

Entropy and Mutual Information of General Random Variables

Radha Krishna Ganti

November 15, 2005

Abstract

Entropy and mutual information of random variables on general spaces are defined.

1 Introduction

The differential entropy of a random variable X with probability density function $f(x)$ is given by $\int f(x) \log f(x) dx$. This definition assumes that the random variable is real and the probability density function exists. But in practice the entropy can be defined for a more general class of random variables.

The report is organized as follows. In Section 2 basic properties of measure spaces are discussed, the definitions of entropy and mutual information are given in Section 3.

2 Basic Measure and Probability Theory

Let X denote a non empty set. A sigma algebra \mathcal{M} on X is a algebra which is closed under countable-unions and complements.

Definition The sigma algebra [3, p.21] generated by $\mathcal{E} \subset \mathcal{P}(X)$ is the smallest sigma algebra containing \mathcal{E}

If X is any metric space, or more generally a topological space

Definition The Borel sigma algebra of X is the algebra generated by the set of all open sets [1] of X , and is denoted by \mathcal{B}_X

For metric spaces X_i which are separable [1, p.48], the Borel algebra on the product space $\prod_i^N X_i$ (N finite) $\mathcal{B}_{\prod_i^N X_i} = \prod_{i=1}^N \mathcal{B}_{X_i}$ [3, p.23].

Let X be equipped with a sigma algebra \mathcal{M} . A measure on \mathcal{M} is a function from $p : \mathcal{M} \rightarrow [0, \infty]$ such that

1. $p(\emptyset) = 0$
2. if $\{E_i\}_1^\infty$ are disjoint sequence of sets in \mathcal{M} , then $p(\cup_1^\infty E_i) = \sum_1^\infty p(E_i)$

The triple (X, \mathcal{M}, p) is called as a measure space.

Definition If $X = \cup_1^\infty E_j$ where $E_j \in \mathcal{M}$ and $p(E_j) < \infty$, then p is called as sigma finite measure. p is called as finite if $p(X) < \infty$

Generally sigma finite measures exhibit good properties.

Examples

1. Let X be a countable set, and the sigma algebra be the power set of X , and let $f : X \rightarrow [0, \infty]$. Then $p(E) = \sum_{x \in E} f(x)$, defines a measure. If $f(x) = 1$ for all X , then it is called as **counting measure**. This is also a sigma finite measure.
2. Let X be the real line and the corresponding sigma algebra be the Borel sigma algebra on the real line.

The following table indicates the similarities and differences between general measure and probability [3, p.314].

Analysts' Term	Probabilists' Term
Measure Space (X, \mathcal{M}, μ)	Sample space (Ω, \mathcal{B}, P)
sigma algebra	sigma field
Measurable set	Event
Measurable real valued function f	Random variable X
Integral of f , $\int f d\mu$	Expectation or mean of X , $E(X)$
L^p	Having finite p th moment
Convergence in measure	Convergence in probability
Almost everywhere a.e.	Almost surely
Borel probability measure on \mathbb{R}	Distribution
Fourier transform of a measure	Characteristic function of a distribution

Observe that probability measure on any space is a finite measure and hence sigma finite. Discrete probability spaces are finite spaces equipped with appropriate counting measure (or $f(x)$ of the example defined appropriately). Let (X, \mathcal{M}) and (Y, \mathcal{N}) , be two probability spaces, and $f : X \rightarrow Y$. f is said to be $(\mathcal{M}, \mathcal{N})$ measurable, if for every $E \in \mathcal{N}$, $f^{-1}(E) \in \mathcal{M}$. In probability f is called as a random variable. If the range of f is finite, f is called as a discrete random variable.

Definition A function ϕ is said to be simple, if the range of the function is finite. It can be represented as $\phi = \sum_{i=1}^N a_i \chi_{E_i}$, where $E_i = \phi^{-1}(a_i)$, and E_i measurable..

Also every positive function can be approximated by a sequence of simple functions, i.e $\forall f > 0$, $\exists \phi_n$ simple and $\phi_n \leq f$ such that $\phi_n \rightarrow f$ uniformly. The integral of the simple function $\int \phi = \sum_{j=1}^N a_j P(E_j)$. This definition is intuitive.

Definition Let (X, \mathcal{M}, P) , be a measure space, and $f : X \rightarrow [0, \infty]$ and be measurable. Then the integral of f with respect to measure P , denoted by $\int f dP$, $\int f(x) dP(x)$, $\int f$, is defined as

$$\int f = \sup\left\{ \int \phi, 0 \leq \phi \leq f, \phi \text{ simple} \right\} \quad (1)$$

i.e, we approximate the positive function by simple functions and find the area under the simple function which approximate it the best. For a general function $\int f = \int f^+ - \int f^-$. The three important theorems for the integral defined above are the **Monotone Convergence Theorem**, **Fubini's Theorem**, **Dominated Convergence Theorem** [3, p. 50,52,54]. These theorems deal with the equality of limit of a sequence of integrals and integrals of limits of sequence of functions. Also if P is a measure and $f \geq 0$, then $Q(E) = \int_E f dP$ is a measure on the same space, i.e, every random variable induces a measure on the range space. For example, consider the Lebesgue measure m on $[0, 1]$, i.e, the measure space $([0, 1], \mathcal{B}_{[0,1]}, m)$, and let $f = 1\chi_{[0,1/3]} + 2\chi_{[1/3,2/3]} + 3\chi_{[2/3,1]}$. This gives a discrete random variable with distribution $P([1, 2, 3]) = [1/3, 1/3, 1/3]$.

3 Defintion of Entropy and Mutual Information

Let P and Q be two probability measures on same space and algebra (X, \mathcal{M}) .

Definition P is said to be **absolutely continuous** with respect to Q and denoted by $P \ll Q$, if $P(E) = 0$ for every $E \in \mathcal{M}$ for which $Q(E) = 0$.

Example of absolutely continuous measures: Let (X, \mathcal{M}, μ_1) be a measure space. Let $f : X \rightarrow [0, \infty]$ be a measurable function and $\int |f| < \infty$. Then the measure $\mu_2(E) = \int_E f d\mu_1$ is absolutely

continuous with respect to μ_1 .

Example (trivial) of not absolutely continuous measures: Consider the measure space $([0, 2\pi], \mathcal{B}_{[0, \pi]}, m)$, where m is the Lebesgue measure. consider $f(x) = \sin(x)$. Define $m_1(E) = \int_E f \cdot \chi_{[0, \pi]} dx$, $m_2(E) = \int_E f \cdot \chi_{[\pi, 2\pi]} dx$. m_1, m_2 are measures (from previous section). These are clearly **not absolutely continuous**.

Theorem 3.1. The Lebesgue-Radon-Nikodym Theorem: [3, p. 91] *Let P and Q be two probability measures on (X, \mathcal{M}) . If $P \ll Q$ then $P = \int f dQ$, where $f \geq 0$ and real valued and f is Q measurable (Any two such functions are almost equal every where).*

f is called the **Radon-Nikodym derivative** and is represented as $f = dP/dQ$.

Definition Divergence (Kullback Leibler distance) of [8, 5] of a discrete (finite) random variable X with distribution $P(x)$ with respect to another random variable Y with distribution $Q(y)$ on the same measure space is given by

$$D(P \parallel Q) = \sum_{i=1}^N p(x) \log \frac{p(x)}{q(x)} \quad (2)$$

It is of general knowledge that the discrete entropy is independent of the values that the random variables take.

Definition Let (X, \mathcal{M}, P) , be a measure space and $C_1 \dots C_N$, $C_i \in \mathcal{M}$ are called a measurable partition of X if $\cup_i^N C_i = X$

Definition The **divergence** [2] of P and Q defined on the same σ -algebra of a space X is given by

$$D(P \parallel Q) = \sup \left\{ \sum_{i=1}^{N(\eta)} p(C_i) \log \frac{p(C_i)}{q(C_i)}; \text{ where, } \eta = [C_1, \dots C_N(\eta)] \text{ is a measurable finite partition of } X \right\} \quad (3)$$

Theorem 3.2. [6, 7, 2] *Given two probability measures P and Q on a common measurable space (X, \mathcal{M}) , if P is not absolutely continuous with respect to Q , then*

$$D(P \parallel Q) = \infty$$

If $P \ll Q$, then the Radon-Nikodym derivative $f = dP/dQ$ exists and

$$D(P \parallel Q) = \int \log f dP = \int f \log f dQ$$

The quantity $\log f$ is called the entropy density or relative entropy density of P with respect to Q .

Proof. The proof is lengthy and the reader is referred to [6, p.80], [7, p.23] □

If P and Q are discrete measures, with pmf's p and q , then the Radon Nikodym derivative $dP/dQ(x) = p(x)/q(x)$, and we get the formula for discrete case. Also if P and Q are probability measures on \mathbf{R}^n . If P and Q are absolutely continuous with respect to the lebegue measure on \mathbf{R}^n , then there exists densities f and g also called as the probability density function and

$$D(P \parallel Q) = \int_{\mathbf{R}^n} f(x) \log \frac{f(x)}{g(x)} dP \quad (4)$$

Let X and Y be two random variables on (A_X, \mathcal{B}_{A_X}) and (A_Y, \mathcal{B}_{A_Y}) . Let P_{XY} and M_{XY} be two distributions on $(A_X \times A_Y, \mathcal{B}_{A_X \times A_Y})$, and $P_X(\cdot) = P_{XY}(\cdot \times A_Y)$, $P_Y(\cdot) = P_{XY}(A_X \times \cdot)$, be the marginal distributions. Let $\mathcal{M} = \{M_{XY}; M_{XY} = M_X \times M_Y\}$, i.e, the set of all product distributions on XY . Then

$$\inf_{M_{XY} \in \mathcal{M}} D(P_{XY} \parallel M_{XY}) = D(P_{XY} \parallel P_X \times P_Y) \quad (5)$$

i.e, $P_X \times P_Y$ is the best product approximation of P_{XY} yielding the minimum divergence. This motivates calling $P_X \times P_Y$ the *independent (memoryless) approximation* to P_{XY}

Mutual Information: Let X and Y be two random variables, the mutual information between them is given by

$$I(X; Y) = D(P_{XY} \parallel P_X \times P_Y) = \int \log \left(\frac{d(P_{XY})}{d(P_X \times P_Y)} \right) d(P_{XY}) \quad (6)$$

Entropy:

$$H(X) = I(X; X) \quad (7)$$

3.1 Properties of $D(P \parallel Q)$, $I(P; Q)$

Let P and Q , with $P \ll Q$ be two probability distributions on the same space (A, \mathcal{M}) . Let X and Y be two random variables on alphabet \mathcal{X} and \mathcal{Y} and the mapping $x \rightarrow V(\cdot|x)$ is measurable for every x and a probability density on Y ($V(\cdot|x)$ represents the channel). The properties are stated here and the reference to the appropriate proofs are given as references.

1. $D(P \parallel Q) \geq 0$, with equality iff $P = M$ [6, p.79].
2. $D(P \parallel Q)$ is convex with respect to (P, Q) [4, Theorem. 3.1].
3. $I(X; Y) \geq 0$, with equality iff X and Y are independent [6].

4. If f is a measurable function of X and g is a measurable function of Y , then [6]

$$I(f(X);g(Y)) \leq I(X;Y)$$

5. If X or Y has a **finite** alphabet, then [6]

$$I(X;Y) = H(X) - H(X|Y)$$

In general this may not be true because when atleast one of the RV is not finite, then we may end up with indeterminate forms like $\infty - \infty$.

6. If $P_{XY}(A \times B) = \int_A V(B|x)P_X(dx)$, and $Q_Y(B) = \int V(B|x)P_X(dx)$,and

$$I(P;V) = D(P_{XY} \parallel P_X \times Q_Y) \tag{8}$$

then $I(P;V)$ is a concave function of P and a convex function of V . If \mathcal{X} is finite then $I(P;V)$ is continous in the distribution P [2]. In the more general case when \mathcal{X} is not finite, the continuity of $I(P;V)$ holds under more restrictive conditions [4, Theorem 3.2].

4 Conclusion

In this report, the definitions of divergence,entropy and mutual information have been presented for general random variables. This is a very introductive report and an indepth analysis of these quantities can be found in [6]. Most of the work in information theory is restricted to the finite alphabet case. Although this is very much sufficient for practical purposes, it is important and nice to understand that the quantites of entropy and information can be defined in a very general sense, and most of the results that hold in the discrete setting can be shown to be true in a more general scenario with a little effort.

References

- [1] S.V. Fomin A.N.Kolmogorov. *Introductory Real Analysis*. Dover, 1970.
- [2] Imre Csiszar. Arbitrarily varying channels with general alphabets and states. *IEEE Trans. on Information Theory*, 38:1725, 1992.
- [3] Gerald B. Folland. *Real Analysis, Modern Techniques and Their Applications*. Wiley, 1999.

- [4] Majid Fozunbal, Steven W. McLaughlin, and Roland W. Schafer. Capacity analysis for continuous-alphabet channels with side information, part 1: A general framework. *IEEE Transactions on Information Theory*, (9):3075, 2005.
- [5] Robert G. Gallager. *Information Theory and Reliable Communication*. Wiley, 1968.
- [6] Robert M. Gray. *Entropy and Information Theory*. <http://www-ee.stanford.edu/~gray/it.pdf>.
- [7] M. S. Pinsker. *Information and Information Stability of Random Variables and Processes*.
- [8] Joy Thomas and Thomas Cover. *Elements of Information Theory*. Wiley-Interscience, 1991.