[10] ——, "The zero error capacity of a noisy channel," *IRE Trans. Inform. Theory*, vol. IT-12, pp. S8–S19, Sept. 1956. Reprinted in D. Slepian, Ed., *Key Papers in the Development of Information Theory*. New York: IEEE Press, 1974.

[11] $\phi$. Ytrehus, "Upper bounds on error-correcting runlength-limited block codes," *IEEE Trans. Inform. Theory*, vol. 37, pp. 941–945, May 1991.

## On the Maximum Entropy of the Sum of Two Dependent Random Variables

Thomas M. Cover, *Fellow, IEEE,* and Zhen Zhang, *Senior Member, IEEE*

***Abstract*—We investigate the maximization of the differential entropy $h(X+Y)$ of arbitrary dependent random variables $X$ and $Y$ under the constraints of fixed equal marginal densities for $X$ and $Y$. We show that $\max h(X+Y) = h(2X)$, under the constraints that $X$ and $Y$ have the same fixed marginal density $f$, if and only if $f$ is log-concave. The maximum is achieved when $X \equiv Y$. If f is not log-concave, the maximum is strictly greater than $h(2X)$. As an example, identically distributed Gaussian random variables have log-concave densities and satisfy $\max h(X+Y) = h(2X)$ with $X \equiv Y$. More general inequalities in this direction should lead to capacity bounds for additive noise channels with feedback.**

***Index Terms*—Differential entropy, maximum entropy, entropy power inequality.**

### I. Introduction

Let $X$ and $Y$ be two random variables. Their joint distribution is subject to the constraint that the two marginal densities are equal to a given density $f$. The differential entropy of a random variable $Z$ with density $g$ is given by $h(Z) = h(g) = -\int g \log g$. We are interested in the maximal entropy of the sum of these two random variables over all joint distributions:

$$\max_{X\sim f, Y\sim f} h(X+Y). \tag{1}$$

If $f$ is Gaussian, then we observe

$$\max_{X\sim f, Y\sim f} h(X+Y) = h(2X), \tag{2}$$

with equality if $X \equiv Y$. In this note, we study the conditions under which the inequality

$$\max_{X\sim f, Y\sim f} h(X+Y) \le h(2X) \tag{3}$$

holds.

We obtain a necessary and sufficient condition (concavity of the logarithm of $f$) for this inequality. These results are similar to the well-known entropy power inequality [1], [2].

$$h(X+Y) \le h(X^*+Y^*)$$

where $X$ and $Y$ are two independent random variables and $X^*$ and $Y^*$ are two independent Gaussian random variables having the same respective entropies. These inequalities were motivated by work on additive feedback channels where the signal can become correlated with the additive noise through feedback.

A function $f$ is called log-concave if, for every $\delta$,

$$\tfrac{1}{2}\log f(x-\delta) + \tfrac{1}{2}\log f(x+\delta) \le \log f(x). \tag{4}$$

If a density is log-concave (a.e.), we can always assume that it is log-concave because densities are defined up to a set of measure zero.

### II. Main Results

*Theorem 1:* The equality

$$\max_{X\sim f, Y\sim f} h(X+Y) = h(2X) \tag{5}$$

holds if and only if $f$ is log-concave. The maximizing joint distribution corresponds to $X \equiv Y$.

*Remark:* Since, by setting $X \equiv Y$, we always have $\max_{X\sim f, Y\sim f} h(X+Y) \ge h(2X)$, consequently, if $f$ is not log-concave, we have

$$\max_{X\sim f, Y\sim f} h(X+Y) > h(2X).$$

*Proof: Sufficiency.*

First, we note that $g = f$ is the only density which achieves the maximal entropy $h(g)$ over all densities $g$ satisfying the constraints

$$\int g = 1 \tag{6}$$

and

$$\int g \log f \ge \int f \log f. \tag{7}$$

This is argued as follows. From the constraint (7) and the information inequality

$$\int g \log\left(\frac{g}{f}\right) \ge 0, \tag{8}$$

we have

$$h(g) \le -\int g \log f \le h(f). \tag{9}$$

Therefore, the maximum of $h(g)$ is at most $h(f)$. To achieve the upper bound $h(f)$, we must have equalities in both (8) and (9). The only density $g$ for which (8) holds with equality is $g = f$.

The density of $Z = 2X$ is $\tilde{f}(z) = \frac{1}{2}f(z/2)$, and is therefore the density which achieves the maximum entropy $h(Z)$ subject to the constraints

$$\int g = 1 \tag{10}$$

and

$$\int g(z)\log \tilde{f}(z)\,dz \ge -h(\tilde{f}) = -h(f) - 1, \tag{11}$$

which is the same as

$$\int g(z)\log f\left(\frac{z}{2}\right)dz \ge -h(f). \tag{12}$$

We consider densities $g$ of $Z = X + Y$ for jointly distributed random variables $(X, Y)$ subject to the constraint $X \sim f$, $Y \sim f$.

Manuscript received November 20, 1992; revised October 13, 1993. This work was supported in part by the National Science Foundation under Grants NCR-9205663 and NCR-9205265, by JSEP Contract DAAL03-91-C-0010, and by DARPA Contract J-FBI-91-218.

T. M. Cover is with the Department of Electrical Engineering and Statistics, Stanford University, Stanford, CA 94305.

Z. Zhang is with the Communication Sciences Institute, Department of Electrical Engineering—Systems, University of Southern California, Los Angeles, CA 90089.

IEEE Log Number 9403838.

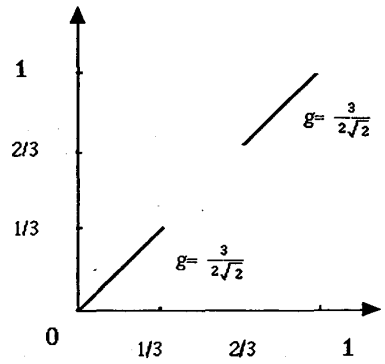


Fig. 1.

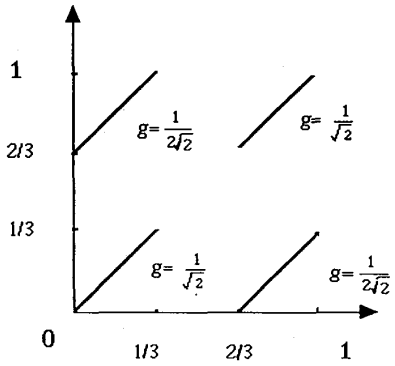


Fig. 2.

To prove sufficiency, we need to prove the validity of (12) for all such $g$. Since $f$ is log-concave, it follows by Jensen's inequality that

$$E_g \log f\left(\frac{Z}{2}\right) = E \log f\left(\frac{X+Y}{2}\right) \geq \frac{1}{2}(E_f \log f(X) + E_f \log f(Y)) = -h(X) = -h(f) \quad (13)$$

where the second expected value is taken with respect to the joint distribution of $X$ and $Y$. This proves sufficiency.

*Necessity.*

Consider an example with $f(x) = \frac{3}{2}$ in the intervals $(0, \frac{1}{3})$ and $(\frac{2}{3}, 1)$. Here, $f$ is not log-concave. The joint distribution $F$ for the random variables $X$ and $Y$ corresponding to $X \equiv Y$ concentrates on the line $x = y$, but has no joint density. The Radon–Nikodym derivative of $F(x, y)$ with respect to Lebesgue measure on the line $x = y$ exists and is shown in Fig. 1. In Fig. 2, we define another joint distribution $G$ of two random variables, say $\hat{X}$ and $\hat{Y}$, by its Radon–Nikodym derivative with respect to one-dimensional Lebesgue measure. $G$ is obtained by moving a part of the value of the Radon–Nikodym derivative of $F$ on the line $x = y$ to the line $x = y + \frac{2}{3}$ and $x = y - \frac{2}{3}$. It is easy to verify that the two marginal densities of the distribution remain the same. But $\hat{X} + \hat{Y}$ is uniformly distributed on the interval $(0, 2)$. Therefore, $h(\hat{X} + \hat{Y}) = 1$. But $h(2X) = \frac{4}{3} \log \frac{4}{3} < 1$. Consequently, the joint distribution corresponding to $X \equiv Y$ does not achieve the maximum entropy. The lack of log-concavity of $f$ enabled us to construct $G$ as above.

We return to the general case. We shall use the same technique. We first find a portion of the density violating log-concavity. If $\log f$ is not concave, then there exists a $\delta > 0$ such that

$$m(\log f(z - \delta) + \log f(z + \delta) > 2 \log f(z)) > 0 \quad (14)$$

where $m$ is Lebesgue measure over $R$. Denote the set satisfying (14) by $\Sigma$. Let $0 < \epsilon < \frac{1}{2}\delta$ and consider all intervals $\Delta(x, \epsilon) = (x, x + \epsilon)$. There exists an $x_0$ such that

$$m(\Sigma \cap \Delta(x_0, \epsilon)) > 0. \quad (15)$$

For all $\delta_0 > 0$, consider

$$\Delta(\delta_0) = \{z \in \Sigma \cap \Delta(x_0, \epsilon) : (f(z - \delta) - \delta_0)(f(z + \delta) - \delta_0) > (f(z) + 2\delta_0)^2\}. \quad (16)$$

There exists a $\delta_0 > 0$ such that

$$m(\Delta(\delta_0)) > 0. \quad (17)$$

Note that for $z \in \Delta(\delta_0)$, we have

$$\begin{aligned} &-(f(z) + 2\delta_0)\log(f(z) + 2\delta_0) - (f(z + \delta) - \delta_0) \\ &\quad \times \log(f(z + \delta) - \delta_0) - (f(z - \delta) - \delta_0) \\ &\quad \times \log(f(z - \delta) - \delta_0) \\ &= -2\delta_0 \log(f(z) + 2\delta_0) + \delta_0 \log(f(z + \delta) \\ &\quad - \delta_0) + \delta_0 \log(f(z - \delta) - \delta_0) \\ &\quad - f(z)\log f(z) - f(z + \delta)\log f(z + \delta) \\ &\quad - f(z - \delta)\log f(z - \delta) \\ &> -f(z)\log f(z) - f(z + \delta)\log f(z + \delta) \\ &\quad - f(z - \delta)\log f(z - \delta). \end{aligned} \quad (18)$$

Let $g$ be the Radon–Nikodym derivative (according to one-dimensional Lebesgue measure) of the joint distribution $G$ of $X$ and $Y$ where $X \equiv Y$. Of course, for this distribution, we have

$$h(X + Y) = h(2X). \quad (19)$$

We now translate part of the derivative to create a new joint distribution with higher entropy $h(\hat{X} + \hat{Y})$. Consider

$$\hat{g} = g + \frac{1}{\sqrt{2}}\delta_0(1_{\{(z-\delta, z+\delta): z \in \Delta(\delta_0)\}} + 1_{\{(z+\delta, z-\delta): z \in \Delta(\delta_0)\}} - 1_{\{(z+\delta, z+\delta): z \in \Delta(\delta_0)\}} - 1_{\{(z-\delta, z-\delta): z \in \Delta(\delta_0)\}}) \quad (20)$$

where $1_A$ is the indicator function of the set $A$. Let $\hat{G}$ be the joint distribution with derivative $\hat{g}$. We check that

1) $\hat{G}$ is a distribution function which satisfies the same marginal conditions as $G$, and

2) $h(\hat{X} + \hat{Y}) > h(2X)$, where $\hat{X}$ and $\hat{Y}$ are two random variables with the joint distribution $\hat{G}$.

Proving 1) is easy. We need only check that the marginals of $G' - G$ are zero and $g' \geq 0$. This is obvious from the definition.

To prove 2), let $Z = X + Y$. If the joint distribution of $(X, Y)$ is $G$, then the density of $Z$ is $\frac{1}{2}f(z/2)$. If the joint distribution of $(\hat{X}, \hat{Y})$ is $\hat{G}$, then the density of $\hat{Z} = \hat{X} + \hat{Y}$ is

$$\hat{f}(z) = \frac{1}{2}f\left(\frac{z}{2}\right) + \frac{\delta_0}{2}(2 \times 1_{\{z: z \in 2\Delta(\delta_0)\}} - 1_{\{z: z - 2\delta \in 2\Delta(\delta_0)\}} - 1_{\{z: z + 2\delta \in 2\Delta(\delta_0)\}}). \quad (21)$$

The three sets $2\Delta(\delta_0)$, $2\Delta(\delta_0) + 2\delta$, and $2\Delta(\delta_0) - 2\delta$ are disjoint. Denote them by $\Delta_1$, $\Delta_2$, and $\Delta_3$. We have

$$h(\hat{X} + \hat{Y}) = \int -\hat{f} \log \hat{f} dz$$

$$= \int_{\Delta_1 \cup \Delta_2 \cup \Delta_3} -\hat{f} \log \hat{f} dz + \int_{\Delta_1^c \cap \Delta_2^c \cap \Delta_3^c} -\hat{f} \log \hat{f} dz. \tag{22}$$

Divide the integral

$$\int -\frac{1}{2} f\left(\frac{z}{2}\right) \log \left[\frac{1}{2} f\left(\frac{z}{2}\right)\right] dz$$

$$= \int_{\Delta_1 \cup \Delta_2 \cup \Delta_3} -\frac{1}{2} f\left(\frac{z}{2}\right) \log \left[\frac{1}{2} f\left(\frac{z}{2}\right)\right] dz + \int_{\Delta_1^c \cap \Delta_2^c \cap \Delta_3^c} -\frac{1}{2} f\left(\frac{z}{2}\right) \log \left[\frac{1}{2} f\left(\frac{z}{2}\right)\right] dz \tag{23}$$

into the corresponding two parts. The second integral is the same for (22) and (23). We investigate the first integral. Using (18), we derive

$$\int_{\Delta_1 \cup \Delta_2 \cup \Delta_3} -\hat{f} \log \hat{f} dz$$

$$= \int_{\Delta_1} -\left(\frac{1}{2} f\left(\frac{z}{2}\right) + \delta_0\right) \log \left(\frac{1}{2} f\left(\frac{z}{2}\right) + \delta_0\right)$$

$$-\left(\frac{1}{2} f\left(\frac{z}{2} + {}_0\right) - \frac{1}{2}\delta_0\right) \log \left(\frac{1}{2} f\left(\frac{z}{2} + \delta\right) - \frac{1}{2}\delta_0\right)$$

$$-\left(\frac{1}{2} f\left(\frac{z}{2} - \delta\right) - \frac{1}{2}\delta_0\right) \log \left(\frac{1}{2} f\left(\frac{z}{2} - \delta\right) - \frac{1}{2}\delta_0\right) dz$$

$$> \int_{\Delta_1} -\frac{1}{2} f\left(\frac{z}{2}\right) \log \left(\frac{1}{2} f\left(\frac{z}{2}\right)\right) - \frac{1}{2} f\left(\frac{z}{2} + \delta\right)$$

$$\times \log \left(\frac{1}{2} f\left(\frac{z}{2} + \delta\right)\right) - \frac{1}{2} f\left(\frac{z}{2} - \delta\right)$$

$$\times \log \left(\frac{1}{2} f\left(\frac{z}{2} - \delta\right)\right) dz$$

$$= \int_{\Delta_1 \cup \Delta_2 \cup \Delta_3} -\frac{1}{2} f\left(\frac{z}{2}\right) \log \left(\frac{1}{2} f\left(\frac{z}{2}\right)\right) dz. \tag{24}$$

This proves 2), and hence completes the proof of necessity.


## Acknowledgments

The authors are grateful to both referees for their careful reading of the manuscript and helpful comments.


## References


[1] A. J. Stam, "Some inequalities satisfied by the quantities of information of Fisher and Shannon," *Inform. Contr.*, vol. 2, pp. 101–112, June 1959.

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.


# The Length of a Typical Huffman Codeword


Rüdiger Schack



***Abstract*—If $p_i (i = 1, \cdots, N)$ is the probability of the $i$th letter of a memoryless source, the length $l_i$ of the corresponding binary Huffman codeword can be very different from the value $-\log p_i$. For a *typical* letter, however, $l_i \approx -\log p_i$. More precisely, $P_m^- = \sum_{j \in \{i | l_i < -\log p_i - m\}} p_j < 2^{-m}$ and $P_m^+ = \sum_{j \in \{i | l_i > -\log p_i + m\}} p_j < 2^{-m/c(m-2)+2}$, where $c \approx 2.27$.**



***Index Terms*—Huffman code, length of a typical codeword.**


## I. Results

Consider a discrete memoryless $N$-letter source ($N \geq 2$) to which a binary Huffman code [1] is assigned. The $i$th letter has probability $p_i < 1$ and codeword length $l_i$. For a dyadic source (i.e., all $p_i$ are negative powers of 2) the Huffman codeword lengths $l_i$ are equal to the self information, $-\log p_i$, for all $i$. More general sources, however, may give rise to *atypical* letters for which the Huffman codeword lengths differ greatly from the self information. Given $p_i$, the length $l_i$ can in principle be as small as 1 and as large as $\{\log[(\sqrt{5} + 1)/2]\}^{-1}(-\log p_i) \approx 1.44(-\log p_i)$ [2].

In this correspondence, bounds on the probability of such atypical letters are derived. It is shown that the probability of the letters for which the Huffman codeword length differs by more than $m$ bits from $-\log p_i$ decreases exponentially with $m$. In this sense, one can say that the Huffman codeword for a *typical* letter satisfies $l_i \approx -\log p_i$. This result has an application to recent fundamental questions in statistical physics [3], [4].

The Huffman code can be represented by a binary tree having the *sibling property* [5] defined as follows. The number of links leading from the root of the tree to a node is called the *level* of that node. If the level-$n$ node $a$ is connected to the level-$(n + 1)$ nodes $b$ and $c$, then $a$ is called the *parent* of $b$ and $c$; $a$'s *children* $b$ and $c$ are called *siblings*. There are exactly $N$ terminal nodes or *leaves*, each leaf corresponding to a letter. Each link connecting two nodes is labeled 0 or 1. The sequence of labels encountered on the path from the root to a leaf is the codeword assigned to the corresponding letter. The codeword length of a letter is thus equal to the level of the corresponding leaf. Each node is assigned a probability such that the probability of a leaf is equal to the probability of the corresponding letter and the probability of each nonterminal node is equal to the sum of the probabilities of its children. A tree has the sibling property if and only if each node except the root has a sibling and the nodes can be listed in order of nonincreasing probability with each node being adjacent to its sibling in the list [5].

*Definition:* A level-$l$ node with probability $p$—or, equivalently, a letter with probability $p$ and codeword length $l$—has the property $X_m^+ (X_m^-)$ if and only if $l > -\log p + m$ ($l < -\log p - m$).

*Theorem 1:* $P_m^- = \sum_{j \in I_m^-} p_j < 2^{-m}$ where $I_m^- = \{i | l_i < -\log p_i - m\}$, i.e., the probability that a letter has property $X_m^-$ is smaller than $2^{-m}$. (This is true for any prefix-free code.)


Manuscript received January 27, 1993; revised August 27, 1993. This work of the author was supported by a fellowship from the Deutsche Forschungsgemeinschaft.



The author is with the Department of Physics and Astronomy, University of New Mexico, Albuquerque, NM 87131.



IEEE Log Number 9403832.