On the Entropy of a Mixture Distribution

by

Dimitris N. Politis
Purdue University

Technical Report #91-67

Department of Statistics
Purdue University

November 1991

On the entropy of a mixture distribution

Dimitris N. Politis

Department of Statistics, Purdue University
West Lafayette, IN 47907-1399

Abstract

Shannon's entropy is usually defined separately for discrete, and for (absolutely) continuous random variables. In this article, a simple expression for the entropy of random variables with mixed (discrete-continuous) distributions is given in terms of the usual entropy definitions. In addition, the Maximum Entropy problem in the setting of mixture distributions is discussed.

Index Terms. Entropy, Maximum Entropy, mixture distributions.

I. Introduction and definitions

Let X be a random variable (r.v.) with absolutely continuous distribution with respect to Lebesgue measure μ on the real line, and let $f_c(x)$ be its probability density function (p.d.f.), and $S_c \subset \mathbf{R}$ its support set. Let $S_d \subset \mathbf{R}$ be a finite (or at most countable) set, and let Y be a discrete r.v. with $f_d(x)$ as its probability mass function (p.m.f.), and S_d as its support set.

The Shannon entropies of X and Y are defined by (cf. [8]) as

$$H(X) = -\int_{x \in S_c} f_c(x) \log f_c(x) d\mu(x) \tag{1}$$

$$H(Y) = -\sum_{x \in S_d} f_d(x) \log f_d(x) \tag{2}$$

It is apparent that the two definitions are special cases of the following general definition of entropy (cf. [6])

$$H(W) = -\int f_W(w) \log f_W(w) d\lambda(w) \tag{3}$$

where the r.v. W has a probability density $f_W(w)$ with respect to some σ-finite measure λ on the real line.

However, the only practical choices for λ are Lebesgue measure, a counting measure, or a linear combination of the two, that correspond to (absolutely) continuous, discrete, or mixed (discrete-continuous) observable r.v.'s respectively. Since the case of mixed r.v.'s includes the other two as special cases, we will focus on that and derive an expression for the corresponding entropy.

So suppose that the r.v. W is defined in the following way. Let the Bernoulli r.v. Z be independent of X and Y and such that $P(Z=1)=p$, and $P(Z=0)=1-p$. Since the p.d.f. $f_c(x)$ is uniquely defined only almost everywhere with respect to Lebesgue measure μ, we can assume without loss of generality that $f_c(x) = 0$, if $x \in S_d$, that is, that the supports S_d and S_c are disjoint. Now let

$$W = \begin{cases} X & \text{if } Z = 0 \\ Y & \text{if } Z = 1 \end{cases} \tag{4}$$

In this case, W is seen to have a probability density

$$f_W(w) = p f_d(w) + (1-p) f_c(w) \tag{5}$$

with respect to the measure ν on the real line, which assigns mass $\nu(A) = \int_A d\mu(x) + \#(A \cap S_d)$, to any Borel set A, where $\#(B)$ denotes the number of elements in set B (cf. [5]). The r.v. W is a mixed discrete-continuous r.v., and it becomes discrete or continuous in the extreme cases $p = 1$ or $p = 0$ respectively.

Using definition (3) and the fact that the supports S_d and S_c are assumed disjoint, it is immediate that the entropy of the general mixed r.v. W is given by

$$H(W) = -\sum_{x \in S_d} p f_d(x) \log(p f_d(x)) - \int_{x \in S_c} (1-p) f_c(x) \log((1-p) f_c(x)) d\mu(x) \tag{6}$$

or, after some simplifications, (and using the fact that $H(Z) = -p \log p - (1-p) \log(1-p)$),

$$H(W) = H(Z) + pH(Y) + (1-p)H(X) \tag{7}$$

It can be immediately verified that at the extreme cases where $p = 0$ or $p = 1$, expression (7) reduces to the standard expressions (1) and (2). In addition, the above expression of entropy can be compared to the notion of ϵ-entropy for mixed r.v.'s (cf. [7]), i.e. the rate distortion function relative to a squared-error fidelity criterion. In particular, the ϵ-entropy and the entropy $H(W)$ asymptotically (as $\epsilon \to 0$) agree except for a factor proportional to $(1-p) \log \epsilon$. The exact same relationship between the ϵ-entropy and the entropy is also observed in the case of (absolutely) continuous r.v.'s, for which $p = 0$ (cf. [1]).

As a further motivation for the entropy expression (7), consider the following example of Bayesian hypothesis testing [4]. Suppose that W is an observed r.v. with probability density $f_0(w)$ under hypothesis H_0, and $f_1(w)$ under hypothesis H_1. Both f_0 and f_1 densities are defined with respect to some measure λ on $\mathbf{R}$. Assume that hypothesis H_1 has a prior probability p of occuring, and H_0 has a prior probability $1 - p$. If one defines

$$Z = \begin{cases} 0 & \text{if } H_0 \text{ occured} \\ 1 & \text{if } H_1 \text{ occured} \end{cases}$$

then Z is the previously mentioned Bernoulli r.v., and the unconditional probability density of W with respect to λ is given by the mixture

$$f_W(w) = p f_1(w) + (1-p) f_0(w) \tag{8}$$

After $W = w$ is observed, the posterior probability of H_1 occuring is

$$P(H_1|W = w) = P(Z = 1|W = w) = \frac{pf_1(w)}{pf_1(w) + (1-p)f_0(w)} = \hat{p}(w)$$

and the posterior uncertainty regarding H_0 or H_1 occuring is

$$H(Z|W = w) = -\hat{p}(w)\log\hat{p}(w) - (1 - \hat{p}(w))\log(1 - \hat{p}(w))$$

The average uncertainty regarding H_0 or H_1 after observing the r.v. W is therefore calculated as

$$H(Z|W) = \int H(Z|W = w)f_W(w)d\lambda(w) = H(Z) + pH(W_1) + (1-p)H(W_0) - H(W) \quad (9)$$

where W_1 and W_0 are two r.v.'s with densities f_1 and f_0 respectively, and the entropies $H(Z), H(W_1), H(W_0)$, and $H(W)$, are calculated using the general definition (3).

Consider now the particular case where f_1 and f_0 have disjoint supports. It is apparent that in that case there is *no* uncertainty regarding H_0 or H_1 after observing $W = w$. Hence, $H(Z|W) = 0$ in equation (9), and we have

$$H(W) = H(Z) + pH(W_1) + (1-p)H(W_0) \quad (10)$$

In case now that $f_1 = f_d$, $f_0 = f_c$, and the measure λ is equal to the aforementioned measure ν, equation (10) provides a different derivation of the entropy expression (7).

II. Maximum Entropy for mixture distributions

The Maximum Entropy problem has been extensively discussed for discrete or continuous r.v.'s (cf. [2]). For example, it is easy to prove [3], that if Y is a discrete r.v. with uniform distribution on a set S_d consisting of m values, then Y has Maximum Entropy among all discrete r.v.'s taking values in S_d, and $H(Y) = \log m$. Similarly, if X is a mean-zero, variance σ^2, normal r.v., then X has Maximum Entropy among all (absolutely) continuous r.v.'s with variance σ^2, and $H(X) = \frac{1}{2}(\log(2\pi\sigma^2) + 1)$. As a last example, if U is an exponential r.v. with mean $\theta > 0$, then U has Maximum Entropy among all (absolutely) continuous positive r.v.'s with mean θ, and $H(U) = \log\theta + 1$.

We will now address the analogous Maximum Entropy problem for the r.v. W possessing the mixture density given in equation (8), where f_1 and f_0 are densities with respect to measure λ, with disjoint supports.

Theorem 1 *Let F_0 and F_1 be two classes of densities with respect to measure λ, such that for any $f_0 \in F_0$, $f_1 \in F_1$, the supports of f_0 and f_1 are disjoint. Also suppose that the Maximum Entropy problem is well-defined for classes F_0 and F_1, i.e. there is $f_0^\star \in F_0$ and $f_1^\star \in F_1$ such that $f_0^\star$ and $f_1^\star$ are the Maximum Entropy distributions in their respective class.*

Let the r.v. $W_0^\star$ have density $f_0^\star$, the r.v. $W_1^\star$ have density $f_1^\star$, and let the Bernoulli r.v. $Z^\star$ be independent of $W_0^\star$ and $W_1^\star$, and such that $P(Z^\star = 1) = p^\star$, and $P(Z^\star = 0) = 1 - p^\star$ with

$$p^\star = \frac{1}{1 + \exp\{H(W_0^\star) - H(W_1^\star)\}} \tag{11}$$

Then the mixture r.v. $W^\star$ defined by

$$W^\star = \begin{cases} W_0^\star & \text{if } Z^\star = 0 \\ W_1^\star & \text{if } Z^\star = 1 \end{cases} \tag{12}$$

has Maximum Entropy in the class spanned by all r.v.'s W that can be obtained by varying $f_0 \in F_0$, $f_1 \in F_1$, and $p \in [0,1]$, in equation (8).

Proof. Let any $f_0 \in F_0$, $f_1 \in F_1$, and $p \in [0,1]$. From equation (10), where W_1 and W_0 are two r.v.'s with densities f_1 and f_0 respectively, we have

$$H(W) = H(Z) + pH(W_1) + (1-p)H(W_0) \leq H(Z) + pH(W_1^\star) + (1-p)H(W_0^\star) \tag{13}$$

Using the fact that $H(Z) = -p\log p - (1-p)\log(1-p)$, standard calculus shows that the right-hand side of (13) is further maximized by letting $p = p^\star$ as given in equation (11). □

As an application of the above theorem, consider the problem where F_1 is the set of all densities with respect to Lebesgue measure μ, that have positive support and mean $\theta > 0$, and F_0 is the set of all densities with respect to μ, that have negative support and mean $-\theta$. Then, the Maximum Entropy mixture distribution is the two-sided exponential density, with scale parameter θ.

The following corollary identifies the mixed (discrete-continuous) r.v. with Maximum Entropy in a certain class.

Corollary 1 *Let F_c be a class of p.d.f.'s with respect to Lebesgue measure μ, and F_d be a class of p.m.f.'s (i.e. densities with respect to a counting measure). Also suppose that the Maximum Entropy problem is well-defined for classes F_c and F_d, i.e. there is $f_c^\star \in F_c$ and $f_d^\star \in F_d$ such that $f_c^\star$ and $f_d^\star$ are the Maximum Entropy distributions in their respective class.*

Let the continuous r.v. $X^\star$ have p.d.f. $f_c^\star$, the dicrete r.v. $Y^\star$ have p.m.f. $f_d^\star$, and let the Bernoulli r.v. $Z^\star$ be independent of $X^\star$ and $Y^\star$,and such that $P(Z^\star = 1) = p^\star$, and $P(Z^\star = 0) = 1 - p^\star$ with

$$p^\star = \frac{1}{1 + \exp\{H(X^\star) - H(Y^\star)\}} \tag{14}$$

Then the mixed (discrete-continuous) r.v. $W^\star$ defined by

$$W^\star = \begin{cases} X^\star & \textit{if } Z^\star = 0 \\ Y^\star & \textit{if } Z^\star = 1 \end{cases} \tag{15}$$

has Maximum Entropy in the class of all mixed r.v.'s W that can be obtained by varying $f_c \in F_c$, $f_d \in F_d$, and $p \in [0,1]$, in equation (5).

As an application of the corollary, consider the problem where F_d is the set of all p.m.f.'s with support on $S_d = \{a_1, a_2, \ldots, a_m\}$, and F_c is the set of all p.d.f.'s with variance σ^2. Then, $Y^\star$ is the uniform r.v. on S_d, and $X^\star$ is the normal $N(0, \sigma^2)$ r.v., and $W^\star$ is defined in (12), with $p^\star = \frac{m}{m + \sqrt{2\pi e \sigma^2}}$. It can be verified that if $\sigma << m$, $p^\star \simeq 1$, and if $m << \sigma$, $p^\star \simeq 0$, which is an intuitive result.

III. Conclusions

A simple expression for entropy (equation (7)) was given that applies equally to discrete, continuous, or mixed (discrete-continuous) random variables. In addition, the mixture distribution possessing Maximum Entropy in a certain class of distributions was identified.

References

[1] Binia, J., Zakai, M., and Ziv, J. (1974), On the ϵ-entropy and the rate distortion function of certain non-Gaussian processes, *IEEE Trans. Info. Theory*, IT-20, 517-524.

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

[3] Dembo, A., Cover, T.M. and Thomas, J. A. (1991), Information Theoretic Inequalities, *IEEE Trans. Info. Theory*, IT-37, 6, 1501-1518.

[4] Kullback, S. (1959), *Information Theory and Statistics*, John Wiley, New York, (reprinted by Peter Smith, 1968).

[5] Loève, M. (1977), *Probability Theory I*, Springer-Verlag.

[6] Pinsker, M.S. (1964), *Information and Information Stability of Random Variables and Processes*, Holden-Day.

[7] Rosenthal, H. and Binia, J. (1988), On the Epsilon Entropy of Mixed Random Variables, *IEEE Trans. Info. Theory*, IT-34, 5, 1110-1114.

[8] Shannon, C.E. and Weaver, W. (1949)), *The Mathematical Theory of Communication*, University of Illinois Press, Urbana.