Convergence of Differential Entropies

Mahesh Godavarti and Alfred O. Hero-III

November 29, 2001

Submitted to *IEEE Transactions on Information Theory*, November 2001

Abstract

Calculation of the differential entropy of the limiting density of a sequence of probability density functions is important in the field of entropy estimation. In such cases it would be of interest to know if the limit of the differential entropies corresponding to the sequence of probability density function is equal to the differential entropy of the limiting probability density function. In this paper, we establish sufficient conditions under which the above is true.

Keywords: differential entropy, convergence, probability density function.

Corresponding author:

Mahesh Godavarti
Altra Broadband, Inc.
Irvine Technology Center
16275 Laguna Canyon Rd
Irvine, CA 92618.
Tel: (949) 341-0106 x230.
Fax: (949) 341-0226.
e-mail: godavarti@altrabroadband.com.

[1]This work was completed while Mahesh Godavarti was a Ph.D. candidate at the University of Michigan, Ann Arbor under the supervision of Prof. Alfred O. Hero-III.

1 Introduction

The concept of convergence of differential entropy can be traced to the problem of entropy estimation [1, 2, 3, 8, 9, 10, 11, 12]. The problem of entropy estimation finds application in independent component analysis and projection pursuit. For more applications see [3].

In [2, 10] the authors use either the Gram Charlier expansion or expansions based on moments to approximate the density and hence the entropy. The drawback in these works is that the approximation to a density function $f(x)$ is of the form $(1-\epsilon)f(x)$ where $\epsilon \to 0$ as the estimate gets better. Most of the approximations [1, 3, 8, 11, 12] are not of this form.

In [1, 3, 8, 11, 12] the density is estimated from a finite number of realizations, $X_1, \ldots, X_n$ of the source X and the estimate is refined as $n \to \infty$. An estimate of the entropy, $\mathcal{H}_n$ is obtained from this estimate of the density f_n and is required to converge to $\mathcal{H}(f)$ as $f_n \to f$. The convergence of the estimate to the actual value happens provided the underlying probability density function $f(x)$, and the corresponding entropy $\mathcal{H}(f)$, satisfy some stringent conditions [3]. The conditions posed in [3] are as follows

1. f is continuous.
2. f is k times differentiable.
3. $\mathcal{H}(\lfloor X \rfloor) < \infty$ where $\lfloor X \rfloor$ is the integer part of X.
4. $\inf_{f(x)>0} f(x) > 0$.
5. $\int f(x)(\log f(x))^2 dx < \infty$.
6. f is bounded.

Some weaker conditions known for the convergence of differential entropies is that $f_n(x)$ be bounded from above and below for all n over the support of $f_n(x)$ [4]. This means that $|\log f_n(x)| \leq A$ for x in the support of $f_n(x)$ $\mathrm{S}f_n$, for all n. Let $\mathrm{S}f$ denote the support of $f(x)$ and let $\mathrm{S}f \backslash \mathrm{S}f_n$ denote the set of all x such that $x \in \mathrm{S}f$ and $x \notin \mathrm{S}f_n$. Then

$$
\begin{aligned}
|\lim_{n\to\infty} \int_{\mathrm{S}f_n} f_n(x)\log f_n(x)dx - \int_{\mathrm{S}f} f(x)\log f(x)dx| \quad &\leq \quad |\lim_{n\to\infty} \int_{\mathrm{S}f_n} f(x)\log\frac{f(x)}{f_n(x)}dx| + \\
&\quad\; |\lim_{n\to\infty} \int_{\mathrm{S}f\backslash\mathrm{S}f_n} f(x)\log f(x)dx| + \\
&\quad\; |\lim_{n\to\infty} \int_{\mathrm{S}f_n} (f(x)-f_n(x))\log f_n(x)dx| \\
&\leq \quad |\lim_{n\to\infty} \int_{\mathrm{S}f_n} f(x)\log\frac{f(x)}{f_n(x)}dx| + \\
&\quad\; |\lim_{n\to\infty} \int_{\mathrm{S}f\backslash\mathrm{S}f_n} f(x)\log f(x)dx| + \\
&\quad\; \lim_{n\to\infty} \int_{\mathrm{S}f_n} |f(x)-f_n(x)|Adx \\
&\to \quad 0.
\end{aligned}
$$

The concept of convergence of entropy can also be found in the area of asymptotic analysis of communication systems [5, 6, 7, 13]. The convergence results are for specific density functions and are not useful for the general problem where a sequence of random variables are converging to a final random variable and we are interested in the convergence of the corresponding differential entropies. In this paper, we tackle the more general problem and we show convergence under fairly weak conditions on the final density function. We start with the following examples:

Example 1 *Consider the sequence of probability density functions $f_n(x)$ defined over the real line as follows*

$$
f_n(x) = \begin{cases} 1-\frac{1}{n} & \text{when } x \in [0,1] \\ \frac{1}{nL^n} & \text{when } x \in (1, 1+L^n] \\ 0 & \text{elsewhere} \end{cases}
$$

where L is a positive number not equal to 1. Then $f_n(x)$ converges to $f(x)$ pointwise everywhere where $f(x)$ is the uniform distribution over the interval $[0,1]$. However, the differential entropy from $f_n(x)$, called $\mathcal{H}_n$, is given by

$$\mathcal{H}_n = -(1-\frac{1}{n})\log(1-\frac{1}{n}) - \frac{1}{nL^n}\left[\log\frac{1}{nL^n}\right]L^n = -(1-\frac{1}{n})\log(1-\frac{1}{n}) + \frac{1}{n}\log n + \log L$$

and therefore, $\lim_{n\to\infty}\mathcal{H}_n = \log L \neq 0 = \mathcal{H}_f$.

Example 2 *Consider the sequence of probability density functions $f_n(x)$ defined over the real line as follows*

$$f_n(x) = \begin{cases} 1-\frac{1}{n} & \text{when } x \in [0,1] \\ \frac{L^n}{n} & \text{when } x \in (1, 1+\frac{1}{L^n}] \\ 0 & \text{elsewhere} \end{cases}$$

where L is a positive number not equal to 1. Then $f_n(x)$ converges to $f(x)$ pointwise almost everywhere where $f(x)$ is the uniform distribution over the interval $[0,1]$. The differential entropy from $f_n(x)$, $\mathcal{H}_n$ is given by

$$\mathcal{H}_n = -(1-\frac{1}{n})\log(1-\frac{1}{n}) - \frac{L^n}{n}\log\frac{L^n}{n}\frac{1}{L^n} = -(1-\frac{1}{n})\log(1-\frac{1}{n}) + \frac{1}{n}\log n - \log L$$

and therefore, $\lim_{n\to\infty} = -\log L \neq 0 = \mathcal{H}_f$.

In both the examples given above we see that convergence of probability density functions doesn't lead to the convergence of the corresponding differential entropies. In Example 1, we see that the second moment $\int |x|^2 f_n(x)dx$ is unbounded whereas in Example 2, we see that the pdf $f_n(x)$ itself is unbounded. It is possible to ask the question if we ensure that the above two quantities are bounded then do we obtain convergence of the differential entropies? The answer is indeed yes and is proved in the following section.

2 Main Results

The main results in this section are Theorems 1, 2 and 3. Lemma 1 is only useful in establishing the proof of Theorem 1 and is not significant otherwise.

Let $\chi_P(x)$ denote the characteristic function over a set P defined as $\chi_P(x) = 0$ if $x \notin P$ and $\chi_P(x) = 1$ if $x \in P$.

Lemma 1 *Let $g : \mathcal{C}^P \to R$ be a positive bounded function whose region of support, $\mathrm{S}g$, is compact. If there exists a constant L such that $\int g(x)dx \leq L < 1/e$ then $|\int g(x)\log g(x)dx| \leq \max\{|L\log L| + |L\log \mathrm{vol}(\mathrm{S}g)|, |L\log A|\}$ where $A = \sup g(x)$.*

Proof: First, $\int g(x)\log g(x)dx \leq \int g(x)\log A dx \leq L\log A$. Let $\int g(x)dx = I_g$. Consider the probability density function $g(x)/I_g$. We know that $\int \frac{g(x)}{I_g}\log\frac{g(x)}{I_g f(x)}dx \geq 0$ for all probability density functions $f(x)$. If

$$f(x) = \frac{\chi_{\mathrm{S}g}}{\mathrm{vol}(\mathrm{S}g)}$$

then

$$\int g(x)\log g(x)dx \geq \int g(x)\log(I_g f(x)) = I_g \log\frac{I_g}{\mathrm{vol}(\mathrm{S}g)}.$$

This implies

$$\begin{aligned} |\int g(x)\log g(x)| &\leq \max\{|L\log A|, |I_g\log\frac{I_g}{\mathrm{vol}(\mathrm{S}g)}|\} \\ &\leq \max\{|L\log A|, |I_g\log I_g| + |I_g\log\mathrm{vol}(\mathrm{S}g)|\} \\ &\leq \max\{|L\log A|, |L\log L| + |L\log\mathrm{vol}(\mathrm{S}g)|\}. \end{aligned}$$

The last inequality follows from the fact that for $x < 1/e$, $|x\log x|$ is an increasing function of x. □

Theorem 1 *Let $\{X_i \in \mathcal{C}^P\}$ be a sequence of continuous random variables with probability density functions, $\{f_i\}$ and $X \in \mathcal{C}^P$ be a continuous random variable with probability density function f such that $f_i \to f$ pointwise. If 1) $\max\{f_i(x), f(x)\} \leq A < \infty$ for all i and 2) $\max\{\int \|x\|^\kappa f_i(x)dx, \int \|x\|^\kappa f(x)dx\} \leq L < \infty$ for some $\kappa > 1$ and all i then $\mathcal{H}(X_i) \to \mathcal{H}(X)$. $\|x\| = \sqrt{x^\dagger x}$ denotes the Euclidean norm of x.*

Proof: The proof is based on showing that given an $\epsilon > 0$ there exists an R such that for all i

$$|\int_{\|x\|>R} f_i(x)\log f_i(x)dx| < \epsilon.$$

This R also works for $f(x)$.

Since $y\log y \to 0$ as $y \to 0$ we have $\max_{f(x)\leq A}|f(x)\log f(x)| \leq \max\{A\log A, e\} \stackrel{\text{def}}{=} K$. Therefore, $f_i(x)\log f_i(x)$ is bounded above by an L^1 function ($g = K\chi_{\|x\|\leq R}$) and by the dominated convergence theorem we have

$$-\int_{\|x\|\leq R} f_i(x)\log f_i(x)dx \to -\int_{\|x\|\leq R} f(x)\log f(x)dx.$$

Now, to show that the integral outside of $\|x\| \leq R$ is uniformly bounded for all f_i and f. Let g denote either f_i or f. We have $\int \|x\|^\kappa g(x)dx \leq L$. Therefore, by Markov's inequality $\int_{R<\|x\|\leq R+1} g(x)dx = I^R \leq L/R^\kappa$. Choose R large enough so that for all $l > R$: $I^l < 1/e$. Now

$$|\int_{\|x\|>R} g(x)\log g(x)dx| \leq \int_{\|x\|>R} |g(x)\log g(x)|dx = \sum_{l=R}^{\infty}\int_{B_l}|g(x)\log g(x)|dx$$

where $B_l = \{x : l < \|x\| \leq l+1\}$.

Consider the term $\int_{B_l}|g(x)\log g(x)|dx = G_l$. Also, define $A_+ = \{x : -\log g(x) > 0\}$ and $A_- = \{x : -\log g(x) < 0\}$ Now,

$$\begin{aligned} G_l &= \int_{A_+\cap B_l}|g(x)\log g(x)|dx + \int_{A_-\cap B_l}|g(x)\log g(x)|dx \\ &= |\int_{A_+\cap B_l} g(x)\log g(x)dx| + |\int_{A_-\cap B_l} g(x)\log g(x)dx|. \end{aligned}$$

From Lemma 1, we have

$$G_l \leq 2\max\{|I^l\log I^l| + |I^l\log \text{vol}(\{B_l\})|, |I^l\log A|\}.$$

We know $\text{vol}(\{x : B_l\}) = o(l^{2P})$. Therefore,

$$\int_{B_l}|g(x)\log g(x)|dx \leq \frac{Q}{l^\kappa}\log l$$

where Q is some sufficiently large constant. Therefore, we have

$$\int_{\|x\|>R}|g(x)\log g(x)|dx \leq \sum_{l=R}^{\infty}\frac{Q}{l^\kappa}\log l = O(\log R/R^{\kappa-1}).$$

Finally, as $\kappa > 1$ we can choose R sufficiently large to have $|\int_{\|x\|>R} g(x)\log g(x)dx| < \epsilon$. □

Theorem 2 *Let $\{X_i \in \mathcal{C}^P\}$ be a sequence of continuous random variables with probability density functions, f_i and $X \in \mathcal{C}^P$ be a continuous random variable with probability density function f. Let $X_i \xrightarrow{P} X$. If 1) $\int \|x\|^\kappa f_n(x)dx \leq L$ and $\int \|x\|^\kappa f(x)dx \leq L$ for some $\kappa > 1$ and $L < \infty$ 2) $f(x)$ is bounded then $\limsup_{i\to\infty}\mathcal{H}(X_i) \leq \mathcal{H}(X)$.*

Proof: We will prove this by constructing a density function g_i corresponding to f_i such that $\mathcal{H}(X_i) \leq \mathcal{H}_{g_i}$ and $\limsup_{i\to\infty}\mathcal{H}_{g_i} \leq \mathcal{H}(X)$ thus concluding $\limsup \mathcal{H}(X_i) \leq \mathcal{H}(X)$ where $\mathcal{H}_{g_i} \stackrel{\text{def}}{=} -\int g_i(x)\log g_i(x)dx$.

First we will show that for all g_i defined above there exists a single real number $R > 0$ such that $-\int_{\|x\|>R} g_i(x)\log g_i(x)dx \leq \epsilon$. Note that this is different from the condition in Theorem 1 where we show $|\int_{\|x\|>R} g_i(x)\log g_i(x)dx| \leq \epsilon$. As in Theorem 1 choose R large enough so that $I^l < 1/e$. Also define the two sets A_+ and A_- as in Theorem 1 then

$$\begin{aligned} -\int_{\|x\|>R} g(x)\log g(x)dx &= -\int_{A_+} g(x)\log g(x)dx - \int_{A_-} g(x)\log g(x)dx \\ &= -\sum_{l=R}^{\infty}\int_{B_l\cap A_+} g(x)\log g(x)dx - \int_{A_-} g(x)\log g(x)dx \end{aligned}$$

where B_l is as defined in Theorem 1. The last line follows from the Monotone Convergence Theorem. From the proof of Lemma 1 we have $-\int_{B_l \cap A_+} g(x) \log g(x) dx \leq -I^l \log I^l + I^l \log \text{vol}(B_l)$ Therefore

$$\begin{aligned} -\int_{\|x\|>R} g(x) \log g(x) dx &\leq \sum_{l=R}^{\infty} [-I^l \log I^l + I^l \log \text{vol}(B_l)] - \int_{A_-} g(x) \log g(x) dx \\ &\leq \sum_{l=R}^{\infty} [-I^l \log I^l + I^l \log \text{vol}(B_l)] \end{aligned}$$

and the sum in the last line is bounded above by $\sum_{l=R}^{\infty} \frac{Q}{l^\kappa} \log l = O(\log R / R^{\kappa-1})$. Therefore,

$$\max_g \{ -\int_{\|x\|>R} g(x) \log g(x) dx \ \} \leq O(\log R / R^{\kappa-1}).$$

From the proof of Theorem 1, $|\int_{\|x\|>R} f(x) \log f(x) dx| = O(\log R / R^{\kappa-1})$.

Now let's concentrate on upperbounding $-\int_{\|x\|\leq R} f_i(x) \log f_i(x) dx$. Let $A = \sup f(x)$. For each n partition the region $\{\|x\| \leq R\}$ into n regions $P_m, m = 1, \ldots, n$ such that $A\frac{m-1}{n} \leq f(x) < A\frac{m}{n}$ for $x \in P_m$, $m < n$ and $A\frac{n-1}{n} \leq f(x) \leq A$ for $x \in P_n$. Now for each n, there exists a number M_n such that $\max_m |\int_{P_m} (f_i(x) - f(x)) dx| < \frac{1}{n} \min_m \int_{P_m} f(x) dx$ for all $i \geq M_n$. If $M_n \leq M_{n-1}$ set $M_n = M_{n-1} + 1$. Now, define the function $M(i)$ such that

$$M(i) = \begin{cases} 1, & 1 \leq i \leq M_2 \\ 2, & M_2 < i \leq M_3 \\ 3, & M_3 < i \leq M_4 \\ \vdots & \end{cases}$$

For each i, divide the region $\{\|x\| \leq R\}$ into $M(i)$ parts as defined in the previous paragraph: $P_n, n = 1, \ldots, M(i)$, and define $g_i(x)$ over $\{\|x\| \leq R\}$ as

$$g_i(x) = \sum_{n=1}^{M(i)} \chi_{P_n}(x) I_{n,i} / V_n$$

where $I_{n,i} = \int_{P_n} f_i(x) dx$, $V_n = \text{vol}(P_n)$.

Now, it is easy to see that $-\int_{\|x\|\leq R} f_i(x) \log f_i(x) \leq -\int_{\|x\|\leq R} g_i(x) \log g_i(x)$. Also, note that $g_i(x) \to f(x)$ pointwise. Since $f(x)$ is bounded there exists a number N and a constant K such that $g_i(x) \leq K$ for all values of $i > N$, also $f(x) \leq K$. Therefore, using Theorem 1 we conclude that $\lim -\int_{\|x\|\leq R} g_i(x) \log g_i(x) dx \to -\int_{\|x\|\leq R} f(x) \log f(x) dx$.

Therefore, $\limsup \mathcal{H}(X_i) \leq \limsup \mathcal{H}_{g_i} \leq \mathcal{H}(X)$. □

Theorem 3 *Let $\{X_i \in \mathcal{C}^P\}$ be a sequence of continuous random variables with probability density functions, f_i and $X \in \mathcal{C}^P$ be a continuous random variable with probability density function f. Let $X_i \xrightarrow{P} X$. If $0 \leq \max\{f(x), f_i(x)\} \leq A < \infty$ then $\liminf_{i\to\infty} \mathcal{H}(X_i) \geq \mathcal{H}(X)$.*

Proof: Proof is similar to the proof of Theorem 2. First, for every $\epsilon > 0$ there exists $R > 0$ such that

$$-\int_{|x|<R} g(x) \log g(x) dx \geq \mathcal{H}(X) - \epsilon$$

where $g(x)$ is defined as

$$g(x) = f(x)\ \chi_{|x|<R}(x) + A\ \chi_{R\leq|x|<R+\Delta R}(x)$$

where ΔR is such that $\int_{|x|\geq R} f(x) dx = A\text{vol}(\{x : R \leq |x| < R + \Delta R\})$. Similarly, define $g_i(x)$ as

$$g_i(x) = f_i(x)\ \chi_{|x|<R}(x) + A\ \chi_{R\leq|x|<R+\Delta_i R}(x)$$

where $\Delta_i R$ is such that $\int_{|x|\geq R} f_i(x) dx = A\text{vol}(\{x : R \leq |x| < R + \Delta_i R\})$. Then from Theorem 1 we have

$$\lim_{i\to\infty} -\int g_i(x) \log g_i(x) \quad = \quad -\int g(x) \log g(x) dx$$

Since $-\int_{|x|\geq R} f_i(x)\log f_i(x)dx \geq -\int_{|x|\geq R} g_i(x)\log g_i(x)dx$,

$$\begin{aligned}\liminf_{i\to\infty} -\int f_i(x)\log f_i(x)dx &\geq \liminf_{i\to\infty} -\int g_i(x)\log g_i(x)dx \\ &= -\int g(x)\log g(x)dx \\ &\geq \mathcal{H}(X)-\epsilon\end{aligned}$$

Since, ϵ is arbitrary we are done. □

3 Discussion and Conclusion

We derived general sufficient conditions for the convergence of differential entropies. The conditions on the density functions needed in this paper are weaker than the conditions assumed by the authors of [3] in the context of entropy estimation. The conditions in this paper are as follows:

1. $\sup_x \max\{\sup_n f_n(x), f(x)\} < \infty$
2. $\max\{\sup_n \int |x|^\kappa f_n(x)dx, \int |x|^\kappa f(x)dx\} < \infty$ for some $\kappa > 1$

Our results find application in the field of entropy estimation for the purposes of independent component analysis and projection pursuit; and calculation of capacity for communication systems in asymptotic regimes. Examples include the capacity calculation of multi-antenna systems for high SNR [7, 13]. The results in this paper were directly applied in [7]. In [13], because of the special nature of the density functions considered, the convergence of the differential entropies could be proved directly. However, the same analysis could have been performed with the results derived in this paper.

References

[1] I. A. Ahmad and P.-E. Lin, "A nonparametric estimation of the entropy for absolutely continuous distributions," *IEEE Trans. on Inform. Theory*, vol. 22, pp. 688–692, May 1976.

[2] S. Amari, A. Cichocki, and H. H. Yang, "A new learning algorithm for blind signal separation," *Advances in Neural Information Processing Systems*, vol. 8, pp. 757–763, MIT Press, Cambridge MA, 1996.

[3] J. Beirlant, E. J. Dudewicz, L. Györfi, and E. van der Meulen, "Nonparametric entropy estimation: an overview," *Intern. J. Math. Stat. Sci.*, vol. 6, no. 1, pp. 17–39, june 1997.

[4] I. Csiszár, Personal communication at ISIT Washington, DC, 2001.

[5] A. Gersho, "Asymptotically optimal block quantization," *IEEE Trans. on Inform. Theory*, vol. 25, no. 4, pp. 373–380, July 1979.

[6] H. Gish and J. N. Pierce, "Asymptotically efficient quantizing," *IEEE Trans. on Inform. Theory*, vol. 14, no. 5, pp. 676–683, September 1968.

[7] M. Godavarti and A. O. Hero-III, "Multiple antenna capacity in a deterministic Rician fading channel," *Submitted to IEEE Trans. on Info. Theory*, 2001.

[8] P. Hall and S. C. Morton, "On the estimation of entropy," *Ann. Inst. Statist. Math.*, vol. 45, pp. 69–88, 1993.

[9] A. O. Hero and B. Ma, "Convergence rates of minimal graphs with random vertices," *Submitted to IEEE Transactions on Information Theory*, Aug. 2001.

[10] A. Hyvărinen, "New approximations of differential entropy for independent component analysis and projection pursuit," *Report A47, Helsinki University of Technology*, Aug. 1997.

[11] H. Joe, "On the estimation of entropy and other functionals of a multivariate density," *Ann. Inst. Statist. Math.*, vol. 41, pp. 683–697, 1989.

[12] A. Mokkadem, "Estimation of the entropy and information of absolutely continuous random variables," *IEEE Trans. on Inform. Theory*, vol. 35, no. 1, pp. 193–196, Jan. 1989.

[13] L. Zheng and D. N. C. Tse, "Packing spheres in the Grassmann manifold: A geometric approach to the non-coherent multi-antenna channel," *To appear in IEEE Trans. on Information Theory.*