Some mutual information inequalities

Milan S. Derpich and Eduardo I. Silva*

September 25, 2008

Abstract

The objective of this note is to report some potentially useful mutual information inequalities.

1 Preliminaries

Throughout this section, and unless otherwise stated, x, x_i, $i \in \mathbb{N}_0$, y, z and n are continuous random variables taking values in appropriate subsets of $\mathbb{R}^n$. We assume that they all have well defined probability density functions (PDFs), which we denote by f_x, f_{x_i}, f_y, f_z and f_n, respectively, and well defined joint PDFs denoted by f_{xy}, f_{xz}, etc.[1] We also use the notation $f_{x|y}$ to refer to the conditional PDF of x, given y. All definitions and results in this section are standard and can be found in [1].

Definition 1 (Differential entropy) *The differential entropy of x is defined via*[2]

$$h(x) \triangleq -\int f_x(u) \ln f_x(u) du. \tag{1}$$

The conditional differential entropy of x, given y, is defined via

$$h(x|y) \triangleq -\int f_{xy}(u,v) \ln f_{x|y}(u,v) du\, dv. \tag{2}$$

□□

The differential entropy has the following properties:

Fact 1 (Properties of h)

- $h(x|y) \leq h(x)$ *with equality if and only if x and y are independent.*
- $h(x+y|y) = h(x|y)$.
- *If $a \in \mathbb{R} \setminus \{0\}$, then* $h(ax) = h(x) + \ln|a|$.

*School of Electrical Engineering and Computer Science, The University of Newcastle, Australia, 2008. Milan.Derpich@studentmail.newcastle.edu.au

[1] We will seldom need to make a distinction between a random variable and its realization values. Thus, we introduce at this moment no additional notation for the *values* of x, y, z or n.

[2] It is understood that the integrals are defined over the support of the functions involved.

- $h(x_0, \cdots, x_{n-1}) = \sum_{i=0}^{n-1} h(x_i|x_0, \cdots, x_{i-1})$ *(this property is called* chain rule for differential entropy*).*
- *If x and y are independent, then $e^{2h(x+y)} \geq e^{2h(x)} + e^{2h(y)}$ ($e^{2h(x)}$ is called* entropy power *of x. This property is called* entropy power inequality*.)*

□□□

Definition 2 (Mutual information) *The mutual information between x and y is defined via*

$$I(x;y) \triangleq \int f_{xy}(u,v) \ln \frac{f_{xy}(u,v)}{f_x(u)f_y(v)} du\, dv \tag{3}$$

The conditional mutual information between x and y, given z, is defined via

$$I(x;y|z) \triangleq \int f_{xyz}(u,v,w) \ln \frac{f_{xyz}(u,v,w)f_z(w)}{f_{xz}(u,w)f_{yz}(v,w)} du\, dv\, dw. \tag{4}$$

□□

Mutual information has the following properties:

Fact 2 (Properties of I)

1. $I(x;y) = h(x) - h(x|y) = h(y) - h(y|x) = I(y;x)$.
2. $I(x;y|z) = h(x|z) - h(x|y,z) = h(y|z) - h(y|x,z) = I(y;x|z)$.
3. $I(x;y) \geq 0$ *with equality if and only if x and y are independent.*
4. $I(x,y;z) = I(x;z) + I(y;z|x)$ *(*chain rule of mutual information*).*

□□□

Definition 3 (Markov chain) *The random variables x, y and z are said to form a Markov chain (in that order) if and only if $f(x,z|y) = f(x|y)f(z|y)$, i.e., if and only if x and z are conditionally independent given y. If that is the case, we write*

$$x \leftrightarrow y \leftrightarrow z. \tag{5}$$

□□

Theorem 1 (Data processing inequality) *If $x \leftrightarrow y \leftrightarrow z$, then $I(x;y) \geq I(x;z)$. Equality holds if and only if, in addition, $x \leftrightarrow z \leftrightarrow y$.* □□□

Definition 4 (Divergence between PDFs) *The divergence of the distribution of x with respect to the distribution of y (in short, the divergence between x and y) is defined by*[3]

$$D(x||y) \triangleq \int f_x(u) \ln \frac{f_x(u)}{f_y(u)} du. \tag{6}$$

□□

Relevant properties of $D(\cdot||\cdot)$ are summarized below:

Fact 3 (Properties of D)

- $D(x||y) \geq 0$ *with equality if and only if* $f_x = f_y$ *almost everywhere*[4] *(a.e.).*
- *If* x_G *is a second order Gaussian random variable and* x *is any other random variable with the same mean and covariance matrix, then*

$$D(x||x_G) = h(x_G) - h(x) = D(ax||ax_G), \tag{7}$$

where $a \in \mathbb{R} \setminus \{0\}$ *is any real number.*

□□□

Remark 1 (Conditional divergence) *It will prove useful to consider an extension of the definition of divergence. Given two joint distributions* f_{xy} *and* f_{wz}*, we define the conditional divergence between them*[5] *via*

$$D(x|y||w|z) \triangleq \int f_{xy}(u,v) \ln \frac{f_{x|y}(u,v)}{f_{w|z}(u,v)} dudv. \tag{8}$$

It is possible to show that the following holds:

- $D(x|y||w|z) \geq 0.$
- *If* x_G *and* y_G *are jointly Gaussian random variables having joint PDF* $f_{x_G y_G}$*, and* x *and* y *are arbitrary random variables having a joint PDF* f_{xy} *with the same first and second order moments as* $f_{x_G y_G}$*, then*

$$D(x|y||x_G|y_G) = h(x_G|y_G) - h(x|y). \tag{9}$$

□□

We end this section with an extension of the notion of differential entropy to random processes.

[3] Also called the Kullback-Leibler "distance" between the distribution of x and the distribution of y.
[4] i.e., $f_x(u) = f_y(u)$ except (perhaps) on a countable set of reals.
[5] Also known as conditional relative entropy (see [1]).

Definition 5 *(Differential entropy rate) Consider an asymptotically stationary process x. The differential entropy rate of x is defined by*[6]

$$\bar{h}(x) \triangleq \lim_{k\to\infty} \frac{h(x^{k-1})}{k}. \tag{10}$$

□□

If x is stationary, then it is clear that $\bar{h}(x) \leq h(x(k))$, with equality if and only if x is a sequence of independent random variables (recall Fact 1).

Theorem 2 (Differential entropy rate (see, e.g., [2, 3])) *If a stationary process $\hat{x}$ is filtered by a stable filter having frequency response $H(e^{j\omega})$, then the filter output x has an entropy rate given by*

$$\bar{h}(x) = \bar{h}(\hat{x}) + \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln\left|H(e^{j\omega})\right| d\omega. \tag{11}$$

□□□

2 Results

This section presents the main results of this note.

Lemma 1 *Consider the situation depicted in Figure 1, where x and n are m-dimensional random variables that have arbitrary distributions. If x and n are independent, and x_G and n_G denote independent m-dimensional Gaussian random variables having the same mean and covariance matrix as x and n, respectively, then*

$$I(x;y) \leq I(x_G;y_G) + D(n||n_G), \tag{12}$$

with equality if and only if x and n are jointly Gaussian.

Proof: Using Facts 2 and 1, the independence of x, n and x_G, n_G, and the definition of $D(\cdot||\cdot)$, it is easy to see that

$$\begin{aligned} I(x;y) - I(x_G;y_G) &= h(y) - h(y|x) - h(y_G) + h(y_G|x_G) \\ &= h(x+n) - h(x+n|x) - h(x_G+n_G) + h(x_G+n_G|x_G) \\ &= h(n_G) - h(n) - h(x_G+n_G) + h(x+n) \\ &\overset{(a)}{=} D(n||n_G) - D(x+n||x_G+n_G) \\ &\leq D(n||n_G), \end{aligned} \tag{13}$$

where the last inequality follows from Fact 3. The result is now immediate. □□□

[6] x^i is shorthand for $x(0), x(1), \cdots, x(i)$.

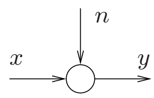

Figure 1: Additive channel.

Lemma 2 *Consider the situation depicted in Figure 1, where x and n are m-dimensional random variables, x is Gaussian and n has an arbitrary distribution. If n_G denotes an m-dimensional Gaussian random variable, jointly Gaussian with x, having the same mean and covariance matrix as n, and such that the cross-covariance between n and x equals the cross-covariance between n_G and x, then*

$$I(x; x + n_G) \leq I(x; x + n), \tag{14}$$

with equality if the covariance matrix of $x+n$ is non-singular, and n is Gaussian and jointly Gaussian with x.

Proof: Using Fact 2 it is possible to write

$$I(x; x + n) - I(x; x + n_G) = h(x|x + n_G) - h(x|x + n). \tag{15}$$

Use of the facts in Remark 1 the first part of the result follows. Clearly, if n is Gaussian, then equality holds in (14). The proof of the converse can be found in [4]. □□□

Lemma 3 *Consider the situation depicted in Figure 1, where x and n are independent scalar random variables with arbitrary distributions. If x_G and n_G denote independent scalar Gaussian random variables having the same mean and covariance matrix as x and n, and $D(x||x_G) \leq D(n||n_G)$, then*

$$D(x + n||x_G + n_G) \leq D(n||n_G) \quad \textit{and} \quad I(x_G; x_G + n_G) \leq I(x; x + n), \tag{16}$$

with equality if and only if x and n are jointly Gaussian.

Proof: We will use the proof of Lemma 1. If the right hand side in equality (a) in (13) were positive, then the result would be true. Thus, we will start examining the difference $D(n||n_G) - D(x + n||x_G + n_G)$:

$$\begin{aligned} D(n||n_G) - D(x + n||x_G + n_G) &= h(n_G) - h(n) - h(x_G + n_G) + h(x + n) \\ &= h(x + n) - h(n) + \frac{1}{2} \ln \frac{2\pi e \sigma_{n_G}^2}{2\pi e \left(\sigma_{x_G}^2 + \sigma_{n_G}^2\right)} \\ &= h(x + n) - h(n) - \frac{1}{2} \ln \left(1 + \frac{\sigma_{x_G}^2}{\sigma_{n_G}^2}\right), \end{aligned} \tag{17}$$

where we have used Fact 3, the independence of x_G, n_G and Gaussianity. On the other hand, the entropy power inequality allows one to conclude that, since x, n are independent,

$$h(x + n) - h(n) \geq \frac{1}{2} \ln \left(e^{2h(x)} + e^{2h(y)}\right) - h(n) = \frac{1}{2} \ln \left(1 + \frac{e^{2h(x)}}{e^{2h(n)}}\right) \tag{18}$$

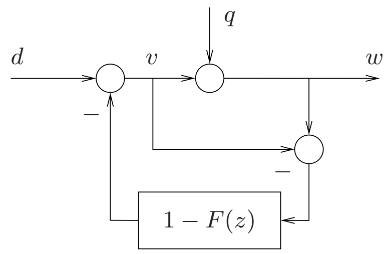

Figure 2: Feedback system considered in Lemma 4.

Use of (18) in (17) yields

$$D(n||n_G) - D(x+n||x_G+n_G) \geq M \triangleq \frac{1}{2}\ln\left(1+\frac{e^{2h(x)}}{e^{2h(n)}}\right) - \frac{1}{2}\ln\left(1+\frac{\sigma^2_{x_G}}{\sigma^2_{n_G}}\right) \tag{19}$$

and, since the variance of the Gaussian and non-Gaussian random variables is the same, we have from (19) that

$$M \geq 0 \Leftrightarrow \frac{e^{2h(x)}}{\sigma_x^2} \geq \frac{e^{2h(n)}}{\sigma_n^2} \overset{(a)}{\Leftrightarrow} h\left(\frac{x}{\sigma_x}\right) \geq h\left(\frac{n}{\sigma_n}\right) \overset{(b)}{\Leftrightarrow} \frac{1}{2}\ln 2\pi e - D(x||x_G) \geq \frac{1}{2}\ln 2\pi e - D(n||n_G) \Leftrightarrow D(x||x_G) \leq D(n||n_G), \tag{20}$$

where (a) follows from Fact 1 and (b) from Facts 3 and 1, and the fact that the variance of the Gaussian and non-Gaussian random variables is the same. The result follows using (20) and (19) in equality (a) in (13). □□□

Definition 6 *Consider two random processes v and w. We define (if the defining limits exist) the* mutual information rate *between v and w as*

$$\bar{I}_\infty(v;w) \triangleq \lim_{k\to\infty}\frac{1}{k}I(v^{k-1};w^{k-1}), \tag{21}$$

and the average mutual information between v and w *as*

$$I_\infty(v\to w) \triangleq \lim_{k\to\infty}\frac{1}{k}\sum_{i=0}^{k-1} I(w(i);v^i|w^{i-1}). \tag{22}$$

□□

Lemma 4 *Consider the feedback system in Figure 2, where $1-F(z)$ is stable and strictly proper (i.e., $\lim_{z\to\infty}F(z)=1$), d is a random process, and q is an i.i.d. sequence that is independent of d and of the initial state of $F(z)$. Then,*

$$\bar{I}_\infty(d;w) = I_\infty(v\to w) - \sum_{i=1}^{n_F}\log\left|p_i^F\right|, \tag{23}$$

where $\{p_1^F, \cdots, p_{n_F}^F\}$ *denotes the set of non minimum phase zeros of* $F(z)$. □□□

Proof: By definition of mutual information rate and the chain rule of mutual information we have that

$$\bar{I}_\infty(d; w) = \lim_{k\to\infty} \frac{1}{k} I(d^{k-1}; w^{k-1}) = \lim_{k\to\infty} \frac{1}{k} \sum_{i=0}^{k-1} I(w(i); d^{k-1}|w^{i-1}). \tag{24}$$

Since w depends causally on d, it follows that $I(w(i); d^{k-1}|w^{i-1}) = I(w(i); d^i|w^{i-1})$. Thus,

$$\bar{I}_\infty(d; w) = I_\infty(d \to w). \tag{25}$$

Define $n \triangleq w - d$ and note that

$$n = F(z)q. \tag{26}$$

We first note that

$$\begin{aligned} I(w(i); d^i|w^{i-1}) - I(w(i); v^i|w^{i-1}) &\stackrel{(a)}{=} h(w(i)|w^{i-1}; v^i) - h(w(i)|w^{i-1}, d^i) \\ &\stackrel{(b)}{=} h(w(i)|w^{i-1}; v^i) - h(n(i)|w^{i-1}, d^i) \\ &\stackrel{(c)}{=} h(w(i)|w^{i-1}; v^i) - h(n(i)|n^{i-1}, d^i) \\ &\stackrel{(c)}{=} h(w(i)|w^{i-1}; v^i) - h(n(i)|n^{i-1}), \end{aligned} \tag{27}$$

where (a) follows from Fact 2, (b) follows from the definition of n and Fact 1, (c) follows from the fact that, by definition of n, $M \leftrightarrow (w^{i-1}, d^i) \leftrightarrow (n^{i-1}, d^i)$ for every random variable M, and (d) follows from the fact that, since d is independent of q and of the initial state of $F(z)$, d is independent of n and, thus, $n(i) \leftrightarrow n^{i-1} \leftrightarrow d^i$ holds.

We also have that

$$\begin{aligned} h(w(i)|w^{i-1}, v^i) &\stackrel{(a)}{=} h(v(i) + q(i)|w^{i-1}, v^i) \\ &\stackrel{(b)}{=} h(q(i)|q^{i-1}, v^i) \\ &\stackrel{(c)}{=} h(q(i)|q^{i-1}), \end{aligned} \tag{28}$$

where (a) follows from the definition of variables in Figure 2, (b) follows from Fact 1 and the fact that, by definition, $M \leftrightarrow (w^{i-1}, v^i) \leftrightarrow (q^{i-1}, d^i)$ for every random variable M, and (c) follows from the fact that both the initial state of $F(z)$ and d being independent of q, q being i.i.d., and $F(z)$ being strictly proper guarantees that $q(i) \leftrightarrow q^{i-1} \leftrightarrow v^i$.

From (25), (27), (28) and Fact 1 it follows that

$$\bar{I}_\infty(d; w) - I_\infty(v \to w) = \bar{h}(q) - \bar{h}(n). \tag{29}$$

Use of Theorem 2, (26) and the Bode integral theorem (see, e.g., [5]) yields the result. □□□

References

[1] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. John Wiley and Sons, Inc., 2nd edition, 2006.

[2] C.E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656, July, October 1948.

[3] Athanasios Papoulis. *Probability, random variables and stochastic process*. McGraw Hill Book Company, New York, 3rd edition, 1991.

[4] M.S. Derpich. *Optimal Source Coding with Signal Transfer Function Constraints*. PhD thesis, School of Electrical Engineering and Computer Science, The University of Newcastle, Australia, 2008.

[5] M.M. Seron, J.H. Braslavsky, and G.C. Goodwin. *Fundamental Limitations in Filtering and Control*. Springer, London, 1997.