A Renyi entropy convolution inequality with application

J.-F. Bercher $^{(1,2)}$ *and C. Vignat* $^{(1)}$

(1) Laboratoire Systèmes de Communications
Université de Marne la Vallée and ENST URA 820
93 166 Noisy-le-Grand, FRANCE
vignat@univ-mlv.fr

(2) Laboratoire Signaux et Télécoms,
Groupe ESIEE
93 162 Noisy-le-Grand, FRANCE
jf.bercher@esiee.fr

ABSTRACT

We present a convolution inequality for Renyi's entropies. An example of application is given in the context of blind deconvolution: an optimization procedure based on the inequality for the quadratic entropy is presented and illustrated.

1 INTRODUCTION

Since the pioneering work by Shannon [1], entropy appears as an interesting tool in many areas of data processing. Donoho, in his paper [4], noticed that entropy based approaches may be valuable for deconvolution of real seismic data. However, the use of Shannon entropy raises the difficult problem of its estimation and analytical manipulation. Some attempts and comparisons can be found in [2]. Hence, this entropy is not amenable to basic estimation methods. In 1953, Renyi introduced a wider class of entropies kwown as Renyi entropies, and defined as

$$H_r(X) = \frac{1}{1-r}\log\int f_X^r(x)dx,$$

where r is real positive. In the special case $r=1$, the Renyi entropy reduces to Shannon entropy:

$$H_1(X) = -\int f_X(x)\log f_X(x)dx.$$

These functionals share the major properties of Shannon's entropy since they were obtained by extending one of the fundating postulates of the notion of entropy. Moreover, at least in the case $r=2$, as remarked in [8], the Renyi entropy can be analytically expressed when the underlying law $f_X(x)$ is estimated using a kernel estimate (for instance using a rectangular window in [8], or a gaussian kernel in [9]): this makes its use attractive in a real context. However, until now, these generalized entropies were used in a restricted number of areas such as database retrieval and image processing.

In this paper, we show that a convolution inequality on the Shannon entropy (the entropy power inequality) can also be expressed for the Renyi class of entropies, based on the extended Young's inequality. The outline of this paper is the following: in the first part, we derive an original convolution inequality in the case of Renyi entropies and characterize the case of equality. In a second part, we show that this inequality can be applied in a context of blind deconvolution.

2 ENTROPIC INEQUALITIES

2.1 Shannon case

In the case of Shannon entropy, convolution of two independent random variables X and Y leads to the entropy power inequality, first stated by Shannon [1, Theorem 15 and Appendix 6], see also [5]:
Entropy power inequality: if X and Y are two independent random variables with entropies $H_1(X)$ and $H_1(Y)$, then

$$e^{2H_1(X+Y)} \geq e^{2H_1(X)} + e^{2H_1(Y)},$$

with equality if and only if X and Y are gaussian variables, or one of them is deterministic. This inequality has been extended to the multivariate case in [3].

The extension of such inequality to the general Renyi entropies (i.e. $r \neq 1$) remains an open problem. We give here an alternate inequality for the entropy of convolved variables, derived from the extended Young's inequality.

2.2 Entropic Young's inequality

The extended Young's inequality [6] states as follows:

Theorem 1 *Let $p,q,r > 0$ satisfy $1/p + 1/q = 1 + 1/r$, and let $f \in L^p\left(R^N\right)$ and $g \in L^q\left(R^N\right)$ be non-negative functions. Let also $C_t = \sqrt{\frac{t^{1/t}}{|t'|^{1/t'}}}$.*

$$\text{If } p,q,r \geq 1: \quad \|f*g\|_r \leq \left(\frac{C_pC_q}{C_r}\right)^N \|f\|_p\,\|g\|_q \quad (1)$$

$$\text{If } p,q,r \leq 1: \quad \|f*g\|_r \geq \left(\frac{C_pC_q}{C_r}\right)^N \|f\|_p\,\|g\|_q \quad (2)$$

Moreover, when $N=1$ and $p,q \neq 1$, there is equality in (1) or (2) if and only if $f(x) = \exp\left(-|p'|\,x^2\right)$ and $g(x) = \exp\left(-|q'|\,x^2\right)$, with $1/p + 1/p' = 1/q + 1/q' = 1$.

In the case of the sum of two independent random vectors, taking $r = p$ and $q = 1$ and noting that $||g||_1 = 1$, we then obtain,

$$H_r(X+Y) \geq H_r(X),$$

and also, exchanging X and Y,

$$H_r(X+Y) \geq H_r(Y), \quad (3)$$

so that finally

$$H_r(X+Y) \geq max(H_r(X), H_r(Y)). \tag{4}$$

The case of equality is more intricate and happens if and only if X or Y is a deterministic vector [6, 7].

2.3 Case of filtered data

In many applications, available data result from a mixture. For instance, consider a filter with impulse response f and an independent identically distributed (iid) input $X(n)$. Its output is $Z(n) = \sum_i f_i X(n-i)$.

Let us first consider the case of Shannon entropy: the classical result on the entropy of rescaled variables, that is $H_1(aX) = H_1(X) + \log|a|$, and the assumption of stationarity give

$$\begin{aligned} H_1(f_i X(n-i)) &= H_1(X(n-i)) + \log|f_i| \\ &= H_1(X) + \log|f_i|. \end{aligned} \tag{5}$$

Now, the entropy power inequality gives

$$e^{2H_1(Z)} \geq \sum_i e^{2H_1(X)+2\log|f_i|} = e^{2H_1(X)} \sum_i |f_i|^2,$$

or equivalently,

$$H_1(Z) \geq H_1(X) + \frac{1}{2}\log\sum_i |f_i|^2, \tag{6}$$

with equality if and only if either $X(n)$ is gaussian or if $Z(n) = X(n-k)$, for some k, that is, the filtering operation is a pure delay.

Let us now consider the Renyi case. The scaling property is unchanged so that (5) remains true. Then, from inequality (3), we deduce

$$H_r(Z) \geq H_r(X) + \frac{1}{2}\log|f_i|^2, \quad \forall i \tag{7}$$

with equality if and only the filter is a pure delay.

Observe that this last relation still holds in the multivariate case, where $X(n)$ is a sequence of iid vectors (not necessarily with independent components) and where $|f_i|$ should be understood as the absolute value of the determinant of matrix coefficient f_i.

In the Shannon case, inequality (6) amounts to the well known result that the linear transformation of a random i.i.d. sequence increases its entropy, under a variance preserving constraint $||f||_2 = 1$. Inequality (7) shows that the same result holds in Renyi's case, but under the constraint $||f||_\infty = \max_i |f_i| = 1$.

In the following section, we show how this tool can be applied to the problem of blind MIMO deconvolution.

3 APPLICATION TO BLIND MIMO DECONVOLUTION

3.1 Model and principle

The problem of blind deconvolution consists in recovering the i.i.d. input $X(n)$ and possibly the parameters of a filter G from the sole observation of its output $Y(n)$.

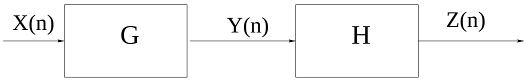

Figure 1: Blind deconvolution setup.

Here, $X(n)$ and $Y(n)$ are respectively $p \times 1$ and $q \times 1$ random vectors, with $q > p$. Let us also denote by f_k the L_f samples ($p \times p$) of the impulse response of the compound filter, $f_k = [h * g]_k$.

According to the last remark of section 2.3, the entropy of the output $Z(n)$ can not be lower than the entropy of $X(n)$ provided that either $||f||_2 = 1$ (Shannon) or $||f||_\infty = 1$ (Renyi). In both cases, the minimum entropy of the output is reached precisely when $Z(n)$ and $X(n)$ coincide, up to inherent scale and delay indeterminacies. [1]

The deconvolution procedure then simply consists in adjusting the "equalizing" filter H with input $Y(n)$, such that its output $Z(n)$ has minimum (estimated) entropy.

In order to minimize the entropy, knowledge of its gradient with respect to the matrix coefficients h_k is helpful. Using a kernel estimate of the probability density, we can obtain an analytical expression of this gradient in the case of the quadratic entropy ($r = 2$).

3.2 Gaussian kernel estimation of the quadratic entropy

Based on the observation of N samples $\{Y(1), \ldots, Y(N)\}$ of $Y(n)$, we estimate the density f_Y of Y as

$$\hat{f}_Y(Y) = \frac{1}{N}\sum_{k=1}^{N} \phi_{Y(k),\sigma^2 I}(Y) \tag{8}$$

where $\phi_{\mu,C}$ denotes a Gaussian kernel with mean μ and covariance matrix C.

Introducing the matrix

$$H = [h_0 \| h_1 \| \ldots \| h_{L-1}]$$

and a $(qL \times 1)$ vector $\bar{Y}(n)$ that collects L time samples of $Y(n)$ as

$$\bar{Y}(n) = [Y(n), Y(n-1), \ldots, Y(n-L+1)]^T,$$

$Z(n)$ can be expressed under a matrix form as $Z(n) = H\bar{Y}(n)$. Thus a natural gaussian kernel estimate of the law of $Z(n)$ follows as:

$$\hat{f}_Z(Z) = \frac{1}{N}\sum_{l=1}^{N} \phi_{Z(l),\sigma^2 HH^T}(Z).$$

The estimate of the quadratic entropy can then be ex-

[1] Indeterminacies. If A is a unitary matrix, in the case where the components are independent then $H_r(AX(n-k)) = H_r(X(n))$, where $X(n-k) = [x_1(n-k_1)\ldots x_m(n-k_m)]^T$; otherwise the delay must be the same on each component.

plicitely computed as

$$H_2\left(Z\right) = -\log \int p_Z^2\left(z\right) dz$$
$$= -\log \left\{ \frac{1}{N^2} \sum_{k,l=1}^{N} \phi_{0,2\sigma^2 P}\left(Z\left(k\right) - Z\left(l\right)\right) \right\} \quad (9)$$

with $P = HH^T$ and the help of the convolution equality

$$\int \phi_{Z_i,\sigma^2 P}\left(z\right) \phi_{Z_j,\sigma^2 P}\left(z\right) dz = \phi_{0,2\sigma^2 P}\left(Z\left(k\right) - Z\left(l\right)\right).$$

3.3 Gradient of the (estimated) quadratic entropy

Computation of the matrix gradient

$$\frac{\partial H_2(Z)}{\partial H}$$

of the quadratic entropy (9) is a formidable adventure. Let us recall that the matrix derivative $\partial A / \partial B$ is a partioned matrix whose (i, j) block is

$$\frac{\partial A}{\partial B_{ij}} = \left[\frac{\partial A_{kl}}{\partial B_{ij}}\right]$$

(see [10] for a review of useful matrix gradient identities).

Let us begin by some notations: we note $R = 2\sigma^2 P = 2\sigma^2 HH^T$, $Z_{k,l} = Z\left(k\right) - Z\left(l\right)$, $Y_{k,l} = Y\left(k\right) - Y\left(l\right)$ and $\Pi = H^T \left(HH^T\right)^{-1} H$. We also introduce the matrices U and $\bar{U}$ as

$$\frac{\partial H}{\partial H} = \bar{U}, \qquad \frac{\partial H^T}{\partial H} = U.$$

Now write the entropy in (9) as $H_2\left(Z\right) = H^{(1)} + H^{(2)}$ with

$$H^{(1)} = \log \left\{ N^2 \left(2\pi\right)^{p/2} \det R \right\} \text{ and } H^{(2)} = \log \sum_{k,l=1}^{N} E_{k,l},$$

where $E_{k,l} = e^{-\frac{1}{2}(Z_{k,l})^T R^{-1} (Z_{k,l})}$. In order to compute the gradient we have derived the following lemmas. Lemma 2 enables to compute the gradient of $H^{(1)}$:

Lemma 2 *For $P = HH^T$, we have*

$$\frac{d}{dH} \log\left|\det P\right| = 2P^{-1}H$$

The delicate step in the computation of the gradient of the second term $H^{(2)}$ is the evaluation of the derivative of the quadratic form $z^T R^{-1} z$, with $z = Hy$. We use the two following lemmas:

Lemma 3 *The derivative of a product of matrices writes*

$$\frac{\partial ABC}{\partial D} = \frac{\partial AB}{\partial D}\left(I_n \otimes C\right) + \left(I_m \otimes AB\right)\frac{\partial C}{\partial D}$$
$$= \frac{\partial A}{\partial D}\left(I_n \otimes BC\right) + \left(I_m \otimes A\right)\frac{\partial B}{\partial D}\left(I_n \otimes C\right) + \left(I_m \otimes AB\right)\frac{\partial C}{\partial D}$$

if D is $m \times n$, (see [10]).

Lemma 4

$$\frac{\partial z}{\partial H} = \bar{U}_{p,qL}\left(I_{qL,qL} \otimes y\right) = vec\{I_p\} \otimes y^T$$
$$\frac{\partial z^T}{\partial H} = \left(I_{p,p} \otimes y^T\right) U_{p,qL} = y^T \otimes I_p$$

Furthermore, the derivative of the inverse of a matrix is given in [10] and leads to

$$\frac{\partial P^{-1}}{\partial H} = -\left(I_p \otimes P^{-1}\right)\frac{\partial P}{\partial H}\left(I_{qL} \otimes P^{-1}\right),$$
$$\frac{\partial P}{\partial H} = \bar{U}_{p,qL}\left(I_{qL} \otimes H^T\right) + \left(I_p \otimes H\right) U_{p,qL}.$$

We finally obtain Lemma 5 using Lemmas 3 and 4 and simplifications from Kronecker products algebra.

Lemma 5

$$\frac{\partial z^T P^{-1} z}{\partial H} = \frac{\partial z^T}{\partial H}\left(I_{qL} \otimes R^{-1} z\right) + \left(I_p \otimes z^T\right)\frac{\partial P^{-1}}{\partial H}\left(I_{qL} \otimes z\right)$$
$$+ \left(I_p \otimes z^T P^{-1}\right)\frac{\partial z}{\partial H}$$

so that

$$\frac{\partial z^T P^{-1} z}{\partial H} = 2\left(P^{-1} H y y^T \left(I_{qL} - \Pi\right)\right).$$

The gradient of $H^{(2)}$ then writes:

$$\frac{d}{dH} H^{(2)} = R^{-1} H \sum_{k,l} \frac{E_{k,l}}{\sum_{pq} E_{pq}} Y_{k,l} Y_{k,l}^T \left(I_{qL} - \Pi\right).$$

Using Lemma 2 we finally obtain

$$\frac{\partial H_2(Z)}{\partial H} = 2P^{-1}H \times$$
$$\left[I_{qL} + \frac{1}{4\sigma^2} \sum_{k,l} \frac{E_{k,l}}{\sum_{pq} E_{pq}} Y_{k,l} Y_{k,l}^T \left(I_{qL} - \Pi\right) \right]. \quad (10)$$

3.4 Simulation results

Based on the Renyi convolution inequality (7) and on the discussion in section 3.1, we use the estimate of the quadratic entropy and its gradient in a blind deconvolution procedure. It amounts to minimize $H_2(Z)$ with respect to the coefficients of the equalizer h while constraining one of the coefficients, say f_k, such that $|f_k| = 1$. This constraint ensures that the minimization of $H_2(Z)$ does not lead to a trivial null output and that equality in (7) is reached for a pure delay filter f. Because the f_k depend both on the equalizer and on the unknown filter g, the constraint $f_k = 1$ can not be ensured in practice. However, one can ensure $|f_k|$ =constant for some k, so that the right hand side of (7) remains constant and it makes sense to minimize $H_2(Z)$. A solution is to fix h_0 so that $f_0 = h_0 g_0$ is also fixed. Then, the RHS of (7) is fixed, and blind deconvolution is achieved up to a scaling matrix f_0 (instantaneous mixture). Thus, the resulting output will have to be post-processed by some standard method of instantaneous source separation.

Note that constraining h_0 may not lead to a robust constraint, especially if $f_0 = h_0 g_0$ is small or (and) badly-conditioned, since the deconvolution procedure may result in a filter with very small coefficients (thus with high sensitivity) and since the quality of the post-processing is highly related to the conditioning of f_0. However, a bad choice of h_0 will be easy to detect: in such a case, $H_2(Z)$ goes to $-\infty$. In practice, the constraint is ensured by using the gradient with respect to all coefficients but h_0 so that h_0 remains to its initial value.

The blind deconvolution procedure was tested in the case of SIMO and MIMO systems. Although the procedure may suffer of spurious local minima, correct blind deconvolution is achieved in most cases. In the experiments below, we used $N = 100$ samples of observations $Y(n)$, and chose $\sigma^2 = 0.1$ for the gaussian kernels bandwidth.

Experiment 1 — We consider a SIMO case, with $q = 2$, where signals $Y_1(n)$ and $Y_2(n)$ are AR filtered outputs (respectively minimum and maximum phase) of a binary signal. We used

$$\begin{aligned} Y_1(n) &= -0.5Y_1(n-1) - 0.2Y_1(n-2) + X(n), \\ Y_2(n) &= -2Y_2(n-1) - 1.5Y_2(n-2) + X(n). \end{aligned}$$

Note that $Y_1(n)$ can be deconvolved using predictive deconvolution, but $Y_2(n)$ can not. Using the Renyi blind deconvolution procedure with $L = 15$, the SIMO system is correctly inversed: the impulse response of the equivalent filter f, is reported figure 2.

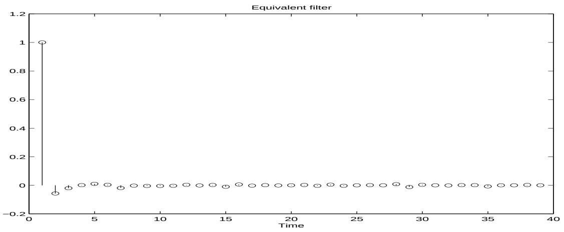

Figure 2: Equivalent impulse response f for experiment 1.

Experiment 2 — An important property of SIMO systems is that FIR SIMO filters can be (perfectly) equalized by a FIR filter, under some technical conditions. So we now consider the equalization of a FIR SIMO filter, with $q = 2$. The transfer function is

$$g(z) = \begin{bmatrix} 1 \\ 0.82 \end{bmatrix} + \begin{bmatrix} 0.7 \\ 1 \end{bmatrix} z^{-1} + \begin{bmatrix} 0.5 \\ 0.4 \end{bmatrix} z^{-2} + \begin{bmatrix} 0.5 \\ -0.3 \end{bmatrix} z^{-3}$$

Using an equalizer of length $L = 4$, the system is also correctly inversed, as indicated figure 3.

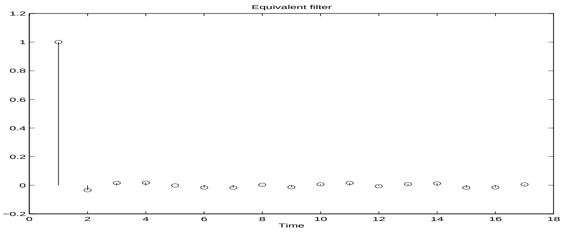

Figure 3: Equivalent impulse response f for experiment 2.

Experiment 3 — Last, we report the results obtained for a MIMO system, with $p = 2$ inputs and $q = 3$ outputs. The transfert function considered is

$$g(z) = \begin{bmatrix} 1 & -0.7 \\ 0.82 & 1 \\ 0.7 & 0.3 \end{bmatrix} + \begin{bmatrix} 0.5 & 0.5 \\ 0.4 & -0.3 \\ 0.1 & -0.2 \end{bmatrix} z^{-1} + \begin{bmatrix} 0.3 & -0.5 \\ 0.6 & -0.5 \\ 1 & 0.8 \end{bmatrix} z^{-2}.$$

With $L = 12$, the blind deconvolution procedure converges again to a good equalizer, providing the complete system impulse response f given in figure 4.

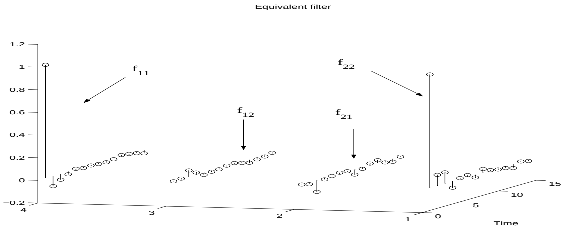

Figure 4: Equivalent impulse responses f for experiment 3.

ACKNOWLEGMENTS

The authors are grateful to F. Barthe for invaluable help.

References

[1] C. E. Shannon, “A mathematical theory of communication,” *Bell System Technical Journal*, vol. 27, pp. 379-423 and 623-656, July and October, 1948. Available at `http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html`.

[2] J.-F. Bercher, C. Vignat, “Estimating the entropy of a signal with applications”, *IEEE Trans. on Signal Processing*, vol. 48, no 6, june 2000.

[3] R. Zamir, “A Generalization of the Entropy Power Inequality with Applications”, *IEEE trans. on Information Theory*, vol. 39, no 5, sept. 1993.

[4] D. Donoho, “On minimum entropy deconvolution”, *Applied Time Series Analysis II*, pp. 565-609, Academic Press, 1981.

[5] A. Dembo and T. M. Cover, “Information theoretic inequalities”, *IEEE trans. on Information Theory*, vol. 37, no 6, pp. 1501-1518, nov. 1991.

[6] F. Barthe, “Optimal Young’s inequality and its converse: a simple proof.” *Geom. Funct. Anal.*, vol. 8, no. 2, pp. 234–242, 1998.

[7] F. Barthe, *Personnal communication*, Laboratoire de Mathématiques, Université de Marne-la-Vallée, dec. 2001.

[8] P. J. Huber, “Projection Pursuit”, *Annals of Statistics*, vol. 13, issue 2, pp. 435-475, june 1985.

[9] J. C. Principe, ‘Information-Theoretic Learning”, in *Unsupervised adaptive Filtering*, edited by S. Haykin, John Wiley and Sons, New York, 2000.

[10] J. W. Brewer, “Kronecker products and Matrix Calculus in System Theory”, *IEEE trans. on Circuits and Systems, vol. CAS-25, no. 9, sept. 1978.*