^{(i)}_{U|X}P_{XY}$, and define the corresponding functional:

$$F_\lambda(i) \triangleq I^{(i)}(X;U) - \lambda I^{(i)}(Y;U) + I^{(i)}(Y;V|U) - \lambda I^{(i)}(X;V|U). \quad (37)$$

Although a proof is omitted due to space constraints, for any $i \geq 1$, we have the inequality

$$F_\lambda(0) - F_\lambda(i) \geq \sum_{j=1}^{i} \left[D\left(P^{(j-1)}_{U|X} \middle\| P^{(j)}_{U|X}\right) + D\left(P^{(j)}_{U} \middle\| P^{(j-1)}_{U}\right) + D\left(P^{(j-1)}_{V|Y} \middle\| P^{(j)}_{V|Y}\right) + D\left(P^{(j)}_{V} \middle\| P^{(j-1)}_{V}\right) \right]. \quad (38)$$

Since $F_\lambda(i)$ is bounded from below according to (18), the sum on the right hand side of (38) must converge (assuming the initial test channels $P^{(0)}_{U|X}, P^{(0)}_{V|Y}$ satisfy $F_\lambda(0) < \infty$). In particular, (38) implies that $F_\lambda(i)$ decreases monotonically and converges to some limit, say $F_\lambda(\infty) \triangleq \lim_{i\to\infty} F_\lambda(i)$.

B. Numerical Experiments

Of course, given the infinite-dimensional nature of the problem, it is impractical to implement Algorithm 1 as stated. However, it is a simple matter to quantize the variables U, X, Y, V to a finite number of values. In this case, the integrals in updates (22) and (24) become sums over their respective variables, and (35) and (36) become KKT conditions for the corresponding discretized optimization problem.

The monotone convergence property discussed in the previous section carries over to the discretized variation of Algorithm 1. Therefore, by Pinsker's inequality, there exists a distribution[2]: $Q_{UVXY} = Q_{U|X}Q_{V|Y}P_{XY}$ such that $P^{(i)}_{UVXY} \xrightarrow{TV} Q_{UVXY}$ and hence, by continuity of mutual information, $F_\lambda(i) \searrow I_Q(X;U) - \lambda I_Q(Y;U) + I_Q(Y;V|U) - \lambda I_Q(X;V|U)$, where $I_Q(\cdot;\cdot)$ indicates mutual information evaluated with respect to the distribution Q_{UVXY}. Note that Q_{UVXY} will be a stationary point of the KKT conditions.

The plot shown in Figure 1 is a typical example of the evolution of $F_\lambda(i)$ when running the discretized variation of Algorithm 1. In particular, over thousands of trials with randomly instantiated test channels $P^{(0)}_{U|X}$ and $P^{(0)}_{V|Y}$, $F_\lambda(i)$ has always converged to the conjectured minimum value given by (15). Moreover, this convergence takes place quite rapidly (usually within a few iterations), as exemplified in Figure 1.

The fact that Algorithm 1 converges monotonically, combined with the empirical observation that it converges to the conjectured optimum without exception, suggests that traditional perturbation techniques for proving entropy power inequalities which construct a monotone path from any starting point to a global optimum (see, e.g., [7], [8]) could be adapted to our setting. Unfortunately, despite several attempts, the technical issue of preserving the long Markov chain has proven to be a significant barrier in doing so.

[2] Abusing notation for simplicity, we use P_{XY} to represent the distribution of the jointly Gaussian variables X, Y and their quantized counterparts.

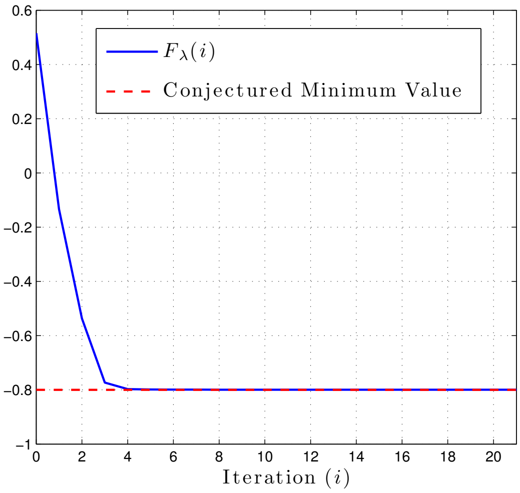

Fig. 1: Evolution of $F_\lambda(i)$ for $\rho = 0.5$, $\lambda = 3/\rho^2$. The variables U, X, Y, V were quantized to 101 evenly spaced values on the interval $[-6, 6]$, and $P^{(0)}_{U|X}, P^{(0)}_{V|Y}$ were randomly instantiated.

V. Concluding Remarks

In summary, Conjecture 1 represents an elegant and natural extension of Lemma 1. Given the widespread use of EPIs in proving converse results, we believe the conjectured extremal inequality (1) could be a useful tool with many applications. As a motivating example, we described an application to the quadratic Gaussian two-encoder source coding problem.

Acknowledgment

The authors gratefully acknowledge conversations with Vignesh Ganapathi-Subramanian, Alon Kipnis, Chandra Nair, and Kartik Venkat which contributed to this work.

References

[1] T. Courtade and T. Weissman, "Multiterminal source coding under logarithmic loss," *IEEE Trans. on Information Theory*, vol. 60, no. 1, pp. 740–761, 2014.

[2] T. Courtade, "Information masking and amplification: The source coding setting," in *Proc. of IEEE Int. Symp. on Information Theory (ISIT)*, 2012, pp. 189–193.

[3] A. Wagner, S. Tavildar, and P. Viswanath, "Rate region of the quadratic gaussian two-encoder source-coding problem," *Information Theory, IEEE Transactions on*, vol. 54, no. 5, pp. 1938–1961, 2008.

[4] 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.

[5] M. Raginsky, "Logarithmic Sobolev inequalities and strong data processing theorems for discrete channels," in *Proc. of IEEE Int. Symp. on Information Theory (ISIT)*, 2013, pp. 419–423.

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

[7] N. Blachman, "The convolution inequality for entropy powers," *IEEE Trans. on Information Theory*, vol. 11, no. 2, pp. 267–271, 1965.

[8] T. Liu and P. Viswanath, "An extremal inequality motivated by multiterminal information-theoretic problems," *IEEE Trans. on Information Theory*, vol. 53, no. 5, pp. 1839–1851, 2007.