

Article

Relative Entropy Derivative Bounds

Pablo Zegers *, Alexis Fuentes and Carlos Alarcón

Universidad de los Andes, Facultad de Ingeniería y Ciencias Aplicadas, Monseñor Álvaro del Portillo 12455, Las Condes, Santiago, Chile; E-Mails: afuentes@miuandes.cl (A.F.); ceag.201175@gmail.com (C.A.)

* Author to whom correspondence should be addressed; E-Mail: pzegers@miuandes.cl; Tel.: +56-226-189-324.

Received: 24 May 2013; in revised form: 12 July 2013 / Accepted: 16 July 2013 /

Published: 23 July 2013

Abstract: We show that the derivative of the relative entropy with respect to its parameters is lower and upper bounded. We characterize the conditions under which this derivative can reach zero. We use these results to explain when the minimum relative entropy and the maximum log likelihood approaches can be valid. We show that these approaches naturally activate in the presence of large data sets and that they are inherent properties of any density estimation process involving large numbers of random variables.

Keywords: relative entropy; Kullback-Leibler divergence; Shannon differential entropy; asymptotic equipartition principle; typical set; Fisher information; maximum log likelihood

1. Introduction

Given a large ensemble of i.i.d. random variables, all of them generated according to some common density function, the asymptotic equipartition principle [1] guarantees that only a fraction of the possible ensembles, which is called the typical set, gathers almost all the probability ([2] [p. 226]). This fact opens interesting possibilities: What if the properties of the typical set impose conditions on any estimation process, such that the parameter search is focused on just a small subset of the available parameter set? We explore this using generalizations of the Shannon differential entropy [1] and the Fisher information [3–5] under typical set considerations. There are other works that take into account both concepts [6–8]. However, this work focuses on something new: defining algebraic bounds on the behavior of the derivative of the relative entropy with respect to their parameters. Furthermore, we

characterize the conditions under which seeking for the extreme of this expression is a valid approach. Most importantly, we prove that these conditions automatically activate if large data sets are used. We finish this work discussing the relation between these bounds and the density function estimation problem. Under the conditions that activate these bounds, we characterize when the minimum relative entropy and maximum log likelihood approaches render the same solution. Furthermore, we show that these equivalences become true by the sole fact of having a large number of random variables.

2. Known Density Functions Case

Let us assume for the ensuing discussion that there exists a known source density function, f_X , with support, $\mathcal{S}$. We also know a set of i.i.d. samples, $\{\mathbf{x}\}_n \equiv \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, that was generated by f_X . There is also another density function, $f_{Y|\theta}(\mathbf{y})$, with the same support, $\mathcal{S}$. This density function is indexed by the parameter vector, $\theta \in \Theta \subseteq \mathbb{R}^m$.

The first part of this work deals with this case, where all the density functions are known. The second part of this work studies the density function estimation problem, where the source density function, f_X , is not known.

3. Mixed Entropy Derivative: A Proxy for the Relative Entropy Derivative

We present the following definition:

Definition 1 The relative entropy ([2][p. 231]), also called Kullback-Leibler divergence, is defined by:

$$D_{RE}(f_X \parallel f_{Y|\theta}) \equiv \int_{\mathcal{S}} f_X(\mathbf{u}) \ln \frac{f_X(\mathbf{u})}{f_{Y|\theta}(\mathbf{u})} d\mathbf{u} \quad (1)$$

Notice that $\lim_{u \rightarrow 0^+} u \ln u = 0$. Even though the definition is valid for any pair of density functions, we are using those useful for the ensuing analysis. The domain of this integral, and that of all the integrals in this work, corresponds to the support, $\mathcal{S}$.

Definition 2 The mixed entropy, $h_M(f_X, f_{Y|\theta})$ of two density functions is defined by:

$$h_M(f_X, f_{Y|\theta}) \equiv - \int_{\mathcal{S}} f_X(\mathbf{u}) \ln f_{Y|\theta}(\mathbf{u}) d\mathbf{u} \quad (2)$$

when the integral exists.

Definition 3 The Shannon differential entropy [1] is defined by:

$$h_S(f_X) \equiv - \int_{\mathcal{S}} f_X(\mathbf{u}) \ln f_X(\mathbf{u}) d\mathbf{u} \quad (3)$$

From these definitions, when $f_{Y|\theta} = f_X$, from a functional point of view, the mixed entropy, $h_M(f_X, f_{Y|\theta})$, equals the differential entropy, $h_S(f_X)$

From examination of the relative entropy definition:

$$\begin{aligned} D_{KL}(f_X \parallel f_{Y|\theta}) &= \int_{\mathcal{S}} f_X(\mathbf{u}) \ln f_X(\mathbf{u}) d\mathbf{u} - \int_{\mathcal{S}} f_X(\mathbf{u}) \ln f_{Y|\theta}(\mathbf{u}) d\mathbf{u} \\ &= -h_S(f_X) + h_M(f_X, f_{Y|\theta}) \end{aligned} \quad (4)$$

Hence:

$$\frac{\partial D_{KL}(f_{\mathbf{x}} \parallel f_{\mathbf{Y}|\theta})}{\partial \theta_k} = \frac{h_M(f_{\mathbf{x}}, f_{\mathbf{Y}|\theta})}{\partial \theta_k} \quad (5)$$

In the previous equation it was assumed that the $\frac{\partial}{\partial \theta_k}$ are defined in the interior of Θ for $k \in \{1, \dots, m\}$. In this work, where we assume that $f_{\mathbf{x}}$ does not depend on any θ_k , we use extensively the fact that the relative entropy derivative is equal to the mixed entropy derivative. Hence, in the following, we focus on studying the properties of the mixed entropy in order to be able to say something about the properties of the relative entropy derivatives.

4. Mixed Entropy Typical Set

An interesting insight related to sequences of i.i.d. random variables, discovered by Shannon [1], is the usefulness of the weak law of large numbers to characterize large ensembles of random variables. In the context of this work, this law implies:

Defining

Definition 4 The likelihood, $f_{\{\mathbf{Y}\}_n|\theta}$, is defined by:

$$f_{\{\mathbf{Y}\}_n|\theta} \equiv \prod_{k=1}^n f_{\mathbf{Y}|\theta}(\mathbf{x}_k) \quad (6)$$

we can state:

Theorem 1 For any positive real number, ε , and fixed parameter vector, θ :

$$\mathbb{P} \left\{ \left| -\frac{1}{n} \ln f_{\{\mathbf{Y}\}_n|\theta} - h_M(f_{\mathbf{x}}, f_{\mathbf{Y}|\theta}) \right| \leq \varepsilon \right\} \xrightarrow{n \rightarrow \infty} 1 \quad (7)$$

Proof: Use that the random variables are i.i.d., and the weak law of large numbers. $\square$

Hence, it makes sense to define the following set ([2][p. 226]):

Definition 5 For any positive real number, ε , fixed parameter vector, θ , and any $n \geq 1$, the mixed entropy typical set, $\mathcal{M}_{\varepsilon}^n$, with respect to the density function, $f_{\mathbf{x}}$, is defined by:

$$\mathcal{M}_{\varepsilon}^n = \left\{ \{\mathbf{x}\}_n \in \mathcal{S}^n : \left| -\frac{1}{n} \ln f_{\{\mathbf{Y}\}_n|\theta} - h_M(f_{\mathbf{x}}, f_{\mathbf{Y}|\theta}) \right| \leq \varepsilon \right\} \quad (8)$$

Furthermore, using the weak law of large numbers, it is possible to prove:

Lemma 1 For any positive real number, ε , and any parameter vector, θ , it is true that:

$$\mathbb{P} \{ \{\mathbf{x}\}_n \in \mathcal{M}_{\varepsilon}^n \} \xrightarrow{n \rightarrow \infty} 1 \quad (9)$$

Proof: Use the weak law of large numbers in the mixed entropy typical set definition. $\square$

This proves that for large values of n , almost all the probability is contained by the mixed entropy typical set and that the probability of being outside this set becomes negligible.

Then, it is possible to prove that:

Lemma 2 For any positive real number, ε , and any fixed parameter vector, θ , it exists $n_0 \equiv n_0(\varepsilon, \theta)$, such that for all $n > n_0$, the total probability of the typical set is lower bounded by:

$$1 - \varepsilon \leq \mathbb{P} \{ \mathcal{M}_\varepsilon^n \} \quad (10)$$

Proof: Check ([2][p. 226], [9][p. 118]) for details. $\square$

5. Micro-Differences in Mixed Entropy Typical Sets

Assuming that $\{x\}_n \in \mathcal{M}_\varepsilon^n$ and from the definition of mixed entropy typical set:

$$\ln f_{\{Y\}_n|\theta} \in] -nh_M(f_X, f_{Y|\theta}) - n\varepsilon, -nh_M(f_X, f_{Y|\theta}) + n\varepsilon[\quad (11)$$

The width of this interval is $2\varepsilon n$. This allows us to compare $\ln f_{\{Y\}_n|\theta}$ with $-nh_M(f_X, f_{Y|\theta})$, the value that is obtained in the limit thanks to the weak law of large numbers. Their ratio is:

$$\left| \frac{2\varepsilon n}{-nh_M(f_X, f_{Y|\theta})} \right| = \left| \frac{2\varepsilon}{h_M(f_X, f_{Y|\theta})} \right| \quad (12)$$

In other words, given that ε can be chosen as small as needed, the range of values that $\ln f_{\{Y\}_n|\theta}$ can take may be indistinguishable from $-nh_M(f_X, f_{Y|\theta})$, with high probability. This is another way of stating the weak law of large numbers.

The previous facts allow us to present the following definition:

Definition 6 When $\{x\}_n \in \mathcal{M}_\varepsilon^n$, its associated micro-difference is defined by the following function:

$$\delta \equiv \delta(f_X, n, \{x\}_n, f_{Y|\theta}, \theta, \varepsilon) \in]0, 1[\quad (13)$$

such that:

$$\ln f_{\{Y\}_n|\theta} = -n(h_M(f_X, f_{Y|\theta}) + \varepsilon) + 2\varepsilon n\delta \quad (14)$$

hence:

$$\delta = \frac{1}{2\varepsilon n} (\ln f_{\{Y\}_n|\theta} + n(h_M(f_X, f_{Y|\theta}) + \varepsilon)) \quad (15)$$

The analysis performed in the preceding paragraphs shows that the behavior of the micro-differences might be negligible in the case of large values of n . However, the following sections of this work deal with the derivative of the micro-difference value with respect to the parameters; so, we cannot neglect it, and we need to study its behavior.

From Equation (14):

$$\frac{\partial \ln f_{\{Y\}_n|\theta}}{\partial \theta_k} = -n \frac{\partial h_M(f_X, f_{Y|\theta})}{\partial \theta_k} + 2\varepsilon n \frac{\partial \delta}{\partial \theta_k} \quad (16)$$

for all $k \in [1, \dots, m]$. Thus:

$$2\varepsilon \frac{\partial \delta}{\partial \theta_k} = \frac{1}{n} \frac{\partial \ln f_{\{Y\}_n|\theta}}{\partial \theta_k} + \frac{\partial h_M(f_X, f_{Y|\theta})}{\partial \theta_k} \quad (17)$$

$$= \frac{\partial}{\partial \theta_k} \left(\frac{1}{n} \ln f_{\{Y\}_n|\theta} + h_M(f_X, f_{Y|\theta}) \right) \quad (18)$$

$$= -\frac{\partial}{\partial \theta_k} \left(-\frac{1}{n} \ln f_{\{Y\}_n|\theta} - h_M(f_X, f_{Y|\theta}) \right) \quad (19)$$

This last expression helps us to understand that the derivative of δ with respect to the parameters is related to changes experienced by quantities upper bounded by ε .

6. Mixed Information

We define:

Definition 7 The mixed information, $i_F(f_{\{X\}_n}, f_{\{Y\}_n|\theta})_k$, is defined by:

$$i_F(f_{\{X\}_n}, f_{\{Y\}_n|\theta})_k \equiv \int_{S^n} f_{\{X\}_n} \left(\frac{\partial \ln f_{\{Y\}_n|\theta}}{\partial \theta_k} \right)^2 d\mathbf{u} \quad (20)$$

with:

$$f_{\{X\}_n} \equiv \prod_{k=1}^n f_X(\mathbf{x}_k) \quad (21)$$

when the integral exists.

7. Mixed Entropy Derivative Bounds

The main contribution of this work is the following pair of inequalities, which are obtained thanks to a combination of the mixed information and the mixed entropy.

Theorem 2 Given a positive real number, ε , any parameter vector, θ , and an i.i.d. sequence with n elements, then the components of $\nabla_{\theta} h_M(f_X, f_{Y|\theta})$ comply with:

$$\frac{\sqrt{I_R^k} - \frac{1}{n} \sqrt{i_F(f_{\{X\}_n}, f_{\{Y\}_n|\theta})_k}}{\sqrt{I_L}} \leq \left| \frac{\partial h_M(f_X, f_{Y|\theta})}{\partial \theta_k} \right| \leq \frac{\sqrt{I_R^k} + \frac{1}{n} \sqrt{i_F(f_{\{X\}_n}, f_{\{Y\}_n|\theta})_k}}{\sqrt{I_L}} \quad (22)$$

where:

$$I_R^k \equiv \int_{\mathcal{M}_{\varepsilon}^n} f_{\{X\}_n} \left| 2\varepsilon \frac{\partial \delta}{\partial \theta_k} \right|^2 d\mathbf{u} \quad (23)$$

$$I_L \equiv \int_{\mathcal{M}_{\varepsilon}^n} f_{\{X\}_n} d\mathbf{u} \quad (24)$$

Proof: From the mixed information definition:

$$i_F(f_{\{X\}_n}, f_{\{Y\}_n|\theta})_k = I_{\mathcal{M}_{\varepsilon}^n}^k + I_{\sim \mathcal{M}_{\varepsilon}^n}^k \geq I_{\mathcal{M}_{\varepsilon}^n}^k > 0 \quad (25)$$

where:

$$I_{\mathcal{M}_{\varepsilon}^n}^k \equiv \int_{\mathcal{M}_{\varepsilon}^n} f_{\{X\}_n} \left(\frac{\partial \ln f_{\{Y\}_n|\theta}}{\partial \theta_k} \right)^2 d\mathbf{u} \quad (26)$$

and the symbol, $\sim$, denotes the complement set.

Replacing Equation (16) in Equation (26), it is obtained:

$$\begin{aligned} I_{\mathcal{M}_{\varepsilon}^n}^k &= n^2 \left(\frac{\partial h_M(f_X, f_{Y|\theta})}{\partial \theta_k} \right)^2 \int_{\mathcal{M}_{\varepsilon}^n} f_{\{X\}_n} d\mathbf{u} - 2n \frac{\partial h_M(f_X, f_{Y|\theta})}{\partial \theta_k} \int_{\mathcal{M}_{\varepsilon}^n} f_{\{X\}_n} \left(2\varepsilon n \frac{\partial \delta}{\partial \theta_k} \right) d\mathbf{u} \\ &\quad + \int_{\mathcal{M}_{\varepsilon}^n} f_{\{X\}_n} \left(2\varepsilon n \frac{\partial \delta}{\partial \theta_k} \right)^2 d\mathbf{u} \end{aligned} \quad (27)$$

Using the Cauchy-Bunyakovsky-Schwarz inequality:

$$\begin{aligned} \left| \int_{\mathcal{M}_\varepsilon^n} f_{\{\mathbf{x}\}_n} \left(2\varepsilon n \frac{\partial \delta}{\partial \theta_k} \right) d\mathbf{u} \right|^2 &= \left| \int_{\mathcal{M}_\varepsilon^n} \sqrt{f_{\{\mathbf{x}\}_n}} \sqrt{f_{\{\mathbf{x}\}_n}} \left(2\varepsilon n \frac{\partial \delta}{\partial \theta_k} \right) d\mathbf{u} \right|^2 \\ &\leq \int_{\mathcal{M}_\varepsilon^n} f_{\{\mathbf{x}\}_n} d\mathbf{u} \cdot \int_{\mathcal{M}_\varepsilon^n} f_{\{\mathbf{x}\}_n} \left| 2\varepsilon n \frac{\partial \delta}{\partial \theta_k} \right|^2 d\mathbf{u} \\ &= I_L \cdot n^2 I_R^k \end{aligned} \quad (28)$$

Therefore:

$$\begin{aligned} I_{\mathcal{M}_\varepsilon^n}^k &\geq n^2 \left(\frac{\partial h_M(f_{\mathbf{x}}, f_{\mathbf{y}|\theta})}{\partial \theta_k} \right)^2 I_L - 2n^2 \left| \frac{\partial h_M(f_{\mathbf{x}}, f_{\mathbf{y}|\theta})}{\partial \theta_k} \right| \sqrt{I_L} \sqrt{I_R^k} + n^2 I_R^k \\ &= n^2 \left(\left| \frac{\partial h_M(f_{\mathbf{x}}, f_{\mathbf{y}|\theta})}{\partial \theta_k} \right| \sqrt{I_L} - \sqrt{I_R^k} \right)^2 \end{aligned} \quad (29)$$

which implies:

$$\frac{1}{n^2} i_F(f_{\{\mathbf{x}\}_n}, f_{\{\mathbf{y}\}_n|\theta})_k \geq \left(\left| \frac{\partial h_M(f_{\mathbf{x}}, f_{\mathbf{y}|\theta})}{\partial \theta_k} \right| \sqrt{I_L} - \sqrt{I_R^k} \right)^2 \quad (30)$$

Hence, using the negative solution of the square root:

$$\frac{1}{n} \sqrt{i_F(f_{\{\mathbf{x}\}_n}, f_{\{\mathbf{y}\}_n|\theta})_k} \geq - \left| \frac{\partial h_M(f_{\mathbf{x}}, f_{\mathbf{y}|\theta})}{\partial \theta_k} \right| \sqrt{I_L} + \sqrt{I_R^k} \quad (31)$$

Therefore:

$$\frac{\sqrt{I_R^k} - \frac{1}{n} \sqrt{i_F(f_{\{\mathbf{x}\}_n}, f_{\{\mathbf{y}\}_n|\theta})_k}}{\sqrt{I_L}} \leq \left| \frac{\partial h_M(f_{\mathbf{x}}, f_{\mathbf{y}|\theta})}{\partial \theta_k} \right| \quad (32)$$

Taking into account the positive solution of Equation (30):

$$\left| \frac{\partial h_M(f_{\mathbf{x}}, f_{\mathbf{y}|\theta})}{\partial \theta_k} \right| \leq \frac{\sqrt{I_R^k} + \frac{1}{n} \sqrt{i_F(f_{\{\mathbf{x}\}_n}, f_{\{\mathbf{y}\}_n|\theta})_k}}{\sqrt{I_L}} \quad (33)$$

□

The previous result allows us to state the following remarks:

Remark 1 This theorem states algebraic bounds on the derivative of the mixed entropy that are valid for any value of n .

Remark 2 The expression:

$$I_R^k \equiv \int_{\mathcal{M}_\varepsilon^n} f_{\{\mathbf{x}\}_n} \left| 2\varepsilon \frac{\partial \delta}{\partial \theta_k} \right|^2 d\mathbf{u} = \int_{\mathcal{M}_\varepsilon^n} f_{\{\mathbf{x}\}_n} \left| -\frac{\partial}{\partial \theta_k} \left(-\frac{1}{n} \ln f_{\{\mathbf{y}\}_n|\theta} - h_M(f_{\mathbf{x}}, f_{\mathbf{y}|\theta}) \right) \right|^2 d\mathbf{u} \quad (34)$$

refers to the accumulation of derivatives of very small differences. This differences become smaller and smaller as the weak law of large numbers is enacted.

8. Unknown Source Density Function, f_X

Now, we study what happens when the source density function, f_X , is not known.

It can be proven that $D_{KL}(f_X \parallel f_{Y|\theta}) \geq 0$ ([2][p. 232]). Thus, from the corresponding definitions given at the start of this work:

$$h_M(f_X, f_{Y|\theta}) \geq h_S(f_X) \quad (35)$$

Therefore, the density estimation problem can be framed as the following optimization program:

$$\theta_\circ \equiv \theta_\circ(f_X, f_{Y|\theta}) = \arg \min_{\theta} h_M(f_X, f_{Y|\theta}) \quad (36)$$

Remark 3 *The solutions of this optimization program pose the following options:*

- $h_M(f_X, f_{Y|\theta_\circ}) = h_S(f_X)$, hence $f_{Y|\theta_\circ} = f_X$.
- $h_M(f_X, f_{Y|\theta_\circ}) > h_S(f_X)$, thus $f_{Y|\theta_\circ} \neq f_X$. This happens when the estimator cannot implement the desired density function, which is beyond its reach. One way of checking whether the global minimum has been reached is to compare the mixed entropy value to that of the Shannon differential entropy. If they differ, the global minimum has not been reached yet.

According to this result, it is straightforward to search for parameters that make true the following expression:

$$\nabla_{\theta} h_M(f_X, f_{Y|\theta})|_{\theta=\theta_\circ} = 0 \quad (37)$$

However, this is out of the limits in our density estimation problem, given that we do not have access to f_X . Thank to the weak law of large numbers:

$$\lim_{n \rightarrow \infty} -\frac{1}{n} \ln f_{\{Y\}_n|\theta} = h_M(f_X, f_{Y|\theta}) \quad (38)$$

in probability. Thus, for sufficiently large values of n :

$$-\frac{1}{n} \ln f_{\{Y\}_n|\theta} \approx h_M(f_X, f_{Y|\theta}) \quad (39)$$

Hence, the optimization program posed in Equation (36) can be approximated by:

$$\theta_\circ \equiv \theta_\circ(n, \{x\}_n, f_{Y|\theta}) = \arg \min_{\theta} \left(-\frac{1}{n} \ln f_{\{Y\}_n|\theta} \right) \quad (40)$$

or:

$$\theta_\circ \equiv \theta_\circ(n, \{x\}_n, f_{Y|\theta}) = \arg \max_{\theta} \left(\frac{1}{n} \ln f_{\{Y\}_n|\theta} \right) \quad (41)$$

which is the *maximum log likelihood* framework ([3][p. 261], [4][260], [5][p. 65]).

Remark 4 *As in the mixed entropy case, the maximum log likelihood framework can exhibit the following cases:*

- The optimization program finds a parameter combination that makes the estimator equal to the unknown source density function, thus making true the following expression for a sufficiently large n :

$$-\frac{1}{n} \ln f_{\{\mathbf{Y}\}_n|\theta_0} \approx h_M(f_{\mathbf{X}}, f_{\mathbf{Y}|\theta_0}) \approx h_S(f_{\mathbf{X}}) \quad (42)$$

- The estimator does not find the desired target function. This happens when the family of functions that can be implemented by the estimator does not include the target function. It also happens when the maximum log likelihood program is riddled with multiple local minima, and the optimization process gets trapped in one of them. Only under very specific conditions, it is possible to obtain the global maximum of the maximum log likelihood problem [10]. In this case, we cannot use a comparison between the value of the log likelihood term and that of the Shannon differential entropy to determine the goodness of our solution, as we could do in the case of the mixed entropy, because calculating the latter is analogous to the very problem we are trying to solve.

Hence, it is natural to work with this approximated program and look for parameters that make the following expression true:

$$\nabla_{\theta} \left(\frac{1}{n} \ln f_{\{\mathbf{Y}\}_n|\theta} \right) \Big|_{\theta=\theta_0} = 0 \quad (43)$$

9. An Example: Mismatched Density Functions

9.1. Known Density Functions Case

Let us define a source density function:

$$f_X(x) = \lambda e^{-\lambda x} H(x) \quad (44)$$

with $x \in \mathbb{R}$ and $H(x)$, the Heaviside step function. This density function is completely defined by the parameter, λ . Its differential entropy is given by:

$$h_S(f_X) = 1 - \ln \lambda = 1 + \ln \frac{1}{\lambda} \quad (45)$$

We also have a data set, $\{\mathbf{x}\}_n$, generated by the source density function.

The density function that depends on parameters is:

$$f_{Y|\mu,\sigma} = \frac{1}{\sigma\sqrt{2\pi}} \exp \left(-\frac{(\mu - x)^2}{2\sigma^2} \right) \quad (46)$$

with $x \in \mathbb{R}$, μ its mean and σ its standard deviation. Its differential entropy is given by:

$$h_S(f_{Y|\mu,\sigma}) = \ln \left(\sigma\sqrt{2\pi e} \right) = 1 + \ln \sigma + \ln \sqrt{\frac{2\pi}{e}} \quad (47)$$

which does not depend on the mean, μ .

9.2. Compliance with Theorem 2

We first analyze the bounds associated with parameter μ . For these density functions:

$$\frac{1}{n} \ln f_{\{Y\}_n|\mu,\sigma} = \frac{1}{n} \sum_{k=1}^n \ln f_{Y|\mu,\sigma}(x_k) = \ln \left(\frac{1}{\sigma\sqrt{2\pi}} \right) - \frac{1}{2n\sigma^2} \sum_{k=1}^n (\mu - x_k)^2 \quad (48)$$

Hence:

$$\frac{1}{n} \frac{\partial \ln f_{\{Y\}_n|\mu,\sigma}}{\partial \mu} = -\frac{1}{\sigma^2} \left(\mu - \frac{1}{n} \sum_{k=1}^n x_k \right) \quad (49)$$

We calculate the mixed information for μ :

$$\begin{aligned} i_F(f_{\{X\}_n}, f_{\{Y\}_n|\mu,\sigma})_\mu &= \int_{\mathbb{R}_+^n} f_{\{X\}_n} \left(\frac{\partial \ln f_{\{Y\}_n|\mu,\sigma}}{\partial \mu} \right)^2 d\mathbf{u} \\ &= \int_{\mathbb{R}_+^n} f_{\{X\}_n} \left(\frac{n}{\sigma^2} \left(\mu - \frac{1}{n} \sum_{k=1}^n u_k \right) \right)^2 d\mathbf{u} \\ &= \frac{n^2}{\sigma^4} \int_{\mathbb{R}_+^n} f_{\{X\}_n} \left(\mu - \frac{1}{n} \sum_{k=1}^n u_k \right)^2 d\mathbf{u} \end{aligned} \quad (50)$$

Hence:

$$\frac{1}{n} \sqrt{i_F(f_{\{X\}_n}, f_{\{Y\}_n|\mu,\sigma})_\mu} = \frac{1}{\sigma^2} \sqrt{\int_{\mathbb{R}_+^n} f_{\{X\}_n} \left(\mu - \frac{1}{n} \sum_{k=1}^n u_k \right)^2 d\mathbf{u}} \quad (51)$$

Also:

$$\begin{aligned} h_M(f_X, f_{Y|\mu,\sigma}) &= - \int_{-\infty}^{\infty} f_X(u) \ln f_{Y|\mu,\sigma}(u) du \\ &= - \int_{-\infty}^{\infty} \lambda e^{-\lambda u} H(u) \ln f_{Y|\mu,\sigma}(u) du \\ &= - \int_0^{\infty} \lambda e^{-\lambda u} \left(\ln \left(\frac{1}{\sigma\sqrt{2\pi}} \right) - \frac{(\mu - u)^2}{2\sigma^2} \right) du \\ &= \ln(\sigma\sqrt{2\pi}) \int_0^{\infty} \lambda e^{-\lambda u} du + \frac{\lambda}{2\sigma^2} \int_0^{\infty} (\mu - u)^2 e^{-\lambda u} du \\ &= \ln(\sigma\sqrt{2\pi}) + \frac{1}{\sigma^2} \left(\frac{\mu^2}{2} - \frac{\mu}{\lambda} + \frac{1}{\lambda^2} \right) \end{aligned} \quad (52)$$

and:

$$\frac{\partial h_M(f_X, f_{Y|\mu,\sigma})}{\partial \mu} = \frac{1}{\sigma^2} \left(\mu - \frac{1}{\lambda} \right) \quad (53)$$

Finally, we calculate the micro-structure term using Equation (17):

$$2\varepsilon \frac{\partial \delta}{\partial \mu} = \frac{1}{\sigma^2} \left(\frac{1}{n} \sum_{k=1}^n x_k - \frac{1}{\lambda} \right) \quad (54)$$

This allows us to calculate:

$$\begin{aligned}
 I_R^\mu &= \int_{\mathcal{M}_\varepsilon^n} f_{\{\mathbf{x}\}_n} \left| 2\varepsilon \frac{\partial \delta}{\partial \mu} \right|^2 d\mathbf{u} \\
 &= \int_{\mathcal{M}_\varepsilon^n} f_{\{\mathbf{x}\}_n} \left| \frac{1}{\sigma^2} \left(\frac{1}{n} \sum_{k=1}^n u_k - \frac{1}{\lambda} \right) \right|^2 d\mathbf{u} \\
 &= \frac{1}{\sigma^4} \int_{\mathcal{M}_\varepsilon^n} f_{\{\mathbf{x}\}_n} \left(\frac{1}{n} \sum_{k=1}^n u_k - \frac{1}{\lambda} \right)^2 d\mathbf{u}
 \end{aligned} \tag{55}$$

Now, for σ , we have:

$$\frac{1}{n} \frac{\partial \ln f_{\{Y\}_n | \mu, \sigma}}{\partial \sigma} = \frac{1}{\sigma^3} \left(-\sigma^2 + \frac{1}{n} \sum_{k=1}^n (\mu - u_k)^2 \right) \tag{56}$$

The mixed information for σ is defined by:

$$\begin{aligned}
 i_F(f_{\{X\}_n}, f_{\{Y\}_n | \mu, \sigma})_\mu &= \int_{\mathbb{R}_+^n} f_{\{X\}_n} \left(\frac{\partial \ln f_{\{Y\}_n | \mu, \sigma}}{\partial \sigma} \right)^2 d\mathbf{u} \\
 &= \int_{\mathbb{R}_+^n} f_{\{X\}_n} \left(\frac{n}{\sigma^3} \left(-\sigma^2 + \frac{1}{n} \sum_{k=1}^n (\mu - u_k)^2 \right) \right)^2 d\mathbf{u} \\
 &= \frac{n^2}{\sigma^6} \int_{\mathbb{R}_+^n} f_{\{X\}_n} \left(-\sigma^2 + \frac{1}{n} \sum_{k=1}^n (\mu - u_k)^2 \right)^2 d\mathbf{u}
 \end{aligned} \tag{57}$$

Hence:

$$\frac{1}{n} \sqrt{i_F(f_{\{X\}_n}, f_{\{Y\}_n | \mu, \sigma})_\mu} = \frac{1}{\sigma^3} \sqrt{\int_{\mathbb{R}_+^n} f_{\{X\}_n} \left(-\sigma^2 + \frac{1}{n} \sum_{k=1}^n (\mu - u_k)^2 \right)^2 d\mathbf{u}} \tag{58}$$

Again, the derivative of the mixed entropy with respect to σ :

$$\frac{\partial h_M(f_X, f_{Y | \mu, \sigma})}{\partial \sigma} = \frac{1}{\sigma^3} \left(\left(\sigma^2 - \frac{1}{\lambda^2} \right) - \left(\mu - \frac{1}{\lambda} \right)^2 \right) \tag{59}$$

Hence, the associated micro-differences expression is:

$$2\varepsilon \frac{\partial \delta}{\partial \sigma} = \frac{1}{\sigma^3} \left(\left(\frac{1}{n} \sum_{k=1}^n (\mu - u_k)^2 - \frac{1}{\lambda^2} \right) - \left(\mu - \frac{1}{\lambda} \right)^2 \right) \tag{60}$$

which allows us to calculate:

$$\begin{aligned}
 I_R^\sigma &= \int_{\mathcal{M}_\varepsilon^n} f_{\{\mathbf{x}\}_n} \left| 2\varepsilon \frac{\partial \delta}{\partial \sigma} \right|^2 d\mathbf{u} \\
 &= \int_{\mathcal{M}_\varepsilon^n} f_{\{\mathbf{x}\}_n} \left| \frac{1}{\sigma^3} \left(\left(\frac{1}{n} \sum_{k=1}^n (\mu - u_k)^2 - \frac{1}{\lambda^2} \right) - \left(\mu - \frac{1}{\lambda} \right)^2 \right) \right|^2 d\mathbf{u} \\
 &= \frac{1}{\sigma^6} \int_{\mathcal{M}_\varepsilon^n} f_{\{\mathbf{x}\}_n} \left(\left(\frac{1}{n} \sum_{k=1}^n (\mu - u_k)^2 - \frac{1}{\lambda^2} \right) - \left(\mu - \frac{1}{\lambda} \right)^2 \right)^2 d\mathbf{u}
 \end{aligned} \tag{61}$$

9.3. Unknown Source Density Function, f_x

If f_x is unknown, then it is not possible to calculate the previous bounds. In this case, we resort to the optimization program specified in Equation (41). Hence, we use Equations (49) and (56), in order to obtain:

$$\mu_n = \frac{1}{n} \sum_{k=1}^n x_k \quad (62)$$

$$\sigma_n^2 = \frac{1}{n} \sum_{k=1}^n (\mu_n - x_k)^2 \quad (63)$$

Using this choice of values immediately makes both mixed information expressions described by Equations (51) and (58) equal to zero, without the need for large values of n .

Furthermore, if we use:

$$\mu_o = \lim_{n \rightarrow \infty} \mu_n = \frac{1}{\lambda} \quad (64)$$

$$\sigma_o = \lim_{n \rightarrow \infty} \sigma_n = \frac{1}{\lambda} \quad (65)$$

then the micro-differences integrals also become equal to zero.

This example shows several things:

- The use of the maximum log likelihood solutions and large values of n guarantees that the derivative of the mixed entropy equals zero; hence, the derivative of the relative entropy with respect to its parameters is zero, too.
- However, this is not enough to guarantee that the density functions will match each other (a Gaussian cannot estimate an exponential). The maximum log likelihood solution only guarantees that the estimator will do the best it can do.
- If not enough data is available, the derivative of the mixed entropy will be different from zero, because the micro-differences terms are not zero.

10. Discussion

If one wanted to estimate a density function out of a data set, one could minimize the relative entropy: once we reach its minimum, we can guarantee that the density function implemented by the estimator will be equal to that which originated the data. This is true only if the family of functions that the estimator can effectively implement includes that of the source density function. Moreover, realizing that it is not possible to measure the relative entropy, one would quickly settle to maximize the log likelihood and obtain the desired density function. Keeping this in mind, the results presented in this work seem just another way of saying the same. However, this is not so. The difference is subtle. Whereas there are many density function estimation principles and it could be discussed whether minimizing the relative entropy is the most convenient one, the bounds presented in this work show that minimizing the relative entropy automatically activates in the presence of large data sets and that one does not need to decide

whether that estimation framework is the most convenient or not. Something similar happens with the weak law of large numbers: there are many ways of determining the expected value of the density function that generated the data, but everybody knows that the weak law of large numbers guarantees that in the presence of large data sets, this value can be effectively approximated by the average of the values. Our bounds, also based on the weak law of large numbers, state something similar: in the presence of large data sets, the source density function is perceived as the solution of the maximum log likelihood solution.

11. Conclusions

Theorem 2, which is the main result of this work, consists in a set of bounds on the partial derivative of the mixed entropy, which is a proxy to the relative entropy, with respect to the parameters of the estimator. These bounds relate the mixed information value, the capacity of the estimator to produce micro-differences and the number of random variables present in the analysis in such a way that it is possible to determine that these bounds activate automatically when the data set is large. If the micro-differences integral is zero, the mixed information derivative is zero for large data sets, independently of any other considerations. In other words, minimizing the mixed entropy is the preferred density estimation framework when large data sets are available.

Given the impossibility of estimating the mixed entropy, it is shown that the correct numerical approximation to this problem is finding the solution to the maximum log likelihood problem. Again, when the estimator is associated to a micro-differences integral equal to zero, this framework becomes the optimal one in the presence of large data sets.

Acknowledgments

The authors thank Jaime Cisternas, Jorge Silva and Jose Principe for their helpful insights about this work.

Conflict of Interest

The authors declare no conflict of interest.

References

1. Shannon, C. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423.
2. Cover, T.; Thomas, J. *Elements of Information Theory*; John Wiley and Sons, Inc.: Hoboken, NJ, USA, 1991.
3. Hogg, R.V.; Craig, A.T. *Introduction to Mathematical Statistics*; Prentice Hall: Upper Saddle River, NJ, USA, 1995.
4. Papoulis, A. *Probability, Random Variables, and Stochastic Processes*; McGraw-Hill: New York, NY, USA, 1991.
5. Van Trees, H.L. *Detection, Estimation, and Modulation Theory: Part I*; John Wiley and Sons, Inc.: Hoboken, NJ, USA, 2001.

6. Vignat, C.; Bercher, J. Analysis of signals in the Fisher-Shannon information plane. *Phys. Lett. A* **2003**, *312*.
7. Romera, E.; Dehesa, J. The Fisher-Shannon information plane, an electron correlation tool. *J. Chem. Phys.* **2004**, *120*, 8906–8912.
8. Dimitrov, V. On Shannon-Jaynes Entropy and Fisher Information. In Proceedings of the 27th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Saratoga Springs, New York, NY, USA, 8–13 July 2007.
9. Taubman, D.; Marcellin, M. *JPEG2000: Image Compression Fundamentals, Standards, and Practice*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 2002.
10. Boyd, S.; Vandenberghe, L. *Convex Optimization*; Cambridge University Press: Cambridge, UK, 2004.

© 2013 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).