Critical Analysis of Some Entropy Characterizations

Vesselin I. Dimitrov
Idaho Accelerator Center, Idaho State University, Pocatello ID 83209, USA (dimivess@isu.edu)

1. Generalized Kolmogorov-Nagumo Averages

∘ Kolmogorov (1930), Nagumo (1930) introduced them for vectors. De Finetti (1931) extended them for simple probability distributions.

$$< A >_g = g_\alpha^{-1}\left(\sum_{k=1}^{n} p_k g_\alpha(A_k)\right)$$

where $g_\alpha(x)$ is a continuous and strictly monotonic function depending on parameter(s) α, such that g_α^{-1} exists with the same properties.

Generalized KN average:

$$< A >_{g_\alpha} = \int d\alpha w(\alpha) g_\alpha^{-1}\left(\sum_{k=1}^{n} p_k g_\alpha(A_k)\right)$$

$$w(\alpha) > 0 \qquad \int d\alpha w(\alpha) = 1$$

Additivity

Requiring that

$$< A + c >_g = < A >_g + c$$

results in restricting the possible functions $g_\alpha(x)$ to ones for which

$$g_\alpha(x+y) = g_\alpha(x) g_\alpha(y) \Rightarrow$$
$$g_\alpha(x) = (\exp\alpha)^x = \exp(\alpha x)$$
$$\alpha \geq 0$$

where $\alpha = 0$ corresponds (through a limiting procedure) to the usual average with $g_0(x) = x$. Thus

$$< A >_{w(\alpha)} = \int_0^\infty d\alpha \frac{w(\alpha)}{\alpha} \ln\left(\sum_{k=1}^{n} p_k \exp(\alpha A_k)\right)$$

$$< A >_{\delta(\alpha)} = \sum_{k=1}^{n} p_k A_k$$

$$< A + c >_{w(\alpha)} = < A >_{w(\alpha)} + c$$

The additivity is a property expected on general grounds for any quantity, hence this restriction of the functional form of $g(x)$ is of universal character.

∘ Entropies as expectation values:

a) Shannon-Jaynes

$$H(p) = < \ln p >_{\delta(\alpha)}$$

b) Renyi / Generalized Renyi

$$R_a(p) = < \ln p >_{\delta(\alpha - a)}$$
$$R(p) = < \ln p >_{w(\alpha)}$$

Homogeneity

The above generalized average scales as

$$< cA >_{w(\alpha)} = \int_0^\infty d\alpha \frac{w(\alpha)}{\alpha} \ln\left(\sum_{k=1}^{n} p_k \exp(\alpha c A_k)\right) =$$
$$= \int_0^\infty d\alpha \frac{w(\frac{\alpha}{c})}{\alpha} \ln\left(\sum_{k=1}^{n} p_k \exp(\alpha A_k)\right)$$

Requiring the scaling behavior

$$< cA >_{w(\alpha)} = c < A >_{w(\alpha)}$$

results in the condition

$$w(\frac{\alpha}{c}) = c w(\alpha) \Rightarrow w(\alpha) = \frac{\alpha_0}{\alpha} \cup w(\alpha) = \delta(\alpha)$$

and

$$< A >_{w(\alpha)} = \lim_{\varepsilon \to 0} \frac{\int_\varepsilon^\infty d\alpha \frac{\alpha_0}{\alpha^2} \ln\left(\sum_{k=1}^{n} p_k \exp(\alpha A_k)\right)}{\int_\varepsilon^\infty d\alpha \frac{\alpha_0}{\alpha}} =$$
$$= \sum_{k=1}^{n} p_k A_k$$

Hence, when quantities expected to scale meaningfully (e.g. dimensional observables that can be measured in different units) are involved, the usual average is *singled out* as the only one to be used. This, however, is generally not the case for quantities which are not required to obey particular scaling behavior, like $\ln p$.

2. Shannon, Fadeev & Khinchin

∘ This class of characterizations is obtained by imposing a set of requirements on a function – the entropy – which measures the amount of "disorder" (or, sometimes called "missing information") in a probability distribution. The condition which singles out the particular log form of the entropy is called "strong additivity", and is equivalent to prescribing its behaviour upon rebinning. Consider the entropies $H(p_1, p_2, \cdots, p_n)$ and $H(w_1, w_2, \cdots, w_m)$ where $w_1 = p_1 + \cdots + p_{m_1}$, $w_2 = p_{m_1+1} + \cdots + p_{m_1+m_2}, \cdots, w_m = p_{n-m_m} + \cdots + p_n$ represent a coarse-grained version of the original bins with $\sum_{i=1}^{m} m_i = n$ and $\sum_{i=1}^{m} w_i = \sum_{i=1}^{n} p_i = 1$. Then the "strong additivity" requirement can be formulated as

$$H(\mathbf{p}) = H(\mathbf{w}) + \sum_{i=1}^{m} w_i H(\frac{\mathbf{p}^{(i)}}{w_i})$$

where $\mathbf{p}^{(i)} = \{p_{m_{i-1}}, p_{m_{i-1}+1}, \cdots, p_{m_{i-1}+m_i}\}$. The last term above looks like the average of the partitions entropies $< H(\frac{\mathbf{p}^{(i)}}{w}) >$. Renyi was the first to point out that nothing compells us to use this particular average, and that the more general class of KN averages with $g(x) = \exp(\alpha x)$ produces what is now known as Renyi's entropies H_α while preserving the additivity property.

Another take on the same problem is to consider the entropy as the average $< s >$ of the "surprisal" $s_i = \ln(\frac{1}{p_i})$. The whole class of *apriori* possible KN averages is narrowed down to the exponential family by the requirement for additivity, and ther is no basis whatsoever for further restrictions on the value of α, so we are left with

$$H(p) = < s >_= \frac{1}{\alpha} \ln \sum_{i=1}^{n} p_i^{\alpha+1}$$

References

A. Kolmogorov, Atti della R. Accademia Nazionale dei Lincei 12, 388 (1930).
M. Nagumo, Japan. Journ. Math. 7, 71 (1930).
B. de Finetti, Giornale di Istituto Italiano del Attuarii 2,369 (1931).
S. N. Karbelkar, Pramana J. Phys., 26, p.301 (1986)
J. Uffink, Studies in Hist. Filosophy of Science B, 26, p.223 (1995)
Y. Tikochinsky, N.Z. Tishby, R.D. Levine, Phys. Rev. A30, p.2638 (1984)
J.E. Shore, R.W. Johnson, IEEE Trans. on Information Theory 26, p.26 (1980)
B. Lesche, J. Stat. Phys. 27, p.419 (1982)
K.H. Knuth, J. Skilling, arXiv:1008.4831 [math.PR] (2010)

3. Tikochinsky, Tishby & Levine

∘ TTL #1

The first approach of TTL to the characterization of the entropy is actually a derivation of what TTL believe to be a unique inference scheme possessing *repetition consistency* and *uniformity*. Starting with a probability domain consisting of n bins and $m < n$ linear constraints, $\sum_{i=1}^{n} p_i A_{ri} = a_r$, $r = 1, 2, \cdots, m$, their hypothetical algorithm assignes a probability p_i of a single measurement returning a result in the i^{th} bin. Then TTL assemple a compound experiment, consisting of N independent repetitions of the above single measurement, whose outcome is a set of sample frequencies N_i rather than a single bin index. Applied to the compound problem with *the appropriate constraints*, the algorithm produces probabilities $P_{\{N\}}$. Then TTL argue that if the algorithm knows nothing about the connection between the two experiments – the elementary and the compound one – and treats the data in the same way (hence the *uniformity*), it should produce the multinomial probabilities (hence the *repetition consistency*)

$$P_{\{N\}} = \frac{N!}{\prod_{i=1}^{n} N_i!} \prod_{i=1}^{n} p_i^{N_i}$$

TTL proved that, if the constraints for the compound experiments are chosen as $\sum_{\{N\}} P_{\{N\}} B_{r\{N\}} = N a_r$ with $B_{r\{N\}} = \sum_{i=1}^{n} N_i A_{ri}$ the above two requirements force the algorithm to assign the probabilities

$$p_i = g_i \exp(\lambda_0 + \sum_{r=1}^{m} \lambda_r A_{ri})$$

where λ_i are adjustable Lagrange multipliers and $g_i > 0$ are arbitrary weights independent of the data. This, according to TTL, coincides with the result of the maximization of the Shannon's entropy under r constraints, which thus singles this procedure out as the only consistent one. We have two objections to this proof. *First*, due to the $g_i - s$, the TTL's algorithm appears to act as a probability updating rather than probability assigning scheme. TTL argue that the g-s are universal weights inherent to the problem at hand, but a close examination of their derivation does not support this claim. General considerations show that any probability distribution obeying the constraints $\sum_{i=1}^{n} p_i A_{ri} = a_r$ can be represented as

$$g_i \phi\left(\lambda_0 + \sum_{r=1}^{m} \lambda_r A_{ri}\right)$$

with a continuous and non-negative function $\phi(x)$. Hence, TTL's result has no implications for probability assignments, but still can be regarded as restricting the form of the function to be used for probability updating to an exponential. *Second*, it can be easily seen that the exponential form results from the particular choice of the compound experiment observables as the sample averages $B_{r\{N\}} = \sum_{i=1}^{n} N_i A_{ri}$. This is a reasonable, but not unique choice. Other reasonable KN sample averages can be used to produce different, yet just as consistent, functions $\phi(x)$.

∘ TTL #2

When produced by the TTL's hypothetical Algorithm, the probabilities become functions of the prescribed expectations $\{a_r\}$ through the requirement $\sum_{i=1}^{n} p_i(a) A_{ri} = a_r$. This effectively makes the $a - s$ to behave like *independent* parameters, as upon varying one of the $a - s$ the Algorithm keeps the others fixed. TTL show that if the Algorithm is required to produce probabilities such that

$$for\ all\ r = 1, 2, \cdots, m$$
$$\sum_{i=1}^{n} p_i(a) [\frac{\partial \ln p_i(a)}{\partial a_r}]^2 \times \sum_{j=1}^{n} p_i(a)(A_{rj} - a_r)^2 = 1$$

(i.e. the minimum of the Kramer-Rao inequality is realized) it is bound to produce a distribution of the MaxEnt form

$$p_i = g_i \exp\left(\lambda_0 + \sum_{r=1}^{m} \lambda_r A_{ri}\right)$$

where the $\{g_i\}$ do not depend on the $a - s$. TTL argue that the above requirement for the product of the individual Fisher informations with the corresponding variances to have its *minimal value* $= 1$ is equivalent to requiring the probability assignment to be the least sensitive to statistical errors in the $a - s$. We object to this on two counts:

1) The procedure does not produce a unique probability assignment - rather, the above corresponds to an arbitrary prior $\{g\}$ updated to $\{p\}$ as a result of observation of the $a - s$. This only reproduces the MaxEnt if we make the choice of uniform $g - s$.

2) TTL's sensitivity criterion is not convincing. Indeed the sensitivity of the produced distribution to errors in the a-s can be measured by one of the Ali-Silvey distances, e.g.

$$\delta D^2 = -\sum_{i=1}^{n} p_i(a) \ln \frac{p_i(a + \delta a)}{p_i(a)}$$

Now, to the lowest non-trivial order in δa

$$p_i(a + \delta a) = p_i(a) + \sum_{r=1}^{m} \frac{\partial p_i(a)}{\partial a_r} \delta a_r + \frac{1}{2} \sum_{r=1}^{m} \frac{\partial^2 p_i(a)}{\partial a_r \partial a_s} \delta a_r \delta a_s$$

which, together with the normalization of the probabilities, results in

$$\delta D^2 = \frac{1}{2} \sum_{r,s=1}^{m} \delta a_r \delta a_s \sum_{i=1}^{n} \frac{1}{p_i(a)} \frac{\partial p_i(a)}{\partial a_r} \frac{\partial p_i(a)}{\partial a_s}$$

The r.h.s. is the trace of the product of two *positive-definite* matrices, hence with $\mathtt{Tr}(AB) \leq (\mathtt{Tr}A)(\mathtt{Tr}B)$

$$\delta D^2 \leq \frac{1}{2} \sum_{r=1}^{m} (\delta a_r)^2 \times \sum_{s=1}^{m} \sum_{i=1}^{n} \frac{1}{p_i(a)} [\frac{\partial p_i(a)}{\partial a_s}]^2$$

Thus, the most natural way to ensure the least sensitivity for a fixed magnitude of the perturbations in the $a - s$ appears to be minimizing the trace of the Fisher information matrix

$$\sum_{s=1}^{m} \sum_{i=1}^{n} \frac{1}{p_i(a)} [\frac{\partial p_i(a)}{\partial a_s}]^2 \to \min$$

4. Shore & Johnson, Knuth & Skilling

∘ Shore & Johnson (SJ) (1984) formulated perfectly reasonable requirements for an inference algorithm in the form of four axioms. They demanded, in their terms, *Uniqueness, Invariance, System Independence and Subset Independence.* The inference algorithm was assumed to produce q from p by minimizing a functional in the constraint's subspace

$$q : H(q', p) \underset{q' \in \mathbb{C}}{\to} \min$$

SJ proved Theorems I and II, showing that the requirements of uniqueness, invariance and subset independence restricted the form of $H(q, p)$ to one *equivalent* to

$$F(q, p) = \int dx q(x) h\left(\frac{q(x)}{p(x)}\right)$$

In SJ parlance, H being equivalent to F means that, upon minimization, they produce the same q. Therefore, one can take $H(q, p) = \phi(F(q, p))$ with any continuous and strictly monotonic $\phi(x)$. Then SJ proceeded to explore the consequences of the system independece, which is equivalent to requiring additivity for factorized distributions. Just below their (eqn. 29) they stated "From Theorem II we may assume that H has the form (22)", effectively but silently making the particular choice $\phi(x) = x$. This forced F to be additive and thus fixed the function $h(x)$ to $\ln(x)$. Clearly, this is but one possibility. In general, the requirement can be satisfied with F which factorizes (whence $h(xy) = h(x)h(y) \Rightarrow h(x) = x^\alpha$) and ϕ such that $\phi(xy) = \phi(x) + \phi(y) \Rightarrow \phi(x) = \ln x$, resulting in the class of Renyi's relative entropies. Both Karbelkar (1986) and Uffink (1995) pointed out that Renyi's relative entropy satisfies all SJ's requirements, but failed to identify the above oversight and attacked the SJ's axioms instead.

More recently, Knuth & Skilling (2010) committed the same sin as SJ. They presented a thorough derivation from first principles, which hinged of a certain function ("variational potential") H. In characterizing H, they stated (just before (eqn. 33)) "Hence ... there exists a "$\bigcirc = +$" grade on which H is additive $H(\mathbf{m}) = \sum_{atoms\ i} H_i(m_i)$". This is certainly so, but what they failed to recognize is that there also exist grades on which $H(m)$ can be additive in a different way

$$H(\mathbf{m}) = \ln\left(\sum_{atoms\ i} H_i^\alpha(m_i)\right)$$

and has all the same desirable properties (order, associativity) as their H, resulting, again, in the class of Renyi's divergences.

4. Lesche's Instability

∘ Lesche argued that Renyi's entropies, with the exception of the Shannon-Jaynes one, were unstable with respect to small changes in the probability distribution. The latter were defined by requiring $\sum_{i=1}^{N} |p_i - p_i'|$ to be small. The distributions in Lesche's "counterexamples" are such that they have at least one $p_i = 0$ but all $p_i' \neq 0$. The used "distance" belongs to a special class of distances which do not produce infinity in such a case. This is unnatural for probabilities motivated by observations - for these the distance between such distributions must be *infinite*. It will suffice to demonstrate that the distributions of Lesche's counterexamples are not close for one of his assignments ($0 < q < 1$). This we do using the well justified and generally accepted geometric distance (common for all Ali-Silvey distances):

$$1 >> \delta > 0$$
$$p_1 = 1 - \delta \qquad p_{i>1} = \frac{\delta}{N - 1}$$
$$D(p(0), p(\delta)) = c \times \delta \sqrt{\sum_{i=1}^{n} p_i(\varepsilon) \left[\frac{\partial \ln p_i(\varepsilon)}{\partial \varepsilon}\right]^2 |_{\varepsilon=0}} + O(\delta^2)$$

Elementary calculation shows that

$$\sum_{i=1}^{n} p_i(\varepsilon) \left[\frac{\partial \ln p_i(\varepsilon)}{\partial \varepsilon}\right]^2 |_{\varepsilon=0} = \lim_{\varepsilon \to 0} \left(\frac{1}{1 - \varepsilon} + \frac{1}{\varepsilon}\right)$$

diverges regardless of the value of n. Therefore, the distributions from Lesche's "counterexamples" cannot be reasonably considered close to each other, hence the objection that their Renyi's entropies are not close to each other is inconsequential.

5. Conclusions

∘ We conclude that some of the most popular authoritative characterizations of Shannon's entropy and/or relative entropy as the unique tool for probability assignments and updating fall short of doing the job they are meant for. There are, of course, many more characterizations out there, but if we are allowed to indulge in inference based on experience, they most probably don't work quite as expected, too. It appears likely that the narrowest justifiable class of probability functionals to be used in MaxEnt is the one of all convex linear combination of Renyi's (relative) entropies

$$H_w[\mathbf{p}, \mathbf{q}] = \int_0^\infty d\alpha \frac{w(\alpha)}{\alpha} \ln\left(\sum_{i=1}^{n} p_i (\frac{p_i}{q_i})^\alpha\right)$$

with arbitrary non-negative weights $w(\alpha)$.

Acknowledgment: This work was partially supported by DTRA grant HDTRA1-11-1-0035