On Convergence Properties of Shannon Entropy

Francisco J. Piera* and Patricio Parada†
Department of Electrical Engineering, University of Chile, Av. Tupper 2007, Santiago, 8370451, Chile.
(Dated: December 30, 2008)

Convergence properties of Shannon Entropy are studied. In the differential setting, it is known that weak convergence of probability measures (convergence in distribution) is not enough for convergence of the associated differential entropies. In that direction, an interesting example is introduced and discussed in light of new general results here provided for the desired differential entropy convergence, results that take into account both compactly and uncompactly supported densities. Convergence of differential entropy is also characterized in terms of the Kullback-Liebler discriminant for densities with fairly general supports, and it is shown that convergence in variation of probability measures guarantees such convergence under an appropriate boundedness condition on the densities involved. Results for the discrete setting are also provided, allowing for infinitely supported probability measures, by taking advantage of the equivalence between weak convergence and convergence in variation in that setting.

Keywords: Shannon entropy, continuous/discrete alphabet sources, compactly/uncompactly supported densities, weak convergence, convergence in variation, Kullback-Liebler discriminant

I. INTRODUCTION

The convergence of a sequence of probability measure entropies plays a key role in information theory, from both theoretical and applied points of view, mostly in the context of estimation of the entropy of an information source [1, 9, 15, 17].

The problem has been partially studied in the context of discrete sources, and proof of that is that some convergence results can be found in today's standard textbooks of information theory [5, 6], and some recent works [11]. A more general approach is found in the works of A. Barron, where a proof of the Central Limit Theorem based on entropy convergence [3] and the entropy convergence of stationary processes [2] are presented. The discussion of information topologies for general sources [10] touches tangentially the problem of convergence in a more general setting.

However, the focus of many of these works has been on continuity rather than convergence properties of Shannon entropy. On the one hand, continuity properties embrace results that guarantee convergence of entropy for all approximating sequences of probability measures converging to a given limiting probability measure. Emphasis is put there in identifying the largest class of probability measures for which the corresponding convergence of entropy takes place. On the other hand, convergence properties are usually related to deciding whether convergence of entropy takes place for a given, fixed family of probability measures, also converging in a certain topology to a limiting probability measure. Whereas in the continuity context all requirements are imposed on the limiting probability measure, in order to ensure convergence of entropy for all possible approximating sequences, in the convergence context one can and should exploit any underlying structure of the particular approximating sequence at hand, as usually done in applied probability problems.

The purpose of this paper is to present general conditions for the convergence of entropy sequences associated to both discrete and continuous sources, over possibly infinite or non-compactly supported alphabets, respectively.

For the case of continuous sources, a result of this nature can be used in applications where one is confronted with the problem of deciding whether the sequence of differential entropies associated with a family of probability densities $\{p_n\}_{n=1}^{\infty}$ on $\mathbb{R}^k$, each term of the sequence being given by

$$-\int_{\mathbb{R}^k} p_n \log [p_n]\, dx \tag{1}$$

with dx denoting Lebesgue measure, converges as n increases to infinity to the respective differential entropy associated to the limiting density of the family (assuming such limiting density exists in some appropriate sense).

---

*Research supported in part by the Millennium Science Nucleus on Information and Randomness, Dept. of Mathematical Engineering, U. of Chile, Chile, Program P04-069-F.; Electronic address: fpiera@ing.uchile.cl
†Electronic address: pparada@ing.uchile.cl

In general, only a numerical computation of the sequence elements (1) is possible, making it difficult to conclude the desired convergence in an abstract sense. Such convergence must be established then by exploiting underlying properties or structures of the sequence $\{p_n\}_{n=1}^{\infty}$ by itself and its limit.

If we assume pointwise convergence of the corresponding integrands, two main convergence-related results from real analysis are at our disposal: the monotone and dominated convergence theorems for Lebesgue integrals. The monotone convergence theorem provides no help for this problem given that if each p_n is a probability density function and, as such, satisfies the normalization condition

$$\int_{\mathbb{R}^k} p_n dx = 1,$$

then the monotonicity in the sequence $\{p_n\}_{n=1}^{\infty}$ is only possible in the trivial case when all densities coincide for almost every x. If we decide to use the dominated convergence theorem, we need to construct a function f such that

$$|p_n(x) \log [p_n(x)]| \leq f(x) \text{ for each } n \text{ and } x \text{ and with } \int_{\mathbb{R}^k} f dx < \infty. \tag{2}$$

This construction, however, is difficult to carry out in general.

A weaker but easier condition to check corresponds to verify if the boundedness condition

$$\sup_{n,x} |p_n(x)| < \infty$$

holds. This implies

$$M \doteq \sup_{n,x} |p_n(x) \log[p_n(x)]| < \infty,$$

but such condition is not enough for the application of the dominated convergence theorem in the case of densities supported over an infinite Lebesgue measure set[1].

We show that appropriate absolute continuity properties of measures provide a suitable boundedness condition that can be used, in conjunction with the dominated convergence theorem, to establish the desired convergence of the associated differential entropies, and of the Kullback-Liebler discriminant as well, for densities with fairly general supports. Our result holds regardless of the non-compactness or even infinite Lebesgue measure nature of the supports involved. This is accomplished by exploiting the fact that for a density p on $\mathbb{X} \subseteq \mathbb{R}^k$, though Lebesgue measure in $\mathbb{X}$ may be infinite if $\mathbb{X}$ is unbounded, $\mu(\cdot) \doteq \int_{\cdot} p dx$ is not. The value of the result lies on the fact that it does not require the construction of any additional function (such as f above), as it relies exclusively on the structure of the densities involved.

In connection with the previous claim, there exists evidence that convergence in distribution of the respective probability measures is not enough to have convergence of the corresponding differential entropies [7]. In that regard, an interesting example is presented and discussed in our work, reinforcing the importance of establishing general conditions for such convergence to take place. The paper also provides a characterization of convergence of differential entropies in terms of the Kullback-Liebler discriminant, for densities with fairly general supports too. Moreover, it is shown that under an appropriate boundedness condition on the densities involved, convergence in variation of probability measures does indeed guarantee the desired differential entropy convergence. In consequence, the uniform convergence conditions established in [7], that guarantee the convergence of entropy under convergence in variation of the underlying probability measures, are adapted in terms of appropriate boundedness requirements, providing a sufficient condition to be checked for the desired entropy convergence that does not require the explicit computation of the sequence terms in (1).

Our results can be applied in the discrete source case, by using the equivalence between convergence in distribution and in variation of probability measures in such situation. In particular, if the probability measures have finite support, the convergence of their respective entropies and the Kullback-Liebler discriminant follows immediately. In the case of probability mass functions with infinite supports, we exploit the afore mentioned equivalence between weak convergence and convergence in variation to establish the convergence of entropies and the Kullback-Liebler discriminant.

---

[1] We cannot define f to be the constant function, because if we do so, we will get $\int_{\mathbb{X}} M dx = M \int_{\mathbb{X}} dx = \infty$, for $M > 0$ and if $\mathbb{X}$ has infinite Lebesgue measure.

We finalize this section presenting the organization of the rest of the paper. In Section II we introduce notational and terminological conventions used throughout, as well as the necessary elements from the theory of convergence of probability measures. (Most of the definitions in this section apply to both the continuous and discrete case, Lebesgue measure not playing a role in the late, of course.) Sections III and IV consider the case of continuous random variables. In Section III we present an example where convergence in distribution of the underlying probability measures is not enough to have convergence of the associated differential entropies. Differential entropy convergence is then characterized in terms of the Kullback-Liebler discriminant for densities with fairly general supports, and it is shown that, under an appropriate boundedness condition on the densities involved, convergence in variation of probability measures does guarantee the desired differential entropy convergence. In Section IV we provide a general result for convergence of differential entropy and Kullback-Liebler discriminant under a pointwise convergence condition, taking into account both compactly and uncompactly supported densities. Finally, in section V we present our results for the discrete case.

II. PRELIMINARY ELEMENTS

In this section we introduce the concepts (and related notation) upon which we elaborate the present work. Our brief presentation includes the notions of weak convergence, convergence in variation and a measure-theoretic definition of entropy of probability measures.

A. Definitions

Let k be a positive integer, $\mathbb{R}^k$ the k-dimensional Euclidian space endowed with the usual Euclidian metric $\|\cdot - \cdot\|_2$, and $\mathcal{B}(\mathbb{R}^k)$ the collection of Borel sets in $\mathbb{R}^k$. Also, let $\mathbb{X} \in \mathcal{B}(\mathbb{R}^k)$, $\mathbb{X}$ closed, be a Polish subspace, i.e., $\mathbb{X}$ is separable (it has a countable dense subset) and complete (every Cauchy sequence in $\mathbb{X}$ converges to a point $x \in \mathbb{X}$) [8, 12]. Let $\mathcal{AC}(\mathbb{X})$ denote the collection of all probability measures μ on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$ which are absolutely continuous with respect to (w.r.t.) the Lebesgue measure in $\mathbb{X}$ (denoted as dx), i.e., having the representation

$$\mu(A) = \int_A \frac{d\mu}{dx} dx,$$

$A \in \mathcal{B}(\mathbb{X})$, with $\frac{d\mu}{dx} : \mathbb{X} \to \mathbb{R}_+ \doteq [0, \infty)$, Borel measurable, the Radon-Nikodym derivative or density of μ w.r.t. dx. Of course, when considering $\mathcal{AC}(\mathbb{X})$ we assume $\mathbb{X}$ is such that $\mathcal{AC}(\mathbb{X}) \neq \emptyset$ (i.e., $\mathbb{X}$ having strictly positive Lebesgue measure). In the same way, we denote as $\mathcal{AC}_+(\mathbb{X})$ the set of all $\mu \in \mathcal{AC}(\mathbb{X})$ for which $\frac{d\mu}{dx} > 0$ Lebesgue-almost everywhere on $\mathbb{X}$. In particular, $\mu \in \mathcal{AC}_+(\mathbb{X})$ implies that μ and dx are mutually absolutely continuous or equivalent, and that

$$\frac{dx}{d\mu}(x) \doteq \begin{cases} \left[\frac{d\mu}{dx}(x)\right]^{-1} & x \in \mathbb{X},\ \frac{d\mu}{dx}(x) > 0 \\ \alpha & x \in \mathbb{X},\ \frac{d\mu}{dx}(x) = 0 \end{cases} \tag{3}$$

with $\alpha \in \mathbb{R}_+$ any constant value, provides indeed a valid expression for the Radon-Nikodym derivative $\frac{dx}{d\mu}$.

In addition, let $f : \mathbb{X} \to \mathbb{R}$ be a real-valued function. Its support is the closure of the set of all $x \in \mathbb{X}$ where $f(x)$ is strictly positive, i.e.,

$$\text{support}\,(f) \doteq \overline{\{x \in \mathbb{X} : f(x) > 0\}},$$

where the overline $\overline{\{\cdot\}}$ denotes closure. In particular, we have that the Lebesgue measure of the sets support$(\frac{d\mu}{dx})$ and $\mathbb{X}$ coincide when $\mu \in \mathcal{AC}_+(\mathbb{X})$.

B. Convergence of probability measures

We now collect some basic definitions and results, in the context needed for the following sections of the paper. Throughout, $\mathcal{P}(\mathbb{X})$ denotes the collection of all probability measures on $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$ and $\mathcal{C}(\mathbb{X})$ (resp., $\mathcal{C}_b(\mathbb{X})$) the space of all continuous (resp., bounded and continuous), real-valued functions on $\mathbb{X}$.

Definition II.1 *A sequence $\{\mu_n\}_{n=1}^{\infty} \subseteq \mathcal{P}(\mathbb{X})$ is said to converge weakly to $\mu \in \mathcal{P}(\mathbb{X})$, denoted $\mu_n \Rightarrow \mu$ as $n \uparrow \infty$, if*

$$\int_{\mathbb{X}} f d\mu_n \to \int_{\mathbb{X}} f d\mu$$

as $n \uparrow \infty$ for each $f \in \mathcal{C}_b(\mathbb{X})$.

Since $\mathbb{X}$ is separable, weak convergence $\mu_n \Rightarrow \mu$ as $n \uparrow \infty$ of $\{\mu_n\}_{n=1}^{\infty} \subseteq \mathcal{P}(\mathbb{X})$ to $\mu \in \mathcal{P}(\mathbb{X})$ is equivalent to convergence $\rho(\mu_n, \mu) \to 0$, as $n \uparrow \infty$ as well, with $\rho(\cdot, \cdot)$ denoting the Prohorov metric on $\mathcal{P}(\mathbb{X}) \times \mathcal{P}(\mathbb{X})$, that is defined by

$$\rho(\sigma_1, \sigma_2) \doteq \inf\{\epsilon > 0 : \sigma_1(A) \leq \sigma_2(A^{\epsilon}) + \epsilon, \sigma_2(A) \leq \sigma_1(A^{\epsilon}) + \epsilon, \forall A \in \mathcal{B}(\mathbb{X})\},$$

$\sigma_1, \sigma_2 \in \mathcal{P}(\mathbb{X})$, where for $A \subseteq \mathbb{X}$, $\epsilon > 0$ and $y \in \mathbb{X}$, $A^{\epsilon} \doteq \{x \in \mathbb{X} : d(x, A) < \epsilon\}$ and $d(y, A) \doteq \inf\{\|y - z\|_2 : z \in A\}$. Note A^{ϵ} is open in $\mathbb{X}$, and hence $A^{\epsilon} \in \mathcal{B}(\mathbb{X})$. In addition, since $\mathbb{X}$ is not just separable but Polish, $(\mathcal{P}(\mathbb{X}), \rho)$ is Polish too [4].

Weak convergence in $\mathcal{P}(\mathbb{R}^k)$ is also equivalent to the standard convergence in distribution. (Note $\sigma \in \mathcal{P}(\mathbb{X})$ can always be looked at as an element of $\mathcal{P}(\mathbb{R}^k)$ when we note that $\sigma(A) = \sigma(A \cap \mathbb{X})$ for $A \in \mathcal{B}(\mathbb{R}^k)$.) Indeed, for $\{\mu_n\}_{n=1}^{\infty} \subseteq \mathcal{P}(\mathbb{R}^k)$ and $\mu \in \mathcal{P}(\mathbb{R}^k)$, we have $\mu_n \Rightarrow \mu$ as $n \uparrow \infty$ if and only if, as $n \uparrow \infty$ as well,

$$F_n(x) \to F(x)$$

at each $x \in \mathbb{R}^k$ point of continuity of F, where F_n and F denote the distribution functions associated to μ_n and μ, respectively, i.e.,

$$F_n(x) \doteq \mu_n\left(\times_{i=1}^{k}(-\infty, x_k]\right) \text{ and } F(x) \doteq \mu\left(\times_{i=1}^{k}(-\infty, x_k]\right)$$

for each $x = (x_1, \ldots, x_k) \in \mathbb{R}^k$. In fact, the following result holds (see Portmanteau's Theorem, [4, Theorem 2.1, p.16]).

Lemma II.1 *Let $\{\mu_n\}_{n=1}^{\infty} \subseteq \mathcal{P}(\mathbb{X})$ and $\mu \in \mathcal{P}(\mathbb{X})$. We have $\mu_n \Rightarrow \mu$ as $n \uparrow \infty$ if and only if, as $n \uparrow \infty$ as well,*

$$\mu_n(A) \to \mu(A)$$

for each $A \in \mathcal{B}(\mathbb{X})$ being a μ-continuity set, i.e., such that $\mu(\partial A) = 0$ with ∂A denoting the boundary of A: $\partial A \doteq \{x \in \mathbb{X} : x \in \overline{A}, x \notin A\}$.

Another important way of convergence for probability measures, stronger than weak convergence, is the convergence in variation associated with the distance in variation between probability measures.

Definition II.2 *The distance in variation between $\sigma_1 \in \mathcal{P}(\mathbb{X})$ and $\sigma_2 \in \mathcal{P}(\mathbb{X})$ is the real number $\|\sigma_1 - \sigma_2\|_V \in [0, 2]$ given by*

$$\|\sigma_1 - \sigma_2\|_V \doteq \sup_{\substack{\psi \in \mathcal{M}(\mathbb{X}) \\ |\psi| \leq \mathbf{1}}} \left| \int_{\mathbb{X}} \psi d\sigma_1 - \int_{\mathbb{X}} \psi d\sigma_2 \right|,$$

where $\mathcal{M}(\mathbb{X})$ denotes the collection of all $\mathbb{R}^*$-valued, Borel measurable functions on $\mathbb{X}$, $\mathbb{R}^* \doteq \mathbb{R} \cup \{\pm\infty\} = [-\infty, \infty]$ is the extended real line, and $\mathbf{1}(x) \doteq 1$, $x \in \mathbb{X}$. In particular, we have that $\| \cdot - \cdot \|_V : \mathcal{P}(\mathbb{X}) \times \mathcal{P}(\mathbb{X}) \to [0, 2]$ is indeed a metric on $\mathcal{P}(\mathbb{X})$. Moreover, a sequence $\{\mu_n\}_{n=1}^{\infty} \subseteq \mathcal{P}(\mathbb{X})$ is said to converge in variation to $\mu \in \mathcal{P}(\mathbb{X})$ if

$$\|\mu_n - \mu\|_V \to 0$$

as $n \uparrow \infty$.

Distance in variation can alternatively be characterized as

$$\|\sigma_1 - \sigma_2\|_V = 2 \sup_{A \in \mathcal{B}(\mathbb{X})} |\sigma_1(A) - \sigma_2(A)|,$$

$\sigma_1, \sigma_2 \in \mathcal{P}(\mathbb{X})$ [14].

As mentioned before, convergence in variation is stronger than weak convergence. Indeed, we have the following result, being a direct consequence of Lemma II.1.

Lemma II.2 *Let $\{\mu_n\}_{n=1}^{\infty} \subseteq \mathcal{P}(\mathbb{X})$ and $\mu \in \mathcal{P}(\mathbb{X})$. If $\|\mu_n - \mu\|_V \to 0$ as $n \uparrow \infty$, then $\mu_n \Rightarrow \mu$ as $n \uparrow \infty$.*

For $\mu \in \mathcal{P}(\mathbb{X})$ and $p \in [1, \infty)$ we define

$$L^p(d\mu) \doteq \left\{ f \in \mathcal{M}(\mathbb{X}) : \left[\int_{\mathbb{X}} |f|^p d\mu \right]^{\frac{1}{p}} < \infty \right\},$$

with the standard convention $0[\pm\infty] = 0$, and the $L^p(d\mu)$-norm of $f \in L^p(d\mu)$ as

$$\|f\|_{L^p(d\mu)} \doteq \left[\int_{\mathbb{X}} |f|^p d\mu \right]^{\frac{1}{p}}.$$

For $\mu \in \mathcal{P}(\mathbb{X})$ we denote as $L^\infty(d\mu)$ the space of all functions $f \in \mathcal{M}(\mathbb{X})$ which are bounded except possibly on a μ-null set, and define the $L^\infty(d\mu)$-norm of $f \in L^\infty(d\mu)$ as usual, i.e.,

$$\|f\|_{L^\infty(d\mu)} \doteq (\mu)\,\mathrm{ess\,sup}_{x \in \mathbb{X}} |f(x)|,$$

where for $g \in \mathcal{M}(\mathbb{X})$, $(\mu)\,\mathrm{ess\,sup}_{x \in \mathbb{X}} g(x)$, the essential supremum of g w.r.t. μ, is the infimum of $\sup_{x \in \mathbb{X}} h(x)$ as h ranges over all functions mapping $\mathbb{X}$ into $\mathbb{R}^*$ which are equal to g μ-almost everywhere. Thus, for $f \in L^\infty(d\mu)$ we have

$$\|f\|_{L^\infty(d\mu)} = \inf \{ M \in \mathbb{R}_+ : \mu \{ x \in \mathbb{X} : |f(x)| > M \} = 0 \}.$$

As known [13], for any given $\mu \in \mathcal{P}(\mathbb{X})$ the spaces $(L^p(d\mu), \|\cdot\|_{L^p(d\mu)})$, $p \in [1, \infty]$, become normed linear spaces with the usual addition and scalar multiplication of functions, and in fact Banach spaces, provided we treat measurable functions coinciding μ-almost everywhere as equivalent.

This notion is useful to determine another characterization of distance in variation, namely

$$\|\sigma_1 - \sigma_2\|_V = \left\| \frac{d\sigma_1}{dx} - \frac{d\sigma_2}{dx} \right\|_{L^1(dx)} = \int_{\mathbb{X}} \left| \frac{d\sigma_1}{dx} - \frac{d\sigma_2}{dx} \right| dx,$$

$\sigma_1, \sigma_2 \in \mathcal{AC}(\mathbb{X})$, [14].

C. Entropy

We conclude this section by writing a general definition of entropy of probability measures, on measure-theoretical grounds. In the sequel all logarithms are understood to be to the base 2.

The space of measures

$$\mathbb{H}(\mathbb{X}) \doteq \left\{ \mu \in \mathcal{AC}(X) : \log \left[\frac{d\mu}{dx} \right] \in L^1(d\mu) \right\},$$

with the convention $\log[0] = -\infty$, represents the set of well-defined entropy measures.

Definition II.3 *The Shannon Differential Entropy, associated to the underlying space $\mathbb{X}$, is the mapping $\mathcal{H} : \mathbb{H}(\mathbb{X}) \to \mathbb{R}$, assigning to each $\mu \in \mathbb{H}(\mathbb{X})$ the value $\mathcal{H}[\mu] \in \mathbb{R}$ given by*

$$\mathcal{H}[\mu] \doteq -\int_{\mathbb{X}} \log \left[\frac{d\mu}{dx} \right] d\mu.$$

$\mathcal{H}[\mu]$ is called the Shannon Differential Entropy of μ.

III. CONVERGENCE OF DIFFERENTIAL ENTROPY: CHARACTERIZATION IN TERMS OF WEAK CONVERGENCE, CONVERGENCE IN VARIATION AND THE KULLBACK-LIEBLER DISCRIMINANT

In this section we discuss an interesting example where weak convergence of probability measures is not enough for convergence of the associated differential entropies. We characterize the desired differential entropy convergence

for fairly general supported densities in terms of the Kullback-Leibler discriminant, also showing that under an appropriate boundedness condition on the densities involved, convergence in variation of the underlying probability measures does indeed guarantee differential entropy convergence.

Consider the space $\mathbb{X} = [0,1]$, and define the probability measures (taken from [4]) μ and μ_n in $\mathcal{AC}([0,1])$ by defining, for each $x \in [0,1]$ and $n \in \{1,2,\ldots\}$,

$$\frac{d\mu}{dx}(x) \doteq 1 \text{ and } \frac{d\mu_n}{dx}(x) \doteq n^2 \mathbf{1}\left\{x \in \bigcup_{k=0}^{n-1}\left(\frac{k}{n}, \frac{k}{n} + \frac{1}{n^3}\right)\right\},$$

where, as customary for $A \subseteq \mathbb{X}$, $\mathbf{1}\{x \in A\} \doteq 1$ if $x \in A$ and $\mathbf{1}\{x \in A\} \doteq 0$ if $x \in \mathbb{X} \setminus A$, with $\mathbb{X} \setminus A \doteq \{x \in \mathbb{X} : x \notin A\}$, the usual set-theoretic difference. Of course,

$$\mu(A) \doteq \int_A d\mu = \int_A \frac{d\mu}{dx} dx \text{ and } \mu_n(A) \doteq \int_A d\mu_n = \int_A \frac{d\mu_n}{dx} dx$$

for each $A \in \mathcal{B}([0,1])$. Note μ is nothing but Lebesgue measure in $[0,1]$. Also, it is easy to see that, for each $n \in \{1,2,\ldots\}$,

$$|\mu_n([0,x]) - \mu([0,x])| \leq \frac{1}{n}\left[1 - \frac{1}{n^2}\right]$$

for all $x \in [0,1]$, and therefore $\mu_n \Rightarrow \mu$ as $n \uparrow \infty$. On the other hand, we obviously have $\mu \in \mathbb{H}([0,1])$ and $\{\mu_n\}_{n=1}^{\infty} \subseteq \mathbb{H}([0,1])$. In fact,

$$\mathcal{H}[\mu] = -\int_{[0,1]} \log\left[\frac{d\mu}{dx}\right] d\mu = -\log[1] \int_{[0,1]} d\mu = 0$$

and, for each $n \in \{1,2,\ldots\}$,

$$\mathcal{H}[\mu_n] = -\int_{[0,1]} \log\left[\frac{d\mu_n}{dx}\right] d\mu_n - \int_{[0,1]} \log\left[\frac{d\mu_n}{dx}\right] \frac{d\mu_n}{dx} dx = -n^2 \log[n^2] \int_{\bigcup_{k=0}^{n-1}(\frac{k}{n}, \frac{k}{n} + \frac{1}{n^3})} dx = -2\log[n],$$

where for the last equality above we have used the fact that Lebesgue measure of the set $\bigcup_{k=0}^{n-1}(\frac{k}{n}, \frac{k}{n} + \frac{1}{n^3})$ is $\frac{1}{n^2}$. Hence, we have $\mu_n \Rightarrow \mu$ as $n \uparrow \infty$, but $\mathcal{H}[\mu_n] \downarrow -\infty$ as $n \uparrow \infty$, i.e., $\mathcal{H}[\mu_n] \downarrow -\infty \neq 0 = \mathcal{H}[\mu]$.

In the previous example weak convergence of probability measures is not enough for convergence of the respective differential entropies. It is interesting to note that in the example, though $\mu_n \Rightarrow \mu$ as $n \uparrow \infty$, pointwise convergence of the family of densities $\{\frac{d\mu_n}{dx}\}_{n=1}^{\infty}$ to $\frac{d\mu}{dx}$ fails to hold Lebesgue-almost everywhere. Indeed, if $A_n \doteq \bigcup_{k=0}^{n-1}(\frac{k}{n}, \frac{k}{n} + \frac{1}{n^3})$ then $\mu(A_n) = \frac{1}{n^2}$ for each $n \in \{1,2,\ldots\}$, and therefore

$$\sum_{n=1}^{\infty} \mu(A_n) = \sum_{n=1}^{\infty} \frac{1}{n^2} < \infty.$$

Hence, by Borel Lemma, [8, Lemma 3, p.78],

$$\mu\left(\limsup_{n\uparrow\infty} A_n\right) = 0,$$

where, as usual $\limsup_{n\uparrow\infty} A_n \doteq \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m$. But then, for μ-almost every $x \in [0,1]$ there exists $n_x \in \{1,\ldots,n\}$ such that $x \notin A_n$ for all $n \in \{n_x, n_x+1, \ldots\}$. Hence, we conclude

$$\frac{d\mu_n}{dx}(x) \to 0$$

as $n \uparrow \infty$, for μ-almost every $x \in [0,1]$ as well. Thus, we have

$$\frac{d\mu_n}{dx}(x) \to 0 \neq 1 = \frac{d\mu}{dx}(x)$$

for μ-almost every $x \in [0,1]$, i.e., pointwise convergence of $\{\frac{d\mu_n}{dx}\}_{n=1}^{\infty}$ to $\frac{d\mu}{dx}$ fails to hold Lebesgue-almost everywhere in $[0,1]$.

Instead of asking for an appropriate pointwise convergence condition, as we do in the next section, we now characterize the desired convergence $\mathcal{H}[\mu_n] \to \mathcal{H}[\mu]$ as $n \uparrow \infty$ in terms of the Kullback-Liebler discriminant. Some definitions are in order before establishing the result.

Let $\mathcal{AC}(\mathbb{X}||\mu)$ denote the set of all $\sigma \in \mathcal{P}(\mathbb{X})$ that are absolutely continuous w.r.t. μ, $\mu \in \mathcal{P}(\mathbb{X})$, i.e., having the representation

$$\sigma(A) = \int_A \frac{d\sigma}{d\mu} d\mu,$$

$A \in \mathcal{B}(\mathbb{X})$, with $\frac{d\sigma}{d\mu} : \mathbb{X} \to \mathbb{R}_+$, Borel measurable, the Radon-Nikodym derivative or density of σ w.r.t. μ. Also, we set

$$\mathbb{H}(\mathbb{X}||\mu) \doteq \left\{ \sigma \in \mathcal{AC}(\mathbb{X}||\mu) : \log\left[\frac{d\sigma}{d\mu}\right] \in L^1(d\sigma) \right\}.$$

Considering $\sigma \in \mathbb{H}(\mathbb{X}||\mu)$ and $\mathbb{X}_\mu^\sigma \doteq \text{support}(\frac{d\sigma}{d\mu})$, we have

$$\int_{\mathbb{X}} \log\left[\frac{d\sigma}{d\mu}\right] d\sigma = \int_{\mathbb{X}} \log\left[\frac{d\sigma}{d\mu}\right] \frac{d\sigma}{d\mu} d\mu = \int_{\mathbb{X}_\mu^\sigma} \log\left[\frac{d\sigma}{d\mu}\right] \frac{d\sigma}{d\mu} d\mu = \int_{\mathbb{X}_\mu^\sigma} \log\left[\frac{d\sigma}{d\mu}\right] d\sigma$$

and, by a standard application of Jensen's Inequality [6],

$$-\int_{\mathbb{X}_\mu^\sigma} \log\left[\frac{d\sigma}{d\mu}\right] d\sigma \leq \log\left[\mu(\mathbb{X}_\mu^\sigma)\right] \leq 0,$$

with equality if and only if $\frac{d\sigma}{d\mu} = \mathbf{1}$, σ-almost everywhere. Having noticed this, we make the following definition.

Definition III.1 *The Shannon Relative Entropy, relative to $\mu \in \mathcal{P}(\mathbb{X})$, is the mapping $\mathcal{D}[\cdot||\mu] : \mathbb{H}(\mathbb{X}||\mu) \to \mathbb{R}_+$, assigning to each $\sigma \in \mathbb{H}(\mathbb{X}||\mu)$ the value $\mathcal{D}[\sigma||\mu] \in \mathbb{R}_+$ given by*

$$\mathcal{D}[\sigma||\mu] \doteq \int_{\mathbb{X}} \log\left[\frac{d\sigma}{d\mu}\right] d\sigma.$$

$\mathcal{D}[\sigma||\mu]$ is called the Shannon Relative Entropy between σ and μ, or the Kullback-Liebler discriminant between σ and μ too.

As known, the Kullback-Liebler discriminant does not constitute a distance between probability measures: it is not symmetric and does not satisfies the triangle inequality; indeed, $\sigma \in \mathbb{H}(\mathbb{X}||\mu)$ does not even imply $\mu \in \mathcal{AC}(\mathbb{X}||\sigma)$. It is widely used as a notion of closeness between probability measures though, mainly because, as discussed above, $\mathcal{D}[\sigma||\mu] \geq 0$ with equality if and only if $\frac{d\sigma}{d\mu} = \mathbf{1}$, σ-almost everywhere.

Before stating the result in the next theorem, we make the following remarks.

Remark III.1 *If $\sigma \in \mathcal{AC}(\mathbb{X})$ and $\mu \in \mathcal{AC}_+(\mathbb{X})$, then $\sigma \in \mathcal{AC}(\mathbb{X}||\mu)$. In fact, we can define*

$$\frac{d\sigma}{d\mu} \doteq \frac{d\sigma}{dx}\frac{dx}{d\mu}$$

with $\frac{dx}{d\mu}$ given by (3), as we do throughout. Then, when $\sigma, \mu \in \mathcal{AC}_+(\mathbb{X})$ we have $\sigma \in \mathcal{AC}(\mathbb{X}||\mu)$ and $\mu \in \mathcal{AC}(\mathbb{X}||\sigma)$, i.e., σ and μ are mutually absolutely continuous or equivalent. Moreover, on the (partial) converse direction, $\sigma \in \mathcal{AC}(\mathbb{X})$ if $\sigma \in \mathcal{AC}(\mathbb{X}||\mu)$ and $\mu \in \mathcal{AC}(\mathbb{X})$, and we have $\frac{d\sigma}{dx} = \frac{d\sigma}{d\mu}\frac{d\mu}{dx}$, Lebesgue-almost everywhere, and $\frac{d\sigma}{d\mu} = \frac{d\sigma}{dx}[\frac{d\mu}{dx}]^{-1}$, μ-almost everywhere. These facts will be used in the sequel without any further comment.

Remark III.2 *From Pinsker's Inequality (see for example [10]), for any $\mu \in \mathcal{P}(\mathbb{X})$ and $\{\mu_n\}_{n=1}^{\infty} \subseteq \mathbb{H}(\mathbb{X}||\mu)$ we have*

$$\|\mu_n - \mu\|_V \leq \sqrt{2\mathcal{D}[\mu_n||\mu]},$$

for each $n \in \{1, 2, \ldots\}$ and therefore, convergence $\mathcal{D}[\mu_n||\mu] \to 0$ as $n \uparrow \infty$ implies convergence $\|\mu_n - \mu\|_V \to 0$, as $n \uparrow \infty$ as well.

Theorem III.1 *Let $\{\mu_n\}_{n=1}^{\infty} \subseteq \mathbb{H}(\mathbb{X})$ and $\mu \in \mathcal{AC}(\mathbb{X})$ be such that $\frac{d\mu}{dx}(x) > 0$, for each $x \in \mathbb{X}$, and $\log[\frac{d\mu}{dx}] \in \mathcal{C}_b(\mathbb{X})$. Then, $\{\mu_n\}_{n=1}^{\infty} \subseteq \mathbb{H}(\mathbb{X}||\mu)$, $\mu \in \mathbb{H}(\mathbb{X})$ and the following assertions are equivalent.*

1. *$\mu_n \Rightarrow \mu$ and $\mathcal{H}[\mu_n] \to \mathcal{H}[\mu]$ as $n \uparrow \infty$.*
2. *$\mathcal{D}[\mu_n||\mu] \to 0$ as $n \uparrow \infty$.*

Proof. Since $\log[\frac{d\mu}{dx}]$ is in particular bounded on $\mathbb{X}$, we obviously have $\mu \in \mathbb{H}(\mathbb{X})$. In addition,

$$\int_{\mathbb{X}} \left| \log\left[\frac{d\mu_n}{d\mu}\right] \right| d\mu_n \leq \int_{\mathbb{X}} \left| \log\left[\frac{d\mu_n}{dx}\right] \right| d\mu_n + \int_{\mathbb{X}} \left| \log\left[\frac{d\mu}{dx}\right] \right| d\mu_n < \infty, \tag{4}$$

since also $\{\mu_n\}_{n=1}^{\infty} \subseteq \mathbb{H}(\mathbb{X})$. In particular, $\{\mu_n\}_{n=1}^{\infty} \subseteq \mathbb{H}(\mathbb{X}||\mu)$. Now, from equation (4) we may write

$$\mathcal{D}[\mu_n||\mu] = \int_{\mathbb{X}} \log\left[\frac{d\mu_n}{d\mu}\right] d\mu_n = \int_{\mathbb{X}} \log\left[\frac{d\mu_n}{dx}\right] d\mu_n - \int_{\mathbb{X}} \log\left[\frac{d\mu}{dx}\right] d\mu_n \tag{5}$$

and, since also $\mu \in \mathbb{H}(\mathbb{X})$, from equation (5) we conclude

$$\mathcal{D}[\mu_n||\mu] = \mathcal{H}[\mu] - \mathcal{H}[\mu_n] + \int_{\mathbb{X}} \log\left[\frac{d\mu}{dx}\right] d\mu - \int_{\mathbb{X}} \log\left[\frac{d\mu}{dx}\right] d\mu_n. \tag{6}$$

But, since $\log[\frac{d\mu}{dx}] \in \mathcal{C}_b(\mathbb{X})$, if $\mu_n \Rightarrow \mu$ as $n \uparrow \infty$ we conclude that

$$\int_{\mathbb{X}} \log\left[\frac{d\mu}{dx}\right] d\mu_n \to \int_{\mathbb{X}} \log\left[\frac{d\mu}{dx}\right] d\mu,$$

as $n \uparrow \infty$ as well, equation (6) proving then the implication (i) $\Rightarrow$ (ii). The converse implication (ii) $\Rightarrow$ (i) also follows from equation (6), in view of Remark III.2 and Lemma II.2. The theorem is then proved. ■

Remark III.3 *Consider $\mu \in \mathcal{AC}(\mathbb{X})$ with $\frac{d\mu}{dx}(x) > 0$ for each $x \in \mathbb{X}$. Then, $\mathbb{X}$ does not necessarily need to be bounded for $\log[\frac{d\mu}{dx}]$ to be bounded on $\mathbb{X}$. Indeed, consider for instance the uniform distribution on any unbounded set $\mathbb{X}$ ($\subseteq \mathbb{R}^k$, $k > 1$) having finite and strictly positive Lebesgue measure, as for example in $\mathbb{R}^2$*

$$\mathbb{X} = \left\{x = (x_1, x_2) \in \mathbb{R}_+^2 : e^{-\lambda x_1} \geq x_2\right\}$$

with $\lambda \in (0, \infty)$. $\mathbb{X}$ so defined is an unbounded subset of $\mathbb{R}^2$. However, since the Lebesgue measure of $\mathbb{X}$ is $\int_{\mathbb{X}} dx = \lambda^{-1} \in (0, \infty)$, the uniform distribution μ_0 on $\mathbb{X}$ satisfies, for all $x \in \mathbb{X}$,

$$\log\left[\frac{d\mu_0}{dx}(x)\right] = \log\left[\left(\int_{\mathbb{X}} dx\right)^{-1}\right] = \log[\lambda],$$

trivially bounded on $\mathbb{X}$. In the same way, the set $\mathbb{X}$ does not necessarily need to be bounded for $\log[\frac{d\mu}{dx}]$ to be an element of $L^{\infty}(dx)$.

Remark III.4 *Since $\mathbb{X}$ is closed, we have $\log[\frac{d\mu}{dx}] \in \mathcal{C}_b(\mathbb{X})$ whenever $\mathbb{X}$ is in addition bounded and $\mu \in \mathcal{AC}(\mathbb{X})$ with $\frac{d\mu}{dx}(x) > 0$, for each $x \in \mathbb{X}$, and $\frac{d\mu}{dx} \in \mathcal{C}(\mathbb{X})$. Indeed, if $\mathbb{X} \subseteq \mathbb{R}^k$ is closed and bounded it is then compact, and therefore for $\mu \in \mathcal{AC}(\mathbb{X})$ with $\frac{d\mu}{dx}(x) > 0$, for each $x \in \mathbb{X}$, and $\frac{d\mu}{dx} \in \mathcal{C}(\mathbb{X})$, there exist m and M in $(0, \infty)$, $m \leq M$, such that $\frac{d\mu}{dx}(x) \in [m, M]$, for each $x \in \mathbb{X}$ as well. Thus, $\log[\frac{d\mu}{dx}] \in \mathcal{C}_b(\mathbb{X})$. Also, note that for the purpose of Theorem III.1 we can always take $\mathbb{X}$ as being bounded for $\{\mu_n\}_{n=1}^{\infty} \subseteq \mathcal{AC}(\mathbb{R}^k)$ and $\mu \in \mathcal{AC}(\mathbb{R}^k)$ when*

$$\bigcup_{n=1}^{\infty} K_n \subseteq K$$

and K is bounded, where $K \doteq \text{support}(\frac{d\mu}{dx})$ and $K_n \doteq \text{support}(\frac{d\mu_n}{dx})$ for each $n \in \{1, 2, \ldots\}$ (all the supports being taken w.r.t. $\mathbb{R}^k$). Indeed, with $\mathbb{X} \doteq K$ we then have $\frac{d\mu}{dx} > 0$ on $\mathbb{X}$ and each μ_n, the same as μ, is concentrated on $\mathbb{X}$, i.e., $\mu_n(\mathbb{X}) = 1$.

Remark III.5 *The probability measures considered in the example at the beginning of this section satisfies all hypotheses of Theorem III.1 and, in addition, $\mu_n \Rightarrow \mu$ as $n \uparrow \infty$. However, similar to the differential entropy convergence failure, we have $\mathcal{D}[\mu_n||\mu] = 2\log[n] \uparrow \infty$ as $n \uparrow \infty$, i.e., the convergence $\mathcal{D}[\mu_n||\mu] \to 0$ as $n \uparrow \infty$ fails to hold. Also,*

$$\sup_{A \in \mathcal{B}(\mathbb{X})} |\mu_n(A) - \mu(A)| \geq |\mu_n(A_n) - \mu(A_n)| = 1 - \frac{1}{n^2},$$

for each $n \in \{1, 2, \ldots\}$, and hence the convergence $\|\mu_n - \mu\|_V \to 0$ as $n \uparrow \infty$ fails to hold too. In light of Theorem III.1, we have failure of differential entropy convergence due to failure of the corresponding convergence for the Kullback-Liebler discriminant, due in turn and in light of Remark III.2 to the respective failure of convergence in variation.

Though weak convergence does not guarantee convergence of differential entropy, convergence in variation does it under an appropriate boundedness condition on the densities involved. The result is the following.

Theorem III.2 *Let $\{\mu_n\}_{n=1}^{\infty} \subseteq \mathcal{AC}_+(\mathbb{X})$ and $\mu \in \mathcal{AC}_+(\mathbb{X})$ be such that $\log[\frac{d\mu}{dx}] \in L^\infty(dx)$ and $\{\log[\frac{d\mu_n}{dx}]\}_{n=1}^{\infty} \subseteq L^\infty(dx)$. Assume that*

$$M \doteq \sup_{n \in \{1,2,\ldots\}} \left\| \log\left[\frac{d\mu_n}{dx}\right] \right\|_{L^\infty(dx)} < \infty. \tag{7}$$

Then, $\{\mu_n\}_{n=1}^{\infty} \subseteq \mathbb{H}(\mathbb{X}) \cap \mathbb{H}(\mathbb{X}||\mu)$, $\mu \in \mathbb{H}(\mathbb{X})$ and, if $\|\mu_n - \mu\|_V \to 0$ as $n \uparrow \infty$, we have both

$$\mathcal{D}[\mu_n||\mu] \to 0 \text{ and } \mathcal{H}[\mu_n] \to \mathcal{H}[\mu],$$

as $n \uparrow \infty$ as well.

Proof. First, since both

$$\log\left[\frac{d\mu}{dx}\right] \in L^\infty(dx) \text{ and } \left\{\log\left[\frac{d\mu_n}{dx}\right]\right\}_{n=1}^{\infty} \subseteq L^\infty(dx),$$

we have $\mu \in \mathbb{H}(\mathbb{X})$ and $\{\mu_n\}_{n=1}^{\infty} \subseteq \mathbb{H}(\mathbb{X})$. Therefore, we may write

$$\begin{aligned} |\mathcal{H}[\mu_n] - \mathcal{H}[\mu]| &= \left| \int_{\mathbb{X}} \log\left[\frac{d\mu_n}{dx}\right] d\mu_n - \int_{\mathbb{X}} \log\left[\frac{d\mu}{dx}\right] d\mu \right| = \left| \int_{\mathbb{X}} \left(\frac{d\mu_n}{dx} \log\left[\frac{d\mu_n}{dx}\right] - \frac{d\mu}{dx} \log\left[\frac{d\mu}{dx}\right] \right) dx \right| \\ &\leq \int_{\mathbb{X}} \left| \frac{d\mu_n}{dx} \log\left[\frac{d\mu_n}{dx}\right] - \frac{d\mu}{dx} \log\left[\frac{d\mu}{dx}\right] \right| dx. \end{aligned} \tag{8}$$

Now, for each $n \in \{1, 2, \ldots\}$ we have

$$\left| \frac{d\mu_n}{dx}(x) \log\left[\frac{d\mu_n}{dx}(x)\right] - \frac{d\mu}{dx}(x) \log\left[\frac{d\mu}{dx}(x)\right] \right| \leq \left| \log\left[\frac{d\mu_n}{dx}(x)\right] \right| \left| \frac{d\mu_n}{dx}(x) - \frac{d\mu}{dx}(x) \right| + \frac{d\mu}{dx}(x) \left| \log\left[\frac{d\mu_n}{d\mu}(x)\right] \right|, \tag{9}$$

for Lebesgue-almost every $x \in \mathbb{X}$. For each $n \in \{1, 2, \ldots\}$ we also have, for Lebesgue-almost every $x \in \mathbb{X}$ as well,

$$\left| \log\left[\frac{d\mu_n}{d\mu}(x)\right] \right| \leq \frac{\log[M']}{M' - 1} \left| \frac{d\mu_n}{d\mu}(x) - 1 \right|, \tag{10}$$

since

$$\frac{d\mu_n}{d\mu}(x) = \frac{d\mu_n}{dx}(x) \left[\frac{d\mu}{dx}(x)\right]^{-1} \geq 2^{-M} 2^{-\left\| \frac{d\mu}{dx} \right\|_{L^\infty(dx)}} = 2^{-\left(M + \left\| \frac{d\mu}{dx} \right\|_{L^\infty(dx)}\right)} \doteq M' \in (0, 1)$$

for each $n \in \{1, 2, \ldots\}$ and Lebesgue-almost every $x \in \mathbb{X}$ too, and

$$|\log[a]| \leq \frac{\log[a_0]}{a_0 - 1} |a - 1|$$

for all $a \in [a_0, \infty)$, with $a_0 \in (0, 1)$. We also have

$$\left| \frac{d\mu_n}{d\mu}(x) - 1 \right| = \left[\frac{d\mu}{dx}(x)\right]^{-1} \left| \frac{d\mu_n}{dx}(x) - \frac{d\mu}{dx}(x) \right| \tag{11}$$

for each $n \in \{1, 2, \ldots\}$ and Lebesgue-almost every $x \in \mathbb{X}$. Hence, from equations (9), (10) and (11) we conclude

$$\left| \frac{d\mu_n}{dx}(x) \log \left[\frac{d\mu_n}{dx}(x) \right] - \frac{d\mu}{dx}(x) \log \left[\frac{d\mu}{dx}(x) \right] \right| \leq \left[M + \frac{\log [M']}{M' - 1} \right] \left| \frac{d\mu_n}{dx}(x) - \frac{d\mu}{dx}(x) \right|,$$

for each $n \in \{1, 2, \ldots\}$ and Lebesgue-almost every $x \in \mathbb{X}$ as well, and therefore, from equation (8),

$$|\mathcal{H}[\mu_n] - \mathcal{H}[\mu]| \leq \left[M + \frac{\log [M']}{M' - 1} \right] \left\| \frac{d\mu_n}{dx} - \frac{d\mu}{dx} \right\|_{L^1(dx)}. \tag{12}$$

In the same way, since

$$\int_{\mathbb{X}} \left| \log \left[\frac{d\mu_n}{d\mu} \right] \right| d\mu_n \leq M + \left\| \log \left[\frac{d\mu}{dx} \right] \right\|_{L^\infty(dx)} < \infty,$$

and therefore $\{\mu_n\}_{n=1}^{\infty} \subseteq \mathbb{H}(\mathbb{X}||\mu)$, from equations (10) and (11) it is easy to see that

$$\mathcal{D}[\mu_n||\mu] \leq \frac{\log [M']}{M' (M' - 1)} \left\| \frac{d\mu_n}{dx} - \frac{d\mu}{dx} \right\|_{L^1(dx)}. \tag{13}$$

The last part of the theorem then follows from equations (12) and (13) since, if $\|\mu_n - \mu\|_V \to 0$ as $n \uparrow \infty$, then

$$\left\| \frac{d\mu_n}{dx} - \frac{d\mu}{dx} \right\|_{L^1(dx)} (= \|\mu_n - \mu\|_V) \to 0,$$

as $n \uparrow \infty$ as well. ■

Remark III.6 *The reader can verify that the arguments leading to the proof of Theorem III.2 require for the supports of the densities $\frac{d\mu}{dx}$ and $\frac{d\mu_n}{dx}$, $n \in \{1, 2, \ldots\}$, when regarded as densities in $\mathbb{R}^k$, to at most pairwise differ by a Lebesgue-null set. The set $\mathbb{X}$ in the statement of the theorem can then be taken as the intersection of all the afore mentioned supports. Indeed, for such a $\mu \in \mathcal{AC}(\mathbb{R}^k)$ and $\{\mu_n\}_{n=1}^{\infty} \subseteq \mathcal{AC}(\mathbb{R}^k)$ we have, with $\mu_0 \doteq \mu$, $K_0 \doteq \text{support}(\frac{d\mu_0}{dx})$, $K_n \doteq \text{support}(\frac{d\mu_n}{dx})$ for each $n \in \{1, 2, \ldots\}$ (all the supports being taken w.r.t. $\mathbb{R}^k$) and*

$$\mathbb{X} \doteq \bigcap_{n=0}^{\infty} K_n,$$

that for each $m \in \{0, 1, 2, \ldots\}$

$$K_m \setminus \mathbb{X} = \bigcup_{n=0}^{\infty} (K_m \setminus K_n),$$

a Lebesgue-null set, and therefore, since $\{\mu_n\}_{n=0}^{\infty} \subseteq \mathcal{AC}(\mathbb{R}^k)$, that each element in the sequence $\{\mu_n\}_{n=0}^{\infty}$ is concentrated on $\mathbb{X}$. Moreover, $\{\mu_n\}_{n=0}^{\infty} \subseteq \mathcal{AC}_+(\mathbb{X})$.

In view of Remark III.2 and Lemma II.2, we have the following direct corollary to Theorems III.1 and III.2.

Corollary III.1 *Let $\{\mu_n\}_{n=1}^{\infty} \subseteq \mathcal{AC}_+(\mathbb{X})$ and $\mu \in \mathcal{AC}(\mathbb{X})$ be such that $\frac{d\mu}{dx}(x) > 0$, for each $x \in \mathbb{X}$, $\log[\frac{d\mu}{dx}] \in \mathcal{C}_b(\mathbb{X})$ and $\{\log[\frac{d\mu_n}{dx}]\}_{n=1}^{\infty} \subseteq L^\infty(dx)$. Assume that*

$$\sup_{n \in \{1, 2, \ldots\}} \left\| \log \left[\frac{d\mu_n}{dx} \right] \right\|_{L^\infty(dx)} < \infty.$$

Then, $\{\mu_n\}_{n=1}^{\infty} \subseteq \mathbb{H}(\mathbb{X}) \cap \mathbb{H}(\mathbb{X}||\mu)$, $\mu \in \mathbb{H}(\mathbb{X})$ and the following assertions are equivalent.

1. *$\mu_n \Rightarrow \mu$ and $\mathcal{H}[\mu_n] \to \mathcal{H}[\mu]$ as $n \uparrow \infty$.*
2. *$\mathcal{D}[\mu_n||\mu] \to 0$ as $n \uparrow \infty$.*
3. *$\|\mu_n - \mu\|_V \to 0$ as $n \uparrow \infty$.*

IV. POINTWISE CONVERGENCE AND DIFFERENTIAL ENTROPY CONVERGENCE

In this section we provide a general result for convergence of Shannon Differential Entropy, and Kullback-Liebler discriminant as well, under an appropriate pointwise convergence condition. We take into account both compactly and uncomplactly supported densities. As mentioned in Section I, the proof is based on exploiting absolute continuity properties of measures, in conjunction with a suitable boundedness condition and the dominated convergence theorem. The result is the following.

Theorem IV.1 *Let $\mu \in \mathbb{H}(\mathbb{X})$ and $\{\mu_n\}_{n=1}^{\infty} \subseteq \mathcal{AC}(\mathbb{X}||\mu)$ be such that $\frac{d\mu_n}{d\mu}(x) \to 1$ as $n \uparrow \infty$, for μ-almost every $x \in \mathbb{X}$, and $\{\frac{d\mu_n}{d\mu}\}_{n=1}^{\infty} \subseteq L^{\infty}(d\mu)$. Assume that*

$$M \doteq \sup_{n\in\{1,2,\ldots\}} \left\| \frac{d\mu_n}{d\mu} \right\|_{L^{\infty}(d\mu)} < \infty. \tag{14}$$

Then, $\{\mu_n\}_{n=1}^{\infty} \subseteq \mathbb{H}(\mathbb{X}) \cap \mathbb{H}(\mathbb{X}||\mu)$ and we have both

$$\mathcal{D}[\mu_n||\mu] \to 0 \text{ and } \mathcal{H}[\mu_n] \to \mathcal{H}[\mu]$$

as $n \uparrow \infty$.

Proof. First, for each $n \in \{1, 2, \ldots\}$ we have

$$\int_{\mathbb{X}} \left| \log\left[\frac{d\mu}{dx}\right] \right| d\mu_n = \int_{\mathbb{X}} \left| \log\left[\frac{d\mu}{dx}\right] \right| \frac{d\mu_n}{d\mu} d\mu \leq M \int_{\mathbb{X}} \left| \log\left[\frac{d\mu}{dx}\right] \right| d\mu < \infty$$

($\mu \in \mathbb{H}(\mathbb{X})$). Condition (14) in the statement of the theorem also implies that $\{\frac{d\mu_n}{d\mu} \log[\frac{d\mu_n}{d\mu}]\}_{n=1}^{\infty} \subseteq L^{\infty}(d\mu)$ with

$$M' \doteq \sup_{n\in\{1,2,\ldots\}} \left\| \frac{d\mu_n}{d\mu} \log\left[\frac{d\mu_n}{d\mu}\right] \right\|_{L^{\infty}(d\mu)} < \infty, \tag{15}$$

and therefore, $\{\mu_n\}_{n=1}^{\infty} \subseteq \mathbb{H}(\mathbb{X}||\mu)$. Indeed, for each $n \in \{1, 2, \ldots\}$,

$$\int_{\mathbb{X}} \left| \log\left[\frac{d\mu_n}{d\mu}\right] \right| d\mu_n = \int_{\mathbb{X}} \left| \log\left[\frac{d\mu_n}{d\mu}\right] \right| \frac{d\mu_n}{d\mu} d\mu \leq \int_{\mathbb{X}} M' d\mu, \tag{16}$$

and $\int_{\mathbb{X}} M' d\mu = M'\mu(\mathbb{X}) = M' < \infty$. Hence, for each $n \in \{1, 2, \ldots\}$ we also have

$$\int_{\mathbb{X}} \left| \log\left[\frac{d\mu_n}{dx}\right] \right| d\mu_n \leq \int_{\mathbb{X}} \left| \log\left[\frac{d\mu_n}{d\mu}\right] \right| d\mu_n + \int_{\mathbb{X}} \left| \log\left[\frac{d\mu}{dx}\right] \right| d\mu_n < \infty,$$

thus $\{\mu_n\}_{n=1}^{\infty} \subseteq \mathbb{H}(\mathbb{X})$, and we may write

$$\begin{aligned} |\mathcal{H}[\mu_n] - \mathcal{H}[\mu]| &= \left| \int_{\mathbb{X}} \log\left[\frac{d\mu_n}{dx}\right] d\mu_n - \int_{\mathbb{X}} \log\left[\frac{d\mu}{dx}\right] d\mu \right| \\ &\leq \left| \int_{\mathbb{X}} \log\left[\frac{d\mu_n}{dx}\right] d\mu_n - \int_{\mathbb{X}} \log\left[\frac{d\mu}{dx}\right] d\mu_n \right| + \left| \int_{\mathbb{X}} \log\left[\frac{d\mu}{dx}\right] d\mu_n - \int_{\mathbb{X}} \log\left[\frac{d\mu}{dx}\right] d\mu \right|, \end{aligned}$$

i.e.,

$$|\mathcal{H}[\mu_n] - \mathcal{H}[\mu]| \leq \mathcal{D}[\mu_n||\mu] + \left| \int_{\mathbb{X}} \log\left[\frac{d\mu}{dx}\right] d\mu_n - \int_{\mathbb{X}} \log\left[\frac{d\mu}{dx}\right] d\mu \right|, \tag{17}$$

for each $n \in \{1, 2, \ldots\}$ as well. But,

$$\mathcal{D}[\mu_n||\mu] = \int_{\mathbb{X}} \log\left[\frac{d\mu_n}{d\mu}\right] d\mu_n = \int_{\mathbb{X}} \log\left[\frac{d\mu_n}{d\mu}\right] \frac{d\mu_n}{d\mu} d\mu \tag{18}$$

for each $n \in \{1, 2, \ldots\}$ and, as already used in equation (16), from (15) it follows that, for each $n \in \{1, 2, \ldots\}$,

$$\frac{d\mu_n}{d\mu}(x) \left| \log\left[\frac{d\mu_n}{d\mu}(x)\right] \right| \leq M'$$

for μ-almost every $x \in \mathbb{X}$. Since also $\{\frac{d\mu_n}{d\mu}\log[\frac{d\mu_n}{d\mu}]\}_{n=1}^{\infty}$ converges pointwise μ-almost everywhere to $\mathbf{0}$ on $\mathbb{X}$ as $n \uparrow \infty$, where $\mathbf{0}(x) \doteq 0$, $x \in \mathbb{X}$, by Lebesgue's Dominated Convergence Theorem (see for example [13]) we conclude

$$\int_{\mathbb{X}} \log\left[\frac{d\mu_n}{d\mu}\right]\frac{d\mu_n}{d\mu}d\mu \to 0, \tag{19}$$

as $n \uparrow \infty$ as well. The claimed convergence $\mathcal{D}[\mu_n||\mu] \to 0$ as $n \uparrow \infty$ then follows from equations (18) and (19). Now, to establish the remaining claimed convergence $\mathcal{H}[\mu_n] \to \mathcal{H}[\mu]$ as $n \uparrow \infty$, we note that for each $n \in \{1, 2, \ldots\}$ we also have

$$\left|\int_{\mathbb{X}} \log\left[\frac{d\mu}{dx}\right]d\mu_n - \int_{\mathbb{X}} \log\left[\frac{d\mu}{dx}\right]d\mu\right| = \left|\int_{\mathbb{X}} \log\left[\frac{d\mu}{dx}\right]\frac{d\mu_n}{d\mu}d\mu - \int_{\mathbb{X}} \log\left[\frac{d\mu}{dx}\right]d\mu\right| \leq \int_{\mathbb{X}}\left|\log\left[\frac{d\mu}{dx}\right]\left(\frac{d\mu_n}{d\mu} - \mathbf{1}\right)\right|d\mu. \tag{20}$$

But, since $\{\frac{d\mu_n}{d\mu}\}_{n=1}^{\infty}$ converges pointwise μ-almost everywhere to $\mathbf{1}$ on $\mathbb{X}$ as $n \uparrow \infty$, we conclude

$$\log\left[\frac{d\mu}{dx}\right]\left(\frac{d\mu_n}{d\mu} - \mathbf{1}\right) \to \mathbf{0},$$

μ-almost everywhere on $\mathbb{X}$ and as $n \uparrow \infty$ as well. In addition, since we obviously also have $\{\frac{d\mu_n}{d\mu} - \mathbf{1}\}_{n=1}^{\infty} \subseteq L^{\infty}(d\mu)$ and

$$M'' \doteq \sup_{n\in\{1,2,\ldots\}}\left\|\frac{d\mu_n}{d\mu} - \mathbf{1}\right\|_{L^{\infty}(d\mu)} \leq \sup_{n\in\{1,2,\ldots\}}\left\|\frac{d\mu_n}{d\mu}\right\|_{L^{\infty}(d\mu)} + 1 = M + 1 < \infty,$$

we conclude that, for each $n \in \{1, 2, \ldots\}$,

$$\left|\log\left[\frac{d\mu}{dx}(x)\right]\left(\frac{d\mu_n}{d\mu}(x) - 1\right)\right| \leq M''\left|\log\left[\frac{d\mu}{dx}(x)\right]\right|$$

for μ-almost every $x \in \mathbb{X}$. But,

$$\int_{\mathbb{X}} M''\left|\log\left[\frac{d\mu}{dx}\right]\right|d\mu = M''\int_{\mathbb{X}}\left|\log\left[\frac{d\mu}{dx}\right]\right|d\mu < \infty$$

($\mu \in \mathbb{H}(\mathbb{X})$). Hence, once again by Lebesgue's Dominated Convergence Theorem we conclude

$$\int_{\mathbb{X}}\left|\log\left[\frac{d\mu}{dx}\right]\left(\frac{d\mu_n}{d\mu} - \mathbf{1}\right)\right|d\mu \to 0$$

as $n \uparrow \infty$, and therefore from equation (20) we also have

$$\int_{\mathbb{X}} \log\left[\frac{d\mu}{dx}\right]d\mu_n \to \int_{\mathbb{X}} \log\left[\frac{d\mu}{dx}\right]d\mu, \tag{21}$$

as $n \uparrow \infty$ as well. The claimed convergence $\mathcal{H}[\mu_n] \to \mathcal{H}[\mu]$ as $n \uparrow \infty$ now follows from equations (17), (18), (19) and (21), proving the theorem. ■

Remark IV.1 *If $\{\mu_n\}_{n=1}^{\infty} \subseteq \mathcal{AC}(\mathbb{X})$ and $\mu \in \mathcal{AC}_+(\mathbb{X})$, then μ-almost everywhere pointwise convergence $\frac{d\mu_n}{d\mu} \to \mathbf{1}$ as $n \uparrow \infty$ is equivalent to Lebesgue-almost everywhere pointwise convergence $\frac{d\mu_n}{dx} \to \frac{d\mu}{dx}$, as $n \uparrow \infty$ as well (both on $\mathbb{X}$, of course).*

Remark IV.2 *If $\{\mu_n\}_{n=1}^{\infty} \subseteq \mathcal{AC}(\mathbb{X})$ and $\mu \in \mathcal{AC}(\mathbb{X})$ with $\{\frac{d\mu_n}{dx}\}_{n=1}^{\infty}$ converging pointwise Lebesgue-almost everywhere to $\frac{d\mu}{dx}$ on $\mathbb{X}$ as $n \uparrow \infty$, then $\|\mu_n - \mu\|_V \to 0$, as $n \uparrow \infty$ as well. Indeed,*

$$\|\mu_n - \mu\|_V = \left\|\frac{d\mu_n}{dx} - \frac{d\mu}{dx}\right\|_{L^1(dx)} = \int_{\mathbb{X}}\left|\frac{d\mu_n}{dx} - \frac{d\mu}{dx}\right|dx \to 0$$

as $n \uparrow \infty$, the convergence following from Scheffé's Lemma, [16, Lemma 5.10, p.55]. Therefore, when $\{\mu_n\}_{n=1}^{\infty} \subseteq \mathcal{AC}_+(\mathbb{X})$ and $\mu \in \mathcal{AC}_+(\mathbb{X})$, by going from convergence in variation in Theorem III.2, to pointwise convergence of the corresponding densities in Theorem IV.1 (see Remark IV.1 above), we are able to relax the corresponding boundedness

condition from (7) to (14). Indeed, for $\{\mu_n\}_{n=1}^{\infty} \subseteq \mathcal{AC}_+(\mathbb{X})$ *and* $\mu \in \mathcal{AC}_+(\mathbb{X})$ *satisfying* $\log[\frac{d\mu}{dx}] \in L^\infty(dx)$ *and* $\{\log[\frac{d\mu_n}{dx}]\}_{n=1}^{\infty} \subseteq L^\infty(dx)$ *with* $\sup_{n\in\{1,2,\ldots\}} \|\log[\frac{d\mu_n}{dx}]\|_{L^\infty(dx)} < \infty$, *we have* $\{\frac{d\mu_n}{d\mu}\}_{n=1}^{\infty} \subseteq L^\infty(d\mu)$ *and*

$$\sup_{n\in\{1,2,\ldots\}} \left\| \frac{d\mu_n}{d\mu} \right\|_{L^\infty(d\mu)} \leq \frac{2^{\sup_{n\in\{1,2,\ldots\}} \left\|\log\left[\frac{d\mu_n}{dx}\right]\right\|_{L^\infty(dx)}}}{2^{-\left\|\frac{d\mu}{dx}\right\|_{L^\infty(dx)}}},$$

condition (7) implying then (14).

Remark IV.3 *For any* $\{\mu_n\}_{n=1}^{\infty} \subseteq \mathcal{AC}(\mathbb{X})$ *and* $\mu \in \mathcal{AC}_+(\mathbb{X})$, *condition (14) in Theorem IV.1 reads as*

$$\frac{d\mu_n}{dx}(x) \left[\frac{d\mu}{dx}(x)\right]^{-1} \leq M < \infty \tag{22}$$

for each $n \in \{1, 2, \ldots\}$ *and Lebesgue-almost every* $x \in \mathbb{X}$, *and therefore, as the reader can easily verify (note* $M \geq 1$ *necessarily), we have*

$$\frac{d\mu_n}{dx}(x) \log\left[\frac{d\mu_n}{dx}(x)\right] \leq M \max\left\{\psi_1(x), \psi_2(x)\right\},$$

for each $n \in \{1, 2, \ldots\}$ *and Lebesgue-almost every* $x \in \mathbb{X}$ *as well, where*

$$\psi_1(x) \doteq \frac{d\mu}{dx}(x) \log[M] \quad \text{and} \quad \psi_2(x) \doteq \frac{d\mu}{dx}(x) \log[M] + \frac{d\mu}{dx}(x) \log\left[\frac{d\mu}{dx}(x)\right].$$

Thus, since also[2] $y \log[y] \geq (-e \ln[2])^{-1}$ *for all* $y \in \mathbb{R}_+$, *condition (22) then implies the existence of* $C_0, C_1, C_2 \in \mathbb{R}_+$, *with* $C_0 > 0$ *necessarily if* $\mathbb{X}$ *has infinite Lebesgue measure (easy to check), such that for each* $n \in \{1, 2, \ldots\}$

$$\left|\frac{d\mu_n}{dx}(x) \log\left[\frac{d\mu_n}{dx}(x)\right]\right| \leq f(x), \tag{23}$$

for Lebesgue-almost every $x \in \mathbb{X}$, *where*

$$f(x) \doteq C_0 + C_1 \frac{d\mu}{dx}(x) + C_2 \frac{d\mu}{dx}(x) \left|\log\left[\frac{d\mu}{dx}(x)\right]\right|.$$

However, even with $\{\mu_n\}_{n=1}^{\infty} \subseteq \mathbb{H}(\mathbb{X})$, $\mu \in \mathbb{H}(\mathbb{X}) \cap \mathcal{AC}_+(\mathbb{X})$ *and* $\{\frac{d\mu_n}{dx}\}_{n=1}^{\infty}$ *converging pointwise Lebesgue-almost everywhere to* $\frac{d\mu}{dx}$ *on* $\mathbb{X}$ *as* $n \uparrow \infty$ *(see Remark IV.1), condition (23) cannot be used in the dominated convergence theorem to conclude the convergence*

$$\mathcal{H}[\mu_n] = -\int_{\mathbb{X}} \log\left[\frac{d\mu_n}{dx}\right] d\mu_n = -\int_{\mathbb{X}} \log\left[\frac{d\mu_n}{dx}\right] \frac{d\mu_n}{dx} dx \to -\int_{\mathbb{X}} \log\left[\frac{d\mu}{dx}\right] \frac{d\mu}{dx} dx = -\int_{\mathbb{X}} \log\left[\frac{d\mu}{dx}\right] d\mu = \mathcal{H}[\mu],$$

as $n \uparrow \infty$, *if* $\mathbb{X}$ *has infinite Lebesgue measure. Indeed,*

$$\int_{\mathbb{X}} f dx \geq \int_{\mathbb{X}} C_0 dx = C_0 \int_{\mathbb{X}} dx = \infty$$

in that case. Therefore the advantage of considering integrals w.r.t. $d\mu$ *(instead of* dx*) in the arguments leading to the proof of Theorem IV.1.*

V. DISCRETE ALPHABET SOURCES

In this section we consider discrete alphabet sources. We show how all convergence results become straightforward for finitely supported probability measures, and we also provide results for the infinitely supported case, by exploiting the equivalence between weak convergence and convergence in variation in this setting.

---

[2] Recall that $0 \log[0] = 0[-\infty] = 0$ by convention.

Though most of the definitions in the previous sections include the discrete case as a particular case when no reference to $\mathcal{AC}(\mathbb{X})$ is made, we will go briefly through the relevant concepts that apply for the case of probability measures with discrete support prior to state the results.

Throughout this section we consider

$$\mathbb{X} \doteq \{x_i\}_{i\in\mathcal{I}} \subseteq \mathbb{R}^k \text{ with } \mathcal{I} \subseteq \{1, 2, \ldots\}.$$

Accordingly, $\mathcal{S}(\mathbb{X})$ denotes the collection of all subsets of $\mathbb{X}$ and $\mathcal{P}(\mathbb{X})$ the collection of all probability measures on $(\mathbb{X}, \mathcal{S}(\mathbb{X}))$. A measure $\mu \in \mathcal{P}(\mathbb{X})$ is now characterized by the sequence $\{p_i^\mu\}_{i\in\mathcal{I}} \subseteq [0, 1]$, satisfying the normalization condition

$$\sum_{i\in\mathcal{I}} p_i^\mu = 1,$$

and given by $p_i^\mu \doteq \mu(\{x_i\})$, $i \in \mathcal{I}$. We will associate the sequence $\{a_i\}_{i\in\mathcal{I}} \subseteq \mathbb{R}$ to the mapping $a : \mathbb{X} \to \mathbb{R}$ by defining $a(x_i) \doteq a_i$ for each $i \in \mathcal{I}$. We shall use the same notation as in the previous sections to denote now

$$\mathbb{H}(\mathbb{X}) \doteq \left\{\mu \in \mathcal{P}(\mathbb{X}) : \{\log[p_i^\mu]\}_{i\in\mathcal{I}} \in l^1(\mu)\right\}, \text{ where } l^1(\mu) \doteq \left\{\{a_i\}_{i\in\mathcal{I}} \subseteq \mathbb{R} : \sum_{i\in\mathcal{I}} |a_i| p_i^\mu < \infty\right\}$$

for $\mu \in \mathcal{P}(\mathbb{X})$. Note that $\{a_i\}_{i\in\mathcal{I}} \in l^1(\mu)$ if and only if $a \in L^1(d\mu)$. In fact,

$$\sum_{i\in\mathcal{I}} a_i p_i^\mu = \int_{\mathbb{X}} a d\mu \text{ and } \|\{a_i\}_{i\in\mathcal{I}}\|_{l^1(\mu)} \doteq \sum_{i\in\mathcal{I}} |a_i|\, p_i^\mu = \int_{\mathbb{X}} |a|\, d\mu = \|a\|_{L^1(d\mu)}$$

for $\{a_i\}_{i\in\mathcal{I}} \in l^1(\mu)$ or, equivalently, for $a \in L^1(d\mu)$.

Consider the measure $\mu \in \mathbb{H}(\mathbb{X})$. Then,

$$\mathcal{H}[\mu] \doteq -\sum_{i\in\mathcal{I}} p_i^\mu \log[p_i^\mu] \in \mathbb{R}_+$$

is the Shannon Entropy of μ. Also, given $\mu \in \mathcal{P}(\mathbb{X})$, $\mathcal{AC}(\mathbb{X}||\mu)$ denotes the set of all probability measures $\sigma \in \mathcal{P}(\mathbb{X})$ that are absolutely continuous w.r.t. μ, i.e., satisfying the condition $p_i^\sigma = 0$ whenever $p_i^\mu = 0$. In the same way,

$$\mathbb{H}(\mathbb{X}||\mu) \doteq \left\{\sigma \in \mathcal{AC}(\mathbb{X}||\mu) : \left\{\log\left[\frac{p_i^\sigma}{p_i^\mu}\right]\right\}_{i\in\mathcal{I}} \in l^1(\sigma)\right\},$$

with the standard convention $0 \log[\frac{0}{0}] = 0$ (motivated by continuity). Finally, if $\sigma \in \mathbb{H}(\mathbb{X}||\mu)$, then

$$\mathcal{D}[\sigma||\mu] \doteq \sum_{i\in\mathcal{I}} p_i^\sigma \log\left[\frac{p_i^\sigma}{p_i^\mu}\right] \in \mathbb{R}_+$$

is the Shannon Relative Entropy between σ and μ, or equivalently the Kullback-Liebler Discriminant between σ and μ too. Similarly than before, $\mathcal{D}[\sigma||\mu] = 0$ if and only if $p^\mu = p^\sigma$ σ-almost everywhere, i.e., if and only if $p_i^\mu = p_i^\sigma$ for each $i \in \mathcal{I}$ such that $p_i^\sigma > 0$.

Weak convergence of $\{\mu_n\}_{n=1}^\infty \subseteq \mathcal{P}(\mathbb{X})$ to $\mu \in \mathcal{P}(\mathbb{X})$ is now characterized as follows. We have $\mu_n \Rightarrow \mu$ as $n \uparrow \infty$ if and only if

$$\sum_{i\in\mathcal{I}} f(x_i) p_i^{\mu_n} \to \sum_{i\in\mathcal{I}} f(x_i) p_i^\mu, \tag{24}$$

as $n \uparrow \infty$ as well, for each bounded, real-valued function f on $\mathbb{X}$.

Distance in variation between $\sigma_1 \in \mathcal{P}(\mathbb{X})$ and $\sigma_2 \in \mathcal{P}(\mathbb{X})$ is

$$\|\sigma_1 - \sigma_2\|_V = \left\|\{p_i^{\sigma_1} - p_i^{\sigma_2}\}_{i\in\mathcal{I}}\right\|_{l^1(\delta)},$$

where δ denotes the counting measure on $(\mathbb{X}, \mathcal{S}(\mathbb{X}))$, i.e., $\delta(\{x_i\}) \doteq 1$ for each $i \in \mathcal{I}$ and $\delta(A) \doteq \sum_{x_i\in A} \delta(\{x_i\})$ for each $A \in \mathcal{S}(\mathbb{X})$. The corresponding convergence in variation of $\{\mu_n\}_{n=1}^\infty \subseteq \mathcal{P}(\mathbb{X})$ to $\mu \in \mathcal{P}(\mathbb{X})$, $\|\mu_n - \mu\|_V \to 0$ as $n \uparrow \infty$, takes place if and only if, as $n \uparrow \infty$ as well,

$$\sum_{i\in\mathcal{I}} |p_i^{\mu_n} - p_i^\mu| \to 0, \text{ since } \|\{p_i^{\mu_n} - p_i^\mu\}_{i\in\mathcal{I}}\|_{l^1(\delta)} = \sum_{i\in\mathcal{I}} |p_i^{\mu_n} - p_i^\mu|.$$

In the discrete setting, the relationship between weak convergence and convergence in variation in Lemma II.2 can be strengthened, as stated in the following result.

Lemma V.1 *Let $\{\mu_n\}_{n=1}^{\infty} \subseteq \mathcal{P}(\mathbb{X})$ and $\mu \in \mathcal{P}(\mathbb{X})$. Then, we have $\mu_n \Rightarrow \mu$ as $n \uparrow \infty$ if and only if $\|\mu_n - \mu\|_V \to 0$, as $n \uparrow \infty$ as well. Moreover, the previous ways of convergence take place if and only if $p_i^{\mu_n} \to p_i^{\mu}$ as $n \uparrow \infty$ for each $i \in \mathcal{I}$, i.e., both the topology of weak convergence and convergence in variation are equivalent to the topology of coordinatewise convergence of the sequence of vectors $\{(p_i^{\mu_n})_{i\in\mathcal{I}}\}_{n=1}^{\infty}$ to the vector $(p_i^{\mu})_{i\in\mathcal{I}}$ as $n \uparrow \infty$ (equivalently, to the topology of pointwise convergence of $\{p^{\mu_n}\}_{n=1}^{\infty}$ to p^{μ} on $\mathbb{X}$ as $n \uparrow \infty$).*

Proof. We obviously have that $\mu_n \Rightarrow \mu$ as $n \uparrow \infty$ implies $p_i^{\mu_n} \to p_i^{\mu}$, as $n \uparrow \infty$ as well, for each $i \in \mathcal{I}$ (just need to consider equation (24) with $f_i : \mathbb{X} \to \{0, 1\}$ defined, for each $i \in \mathcal{I}$, by letting $f_i(x) \doteq 1$ if $x = x_i$ and $f_i(x) \doteq 0$ if $x \in \mathbb{X} \setminus \{x_i\}$). Now, since

$$\|\mu_n - \mu\|_V = \left\|\{p_i^{\mu_n} - p_i^{\mu}\}_{i\in\mathcal{I}}\right\|_{l^1(\delta)} = \|p^{\mu_n} - p^{\mu}\|_{L^1(d\delta)} = \int_{\mathbb{X}} |p^{\mu_n} - p^{\mu}| d\delta,$$

if the sequence $\{p^{\mu_n}\}_{n=1}^{\infty}$ converges pointwise to p^{μ} on $\mathbb{X}$ as $n \uparrow \infty$, then Scheffé's Lemma gives us the convergence $\|\mu_n - \mu\|_V \to 0$, as $n \uparrow \infty$ too, the same as in the differential case (see Remark IV.2). The result then follows from Lemma II.2. ■

Remark V.1 *In the differential setting and from Remark IV.2 and Lemma II.2, we have the chain of implications: pointwise convergence of densities (Lebesgue-almost everywhere pointwise convergence in fact) ⇒ convergence in variation ⇒ weak convergence. As Lemma V.1 shows, the corresponding three ways of convergence in the discrete setting are indeed equivalent.*

In view of Lemma V.1, it is a straightforward exercise to check that in the case when the set $\mathbb{X}$ (equivalently the index set $\mathcal{I}$) can be taken to be finite (i.e., when the supports of all probability measures involved are contained in a finite set), the convergence $\mu_n \Rightarrow \mu$ as $n \uparrow \infty$ implies both

$$\mathcal{H}[\mu_n] \to \mathcal{H}[\mu] \text{ and } \mathcal{D}[\mu_n||\mu] \to 0,$$

as $n \uparrow \infty$ as well, being in fact, from Remark III.2, $\mu_n \Rightarrow \mu$ as $n \uparrow \infty$ and $\mathcal{D}[\mu_n||\mu] \to 0$ as $n \uparrow \infty$ equivalent (of course with $\{\mu_n\}_{n=1}^{\infty} \subseteq \mathcal{AC}(\mathbb{X}||\mu)$ for the convergence $\mathcal{D}[\mu_n||\mu] \to 0$ as $n \uparrow \infty$).

Given $\mu \in \mathcal{P}(\mathbb{X})$ and $\{\mu_n\}_{n=1}^{\infty} \subseteq \mathcal{P}(\mathbb{X})$, the set $\mathbb{X}$ can be made into a finite set whenever μ is finitely supported and $\{\mu_n\}_{n=1}^{\infty} \subseteq \mathcal{AC}(\mathbb{X}||\mu)$, by just redefining it as $\mathbb{X}_{\mu}$ with

$$\mathbb{X}_{\mu} \doteq \text{support}\,(p^{\mu}) = \{x \in \mathbb{X} : p^{\mu}(x) > 0\} = \{x_i\}_{i\in\mathcal{I}_{\mu}}$$

and $\mathcal{I}_{\mu} \doteq \{i \in \mathcal{I} : p_i^{\mu} > 0\}$. Note that if $\mu \in \mathcal{P}(\mathbb{X})$ is finitely supported and $\{\mu_n\}_{n=1}^{\infty} \subseteq \mathcal{AC}(\mathbb{X}||\mu)$, then $\mu \in \mathbb{H}(\mathbb{X})$ and $\{\mu_n\}_{n=1}^{\infty} \subseteq \mathbb{H}(\mathbb{X}||\mu)$. The discrete setting versions of Theorems III.1 and III.2 and Corollary III.1 are trivial in that case. They cannot be stated for $\mu \in \mathcal{P}(\mathbb{X})$ being infinitely supported however, as clear from the following remark.

Remark V.2 *Unlike in the differential setting (see Remark III.3), in the discrete setting we have for $\mu \in \mathcal{P}(\mathbb{X})$ that $p_i^{\mu} \to 0$ as $i \uparrow \infty$, $i \in \mathcal{I}_{\mu}$, whenever $\mathcal{I}_{\mu}$ (equivalently $\mathbb{X}_{\mu}$) is infinite ($\sum_{i\in\mathcal{I}_{\mu}} p_i^{\mu} = 1 < \infty$), and therefore the subsequence $\{\log[p_i^{\mu}]\}_{i\in\mathcal{I}_{\mu}}$ cannot be bounded in that case (even when $\mathbb{X}_{\mu}$ is a bounded subset of $\mathbb{R}^k$).*

We consider the general case, covering infinitely supported probability measures, in the following theorem (which corresponds to the discrete version of Theorem IV.1) and two corresponding corollaries. The proof of the theorem will be omitted, following by similar corresponding arguments as those in the proof of Theorem IV.1.

Theorem V.1 *Let $\mu \in \mathbb{H}(\mathbb{X})$ and $\{\mu_n\}_{n=1}^{\infty} \subseteq \mathcal{AC}(\mathbb{X}||\mu)$ be such that $p_i^{\mu_n} \to p_i^{\mu}$ as $n \uparrow \infty$, for each $i \in \mathcal{I}_{\mu>0}$, and*

$$M \doteq \sup_{\substack{n\in\{1,2,\ldots\} \\ i\in\mathcal{I}_{\mu>0}}} \frac{p_i^{\mu_n}}{p_i^{\mu}} < \infty.$$

Then, $\{\mu_n\}_{n=1}^{\infty} \subseteq \mathbb{H}(\mathbb{X}) \cap \mathbb{H}(\mathbb{X}||\mu)$ and we have both

$$\mathcal{D}[\mu_n||\mu] \to 0 \text{ and } \mathcal{H}[\mu_n] \to \mathcal{H}[\mu]$$

as $n \uparrow \infty$.

We have the following two direct corollaries to Theorem V.1 (see also Lemma V.1 and Remark III.2). For the first, we define $[\mathbb{H} \cap \mathcal{P}_+](\mathbb{X}) \doteq \mathbb{H}(\mathbb{X}) \cap \mathcal{P}_+(\mathbb{X})$ with $\mathcal{P}_+(\mathbb{X})$ the collection of all $\mu \in \mathcal{P}(\mathbb{X})$ satisfying $p_i^{\mu} > 0$ for each $i \in \mathcal{I}$.

Corollary V.1 *Let $\mu \in [\mathbb{H} \cap \mathcal{P}_+](\mathbb{X})$ and $\{\mu_n\}_{n=1}^{\infty} \subseteq \mathcal{P}(\mathbb{X})$ be such that $\mu_n \Rightarrow \mu$ as $n \uparrow \infty$ and*

$$\sup_{\substack{n\in\{1,2,\ldots\} \\ i\in\mathcal{I}}} \frac{p_i^{\mu_n}}{p_i^{\mu}} < \infty.$$

Then, $\{\mu_n\}_{n=1}^{\infty} \subseteq \mathbb{H}(\mathbb{X}) \cap \mathbb{H}(\mathbb{X}||\mu)$ and we have both

$$\mathcal{D}[\mu_n||\mu] \to 0 \text{ and } \mathcal{H}[\mu_n] \to \mathcal{H}[\mu]$$

as $n \uparrow \infty$.

Corollary V.2 *Let $\mu \in \mathbb{H}(\mathbb{X})$ and $\{\mu_n\}_{n=1}^{\infty} \subseteq \mathcal{AC}(\mathbb{X}||\mu)$ be such that*

$$\sup_{\substack{n\in\{1,2,\ldots\} \\ i\in\mathcal{I}_{\mu>0}}} \frac{p_i^{\mu_n}}{p_i^{\mu}} < \infty. \tag{25}$$

Then, $\{\mu_n\}_{n=1}^{\infty} \subseteq \mathbb{H}(\mathbb{X}) \cap \mathbb{H}(\mathbb{X}||\mu)$ and, if $\mathcal{D}[\mu_n||\mu] \to 0$ as $n \uparrow \infty$, we have

$$\mathcal{H}[\mu_n] \to \mathcal{H}[\mu],$$

as $n \uparrow \infty$ as well.

Remark V.3 *In the context of continuity versus pure convergence properties of Shannon entropy discussed in Section I, note Corollary V.2 establishes the convergence $\mathcal{H}[\mu_n] \to \mathcal{H}[\mu]$ as $n \uparrow \infty$, under the convergence $\mathcal{D}[\mu_n||\mu] \to 0$ as $n \uparrow \infty$ as well, by exploiting an underlying structure relating $\{\mu_n\}_{n=1}^{\infty}$ to μ (condition (25)). In contrast, by imposing the stronger requirement on μ of being power dominated (stronger than just $\mu \in \mathbb{H}(\mathbb{X})$; see [10] for the definition of a power dominated distribution), the continuity result [10, Theorem 21, p.16] establishes the corresponding entropy convergence, in a discrete setting too, for all approximating sequences converging in the above Kullback-Liebler discriminant sense.*

Acknowledgments

The authors would like to thank the anonymous referee for helpful suggestions which helped to improve the paper.

[1] Antos, A. and Kontoyiannis, I. (2002). Convergence properties of functional estimates of discrete distributions. *Random Structures and Algorithms* **19**, 3-4, 163–193.
[2] Barron, A. R. (1985). The strong ergodic theorem for densities: Generalized Shannon-McMillan-Breiman theorem. *The Annals of Probability* **13**, 4, 1292–1303.
[3] Barron, A. R. (1986). Entropy and the central limit theorem. *The Annals of Probability* **14**, 1, 336–342.
[4] Billingsley, P. (1999). *Convergence of Probability Measures*, Second ed. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., New York.
[5] Blahut, R. E. (1987). *Principles of Information Theory.* Addison-Wesley, New York.
[6] Cover, T. and Thomas, J. (1991). *Elements of Information Theory.* Wiley Series in Telecommunications. John Wiley & Sons, Inc., New York.
[7] Dobrushin, R. L. (1960). Passage to the limit under the information and entropy signs. *Theory Prob. Applications* **V**, 1, 25–32.
[8] Fristedt, B. and Gray, L. (1997). *A Modern Approach to Probability Theory.* Probability and its Applications. Birkhäuser, Boston.
[9] Hall, P. and Morton, S. C. (1993). On the estimation of entropy. *Ann. Inst. Statist. Math.* **45**, 1, 69–88.
[10] Harremoës, P. (2007). Information topologies with applications. *Entropy, Search, Complexity (Springer-Verlag) 16*, 113–150.
[11] Ho, S.-W. and Yeung, R. W. (2005). On the discontinuity of the Shannon information measures. In *Proc. IEEE International Symposyum on Information Theory (ISIT 2005).* Adelaide, Australia, 159–163.
[12] Kallenberg, O. (2002). *Foundations of Modern Probability*, Second ed. Probability and its Applications. Springer-Verlag, New York.
[13] Royden, H. L. (1991). *Real Analysis*, Third ed. Prentice Hall, New Jersey.

[14] Shiryaev, A. N. (1986). *Probability*, Second ed. Number 95 in Graduate Texts in Mathematics. Springer-Verlag, New York.
[15] Wieczorkowski, R. and Grzegorzewski, P. (1999). Entropy estimators: Improvements and comparisons. *Commun. Stat., Simul. Comput.* **28**, 2, 541–567.
[16] Williams, D. (1991). *Probability with Martingales*. Cambridge University Press, Cambridge.
[17] Wyner, A. and Foster, D. (2003). On the lower limits of entropy estimation. *Submitted to IEEE Trans. Inf. Theory*.