Extending Maximum Entropy Techniques to Entropy Constraints

Gang Xiang
Philips Healthcare
El Paso, Texas 79912
gxiang@sigmaxi.net

Vladik Kreinovich
Department of Computer Science
University of Texas at El Paso
El Paso, Texas 79968
vladik@utep.edu

***Abstract*—In many practical situations, we have only partial information about the probabilities. In some cases, we have *crisp* (interval) bounds on the probabilities and/or on the related statistical characteristics. In other situations, we have *fuzzy* bounds, i.e., different interval bounds with different degrees of certainty.**

In a situation with uncertainty, we do not know the exact value of the desired characteristic. In such situations, it is desirable to find its worst possible value, its best possible value, and its "typical" value – corresponding to the "most probable" probability distribution. Usually, as such a "typical" distribution, we select the one with the largest value of the entropy. This works perfectly well in usual cases when the information about the distribution consists of the values of moments and other characteristics. For example, if we only know the first and the second moments, then the distribution with the largest entropy if the normal (Gaussian) one.

However, in some situations, we know the entropy (= amount of information) of the distribution. In this case, the maximum entropy approach does not work, since all the distributions which are consistent with our knowledge have the exact sam e entropy value. In this paper, we show how the main ideas of the maximum entropy approach can be extended to this case.

I. Probabilities are Usually Imprecise: A Reminder

Probabilities are imprecise. In many practical situations, we have only *partial* information about the probabilities, i.e., our information about the probabilities is imprecise.

Types of imprecise probability. In different situations, we have different types of partial information.

- In some cases, we have *crisp* (interval) bounds on the probabilities and/or on the related statistical characteristics.
- In other situations, we have *fuzzy* bounds, i.e., in effect, different interval bounds with different degrees of certainty.

Processing imprecise probabilities: the Maximum Entropy approach. We are interested in finding the values of certain statistical characteristics. In the ideal case when we know the exact values of all the probabilities, we can determine the exact values of the corresponding statistical characteristic – e.g., the mean (expected value) of a certain quantity (like gain or loss in economic situations).

In situations when probabilities are only known with uncertainty, we cannot predict the exact value of the desired characteristic. In such situations, from the common sense viewpoint, it is desirable to find:

- the worst possible value of this characteristic,
- the best possible value of this characteristic, and
- the "typical" value of this characteristic.

Intuitively, by the "most typical" characteristic, we mean the value of this characteristic that corresponds to the "most probable" probability distribution. Usually, as such a "typical" distribution, we select the one with the largest value of the *entropy* (see, e.g., [1], [2]).

For a discrete probability distribution, in which we have n different values $x_1, \ldots, x_n$ with probabilities $p_1, \ldots, p_n$, the entropy $S(p)$ is defined as

$$S = -\sum_{i=1}^{n} p_i \cdot \log_2(p_i). \tag{1}$$

For a continuous distribution with a probability density function $\rho(x)$, the entropy is defined as the integral (crudely speaking, the limit of the above sum):

$$S = -\int \rho(x) \cdot \log_2(\rho(x))\, dx. \tag{2}$$

The meaning of the entropy is that it represents the average number of "yes"-"no" questions that we need to ask to determine the exact value x_i (in the discrete case) or the value x with a given accuracy $\varepsilon > 0$ (in the continuous case). In other words, the entropy represents the amount of information as measured in *bits* (i.e., numbers of binary, "yes"-"no" questions).

Successes of the Maximum Entropy approach: a brief reminder. The Maximum Entropy approach works perfectly well in usual cases when the information about the distribution consists of the values of moments and other characteristics.

For example, if we only know the first and the second moments, then the distribution with the largest entropy if the normal (Gaussian) one.

A problem: situations in which the Maximum Entropy Approach is not applicable. As we have mentioned, the entropy itself is – like moments – a reasonable statistical characteristic of the probability distribution, a characteristic that has a clear meaning: it describes the amount of information.

It is therefore not surprising that in some practical situations, in addition to knowing the values of the moments etc., we also know the value of the entropy of the (unknown) distribution. In this case, the Maximum Entropy approach does not work. Indeed,

- In the Maximum Entropy approach, among all the distributions which are consistent with the our (partial) knowledge, we select a one for which the entropy is the largest possible. This requirement usually selects a unique distribution for which the entropy is the largest. (In rare cases, there may be a few different distributions with the same largest value of entropy.)
- However, in our case, our partial knowledge includes the knowledge of the entropy value S. Thus, all the probability distributions which are consistent with this partial information have the exact same value of the entropy – the value S. So, *all* these distributions will be kept intact by the Maximum Entropy approach – so this approach does not allow us to select any distribution as "most typical".

What we plan to do. In this paper, we show how the maximum entropy approach can be naturally extended – so that it will also be able to cover the case when entropy is one of the constraints.

II. Main Idea and Its Consequences

In practice, we always have uncertainty. In practice, even in the ideal case when we observe a large number N of situations corresponding to the same probability distribution, we still cannot determine the corresponding probabilities $p_1, \ldots, p_n$ with an absolute accuracy – we can only determine the *frequencies* $f_1, \ldots, f_n$ with which the values $x_1, \ldots, x_n$ have been observed.

In the limit, when the sample size N tends to infinity, the frequencies f_i tend to the corresponding probabilities, but for finite N, they differ.

Resulting idea. We are considering a typical case when we do not have the full information neither about the probabilities p_i nor about the frequencies f_i. When N is given, we still cannot uniquely determine the probabilities even when we know the frequencies exactly. Therefore, instead of selecting "typical" probabilities, let us select "typical" frequencies.

When N is large, the probabilities and frequencies are close, so for all computational purposes, we can use the frequencies instead of the probabilities to compute the desired statistical characteristics.

However, from the viewpoint of selecting the frequencies, the difference between the frequencies and the probabilities open the possibility of using the Maximum Entropy approach. Specifically, for each collection of frequencies $f = (f_1, \ldots, f_n)$, we have different possible tuples of probabilities $(p_1, \ldots, p_n)$, with slightly different values of the entropy. Among all these possible values of the probabilities, we can now select the one for which the entropy is the largest.

Let us show how this idea can be transformed into the exact formulas.

Relation between frequencies and probabilities: a reminder. It is known (see, e.g., [3]) that asymptotically, for large N, the differences $\delta_i \stackrel{\text{def}}{=} p_i - f_i$ are independent normally distributed random variables, with mean 0 and variance $\sigma_i^2 = \dfrac{f_i \cdot (1 - f_i)}{N}$.

What can we say, based on this information, about the relation between frequencies and probabilities? In statistics, there is a standard χ^2-criterion for checking whether the given set of observables $t_1, \ldots, t_n$ is consistent with the assumption that they are normally distributed with mean 0 and variance σ_i^2: the value

$$\sum_{i=1}^{n} \frac{t_i^2}{\sigma_i^2} \tag{3}$$

must be (approximately) equal to n; see, e.g., [3].

Thus, the above information about the relation between the frequencies and probabilities means that we must have

$$\sum_{i=1}^{n} \frac{\delta_i^2}{\sigma_i^2} = \sum_{i=1}^{n} \frac{\delta_i^2}{f_i \cdot (1 - f_i)/N} \approx n. \tag{4}$$

Dividing both sides of this approximate inequality by N, we conclude that the constraint (4) can be described in the equivalent form, as

$$\sum_{i=1}^{n} \frac{\delta_i^2}{f_i \cdot (1 - f_i)} \approx \frac{n}{N}. \tag{5}$$

Let us maximize entropy under this constraint. For each selection of frequencies f_i, we want to maximize the entropy S under the above constraint. In terms of the probabilities p_i, the entropy has the form

$$S(p_1, \ldots, p_n) = -\sum_{i=1}^{n} p_i \cdot \log_2(p_i). \tag{6}$$

By definition of the differences $\delta_i = p_i - f_i$, we have $p_i = f_i + \delta_i$, so

$$S(p_1, \ldots, p_n) = S(f_1 + \delta_1, \ldots, f_n + \delta_n) = -\sum_{i=1}^{n} p(f_i + \delta_i) \cdot \log_2(f_i + \delta_i). \tag{7}$$

For large N, the differences δ_i are small, so we can expand the expression for S in Taylor series in terms of δ_i and keep only linear terms in this expansion – ignoring quadratic and higher order terms. As a result, we get the formula

$$S(f_1 + \delta_1, \ldots, f_n + \delta_n) = S(f_1, \ldots, f_n) + \sum_{i=1}^{n} \frac{\partial S}{\partial f_i} \cdot \delta_i. \tag{8}$$

Here, the entropy $S(f_1, \ldots, f_n)$ is equal to the given value S_0, and

$$\frac{\partial S}{\partial f_i} = -\log_2(f_i) - \log_2(e), \tag{9}$$

so the maximized expression (8) has the form

$$S_0 - \sum_{i=1}^{n} (\log_2(f_i) + \log_2(e)) \cdot \delta_i. \tag{10}$$

To maximize this expression under the constraint (5), we can the Lagrange multiplier technique, i.e., for an appropriate value λ, we solve the unconstrained optimization problem of maximizing the following function:

$$S_0 - \sum_{i=1}^{n} (\log_2(f_i) + \log_2(e)) \cdot \delta_i + \lambda \cdot \sum_{i=1}^{n} \frac{\delta_i^2}{f_i \cdot (1 - f_i)}. \tag{11}$$

Differentiating this expression with respect to δ_i and equating the derivative to 0, we conclude that

$$-(\log_2(f_i) + \log_2(e)) + \lambda \cdot \frac{2\delta_i}{f_i \cdot (1 - f_i)} = 0, \tag{12}$$

hence

$$\delta_i = c \cdot (\log_2(f_i) + \log_2(e)) \cdot f_i \cdot (1 - f_i) \tag{13}$$

for an appropriate constant

$$c = \frac{1}{2 \cdot \lambda}. \tag{14}$$

Substituting this formula into the expression (10), we conclude that the entropy is equal to $S_0 + \Delta S$, where

$$\Delta S = c \cdot \sum_{i=1}^{n} (\log_2(f_i) + \log_2(e))^2 \cdot f_i \cdot (1 - f_i). \tag{15}$$

As usual for the Lagrange multiplier method, the value c of the parameter (i.e., in effect, the value λ of the Lagrange multiplier) can be determined from the constraint (5). Substituting the formula (13) into the constraint (5), we conclude that

$$c^2 \cdot \sum_{i=1}^{n} \frac{(\log_2(f_i) + \log_2(e))^2 \cdot f_i^2 \cdot (1 - f_i)^2}{f_i \cdot (1 - f_i)} =$$

$$c^2 \cdot \sum_{i=1}^{n} (\log_2(f_i) + \log_2(e))^2 \cdot f_i \cdot (1 - f_i) = \frac{n}{N}. \tag{16}$$

Thus, we conclude that

$$c^2 = \frac{n}{N \cdot \sum\limits_{i=1}^{n} (\log_2(f_i) + \log_2(e))^2 \cdot f_i \cdot (1 - f_i)} \tag{17}$$

and that

$$c = \frac{\sqrt{n}}{\sqrt{N} \cdot \sqrt{\sum\limits_{i=1}^{n} (\log_2(f_i) + \log_2(e))^2 \cdot f_i \cdot (1 - f_i)}}. \tag{18}$$

Substituting this expression for c into the formula (15), we conclude that

$$\Delta S = \frac{\sqrt{n} \cdot \sum\limits_{i=1}^{n} (\log_2(f_i) + \log_2(e))^2 \cdot f_i \cdot (1 - f_i)}{\sqrt{N} \cdot \sqrt{\sum\limits_{i=1}^{n} (\log_2(f_i) + \log_2(e))^2 \cdot f_i \cdot (1 - f_i)}} =$$

$$\frac{\sqrt{n}}{\sqrt{N}} \cdot \sqrt{\sum_{i=1}^{n} (\log_2(f_i) + \log_2(e))^2 \cdot f_i \cdot (1 - f_i)}. \tag{19}$$

Thus, the largest value of ΔS is attained when the following characteristic is the largest:

$$S_2 \stackrel{\text{def}}{=} \sum_{i=1}^{n} (\log_2(f_i) + \log_2(e))^2 \cdot f_i \cdot (1 - f_i). \tag{20}$$

So, we arrive at the following conclusion:

How to extend the maximum entropy approach to the case when entropy is one of the constraints: resulting formula. Let us consider the situations in which the fixed value of entropy is one of the constraints on the probability distribution. In this situation, among all distributions that satisfy this (and other) constraints, we must select a one for which the following characteristic takes the largest possible value:

$$S_2 = \sum_{i=1}^{n} (\log_2(p_i) + \log_2(e))^2 \cdot p_i \cdot (1 - p_i). \tag{21}$$

III. Continuous Case

Formulation of the problem. The above formulas relate to the case when we have finite number (n) of possible outcomes, and we want to find the probabilities $p_1, \ldots, p_n$ of these n outcomes.

In many practical problems, we have a *continuous* case, when all real numbers x from a certain interval (finite or infinite) are possible. In this continuous case, we are interested in finding the *probability density* $\rho(x)$ characterizing this distribution.

Reminder. The probability density is defined as the limit

$$\rho(x) = \lim_{\Delta x \to 0} \frac{p([x, x + \Delta x])}{\Delta x}, \tag{22}$$

when the width Δx of the corresponding interval tends to 0. Here, $p([x, x + \Delta x])$ is the probability to find the value of a random variable inside the interval $[x, x + \Delta x]$.

How we usually deal with the continuous case: general idea (reminder). A usual way to deal with the continuous case is to divide the interval of possible values of x into several small intervals $[x_i, x_i + \Delta x]$ of width Δx and consider the discrete distribution with these intervals as possible values. When Δx is small, by the definition of the probability density, the probability $p_i = p([x_i, x_i + \Delta x])$ for x to be inside the i-th interval $[x_i, x_i + \Delta x]$ is approximately equal to

$$p_i \approx \rho(x_i) \cdot \Delta x. \tag{23}$$

We use these values and then consider the limit case, when $\Delta x \to 0$.

Example. This is how we go from the discrete entropy

$$S = -\sum_{i=1}^{n} p_i \cdot \log_2(p_i). \tag{24}$$

to the corresponding continuous formula

$$S = -\int \rho(x) \cdot \log_2(\rho(x))\, dx. \tag{25}$$

Indeed, substituting $p_i \approx \rho(x_i) \cdot \Delta x$ into the formula (24), we conclude that

$$S = -\sum_{i=1}^{n} \rho(x_i) \cdot \Delta x \cdot \log_2(\rho(x_i) \cdot \Delta x). \tag{26}$$

Since the logarithm of the product is equal to the sum of the logarithms, we conclude that

$$S = -\sum_{i=1}^{n} \rho(x_i) \cdot \Delta x \cdot \log_2(\rho(x_i)) - \sum_{i=1}^{n} \rho(x_i) \cdot \Delta x \cdot \log_2(\Delta x). \tag{27}$$

Let us analyze these two terms one by one.

The first term in this sum is the integral sum for the integral $\int \rho(x) \cdot \log_2(\rho(x))\, dx$. When $\Delta x \to 0$, this term tens to this integral. Thus, when Δx is small, the term

$$\sum_{i=1}^{n} \rho(x_i) \cdot \Delta x \cdot \log_2(\rho(x_i)) \tag{28}$$

is close to the integral $\int \rho(x) \cdot \log_2(\rho(x))\, dx$.

The term $\sum\limits_{i=1}^{n} \rho(x_i) \cdot \Delta x$ is the integral sum for the integral $\int \rho(x)\, dx$. This integral describes the probability of finding x somewhere, and is, therefore, equal to 1. Thus, when Δx is small, this sum is approximately equal to 1. Hence, when Δx is small, the second term

$$\sum_{i=1}^{n} \rho(x_i) \cdot \Delta x \cdot \log_2(\Delta x) \tag{29}$$

is approximately equal to $\log_2(\Delta x)$.

Substituting these approximate expressions into the formula (27), we conclude that for small Δx, we have

$$S = -\int \rho(x) \cdot \log_2(\rho(x))\, dx - \log_2(\Delta x). \tag{30}$$

The second term in this sum does not depend on the probability distribution at all. Thus, maximizing the entropy S is equivalent to maximizing the integral in the expression (30).

This is exactly what is called the entropy of the continuous distribution.

Towards adjusting the new formula for the case of a continuous distribution. Let us apply the same procedure to our new characteristic

$$S_2 = \sum_{i=1}^{n} (\log_2(p_i) + \log_2(e))^2 \cdot p_i \cdot (1 - p_i). \tag{31}$$

Substituting $p_i = \rho(x_i) \cdot \Delta x$ into this expression, we get

$$S_2 = \sum_{i=1}^{n} A_i \cdot \rho(x_i) \cdot \Delta x \cdot (1 - \rho(x_i) \cdot \Delta x), \tag{32}$$

where we denoted

$$A_i \stackrel{\text{def}}{=} (\log_2(\rho(x_i) \cdot \Delta x) + \log_2(e))^2. \tag{33}$$

When Δx is small, we have $\rho(x_i) \cdot \Delta x \ll 1$. Thus, asymptotically,

$$1 - \rho(x_i) \cdot \Delta x \approx 1, \tag{34}$$

and the formula (32) can simplified into

$$S_2 = \sum_{i=1}^{n} (\log_2(\rho(x_i) \cdot \Delta x) + \log_2(e))^2 \cdot \rho(x_i) \cdot \Delta x. \tag{35}$$

Here,

$$\log_2(\rho(x_i) \cdot \Delta x) + \log_2(e) = \log_2(\rho(x_i)) + \log_2(\Delta x) + \log_2(e). \tag{36}$$

Thus,

$$\begin{aligned}(\log_2(\rho(x_i) \cdot \Delta x) + \log_2(e))^2 = \\ (\log_2(\rho(x_i)) + \log_2(\Delta x) + \log_2(e))^2 = \\ (\log_2(\rho(x_i)) + (\log_2(\Delta x) + \log_2(e)))^2 = \\ \log_2^2(\rho(x_i)) + \\ 2 \cdot \log_2(\rho(x_i)) \cdot (\log_2(\Delta x) + \log_2(e)) + \\ (\log_2(\Delta x) + \log_2(e))^2.\end{aligned} \tag{37}$$

Therefore, the expression (35) can be represented as the sum of the following three terms:

$$S_2 = \sum_{i=1}^{n} \log_2^2(\rho(x_i) \cdot \rho(x_i)) \cdot \Delta x + 2 \cdot \sum_{i=1}^{n} \log_2(\rho(x_i)) \cdot (\log_2(\Delta x) + \log_2(e)) \cdot \rho(x_i) \cdot \Delta x + \sum_{i=1}^{n} (\log_2(\Delta x) + \log_2(e))^2 \cdot \rho(x_i) \cdot \Delta x. \tag{38}$$

Let us analyze these terms one by one.

The first term is an integral sum for the integral

$$\int \log^2(\rho(x)) \cdot \rho(x)\, dx. \tag{39}$$

So, for small Δx, this term is very close to this integral – and the smaller Δx, the closer the first sum to this integral (39).

The second term is proportional to the sum

$$\sum_{i=1}^{n} \log_2(\rho(x_i)) \cdot \rho(x_i) \cdot \Delta x, \tag{40}$$

with the coefficient proportionality

$$2 \cdot (\log_2(\Delta x) + \log_2(e)). \tag{41}$$

The sum (40) is an integral sum for the integral

$$\int \log_2(\rho(x)) \cdot \rho(x)\, dx, \tag{42}$$

which is simply $-S$, where S is the entropy of the corresponding continuous distribution. Thus, for small Δx, the sum (40)

is approximately equal to $-S$, and the second term is thus asymptotically equal to

$$-2 \cdot (\log_2(\Delta x) + \log_2(e)) \cdot S. \tag{43}$$

The third term is proportional to the sum

$$\sum_{i=1}^{n} \rho(x_i) \cdot \Delta x, \tag{44}$$

with the coefficient proportionality

$$(\log_2(\Delta x) + \log_2(e))^2. \tag{45}$$

Similarly to the entropy case, the sum (44) is an integral sum for the integral

$$\int \rho(x)\, dx = 1. \tag{46}$$

Thus, for small Δx, the sum (44) is approximately equal to 1, and the third term is thus asymptotically equal to

$$(\log_2(\Delta x) + \log_2(e))^2. \tag{47}$$

Adding up the expressions (39), (43), and (47) for the three terms that form the expression (38), we conclude that

$$S_2 = \int \log^2(\rho(x)) \cdot \rho(x)\, dx - 2 \cdot (\log_2(\Delta x) + \log_2(e)) \cdot S + (\log_2(\Delta x) + \log_2(e))^2. \tag{48}$$

The third term does not depend on the probability distribution at all, so it can be ignored when we compare the values S_2 of different distributions – to select the most "typical" one.

The second term depends only on the step size Δx and on the entropy S. Since we consider the situations in which the value of the entropy is known (since this value is one of the constraints), this term also has the same value for all the distributions that are consistent with all these constraints. Thus, this term can also be ignored when we compare the values S_2 of different distributions – to select the most "typical" one.

Thus, in this selection, the only important term is the first (integral) term: selecting the distribution with the largest possible value of S_2 is equivalent to selecting the distribution with the largest possible value of the integral (39).

Thus, we arrive at the following recommendation:

How to extend the maximum entropy approach to the case when entropy is one of the constraints: resulting formula for the continuous case. Let us consider the situations in which the fixed value of entropy is one of the constraints on the probability distribution. In this situation, among all distributions that satisfy this (and other) constraints, we must select a one for which the following characteristic takes the largest possible value:

$$S_2 = \int \log_2^2(\rho(x)) \cdot \rho(x)\, dx. \tag{49}$$

IV. Meaning of the New Formula

Meaning of entropy: reminder. In order to come up with a reasonable interpretation of this characteristic, let us recall the usual interpretation of the entropy S. This interpretation is related to the average number of binary ("yes"-"no") questions that we need to ask to locate the value x with a given accuracy $\varepsilon > 0$.

In general, when we have N elements and we ask one yes-no question, then it is reasonable to divide the elements into two equal groups of size $N/2$ and ask whether an element belongs to the first group. After this answer, we can decrease the number of possible elements by a factor of two, from N to $N/2$. If we can ask the second question, we can further divide this number by two, to $(N/2)/2 = N/2^2$. After q questions, we have $N/2^q$ possible elements, etc. Thus, if we have a group of N elements and we want to reduce to a group of $n \ll N$ elements, than the required number q of binary questions can be found from the condition that $N/2^q \approx n$, i.e., $2^q \approx N/n$, and $q \approx \log_2(N/n)$.

When we have a value which is close to x, and we want to locate it with accuracy ε, this means that we want, starting from the original list of N possible elements, to restrict ourselves to the list of all the elements which are ε-close to x, i.e., which are located in the interval $[x-\varepsilon, x+\varepsilon]$ of width 2ε. The proportion of elements in this interval is approximately equal to $\rho(x) \cdot 2\varepsilon$. Thus, to locate this element, we must get down from the original number of N elements to the number $n = N \cdot (\rho(x) \cdot 2\varepsilon)$ elements in this interval. This reduction requires

$$q(x) \approx \log_2\left(\frac{N}{n}\right) = \log_2\left(\frac{N}{N \cdot (\rho(x) \cdot 2\varepsilon)}\right) = \log_2\left(\frac{1}{\rho(x) \cdot 2\varepsilon}\right) = -\log_2(\rho(x)) - \log_2(2\varepsilon). \tag{50}$$

Thus, to locate a value close to x with a given accuracy, we need to ask $k(x) \approx -\log_2(\rho(x))$ binary questions.

The average number of such questions is equal to

$$E[q] = \int q(x) \cdot \rho(x)\, dx = -\int \log_2(\rho(x)) \cdot \rho(x)\, dx. \tag{51}$$

Thus, the entropy is simply the average number of questions that we need to ask to locate the corresponding random value with a given accuracy.

In these terms, the idea of the Maximum Entropy method is that among all possible distribution, we select the one which is the least certain, i.e., for which we need, on average, the largest number of binary questions to locate the corresponding value.

From the meaning of entropy to the meaning of the new function. We consider the situations in which entropy is one of the constraints, i.e., situations when we know the entropy of the distribution. In such situations, because of the meaning of entropy, for all the distributions which are consistent with the given information, the *average* number of questions is the same.

What may differ is the "spread" of these values. Some distributions may require the same number of questions for all x, for others, for some x, we need many more questions, for some, much less. The number of questions $q(x)$ corresponding to a randomly selected x is s random variable itself. Usually, the spread of a random variable can be described if we know its mean $E[q]$ and its standard deviation $\sigma[q]$: with high confidence, the actual value of $q(x)$ is within the interval $[E[q]-k\cdot\sigma[q], E[q]+k\cdot\sigma[q]]$, where the parameter k (usually, 2, 3, or 6) depends on the confidence with which we want to conclude that $q(x)$ belongs to this interval.

In our situation, all distributions have the same mean $E[q]$ (entropy), but they may have different standard deviations $\sigma[q]$. When we select among several possible distributions, the one which is the least certain – i.e., the one that requires the largest number of binary questions – it is therefore reasonable to select the one for which the largest possible number of questions $E[q]+k\cdot\sigma[q]$ is the largest possible. Since the mean $E[q]$ is the same for all these distributions, this means that we select a distribution with the largest standard deviation $\sigma[q]$.

It is known that the standard deviation is related to the second moment $E[q^2]$ by the formula $\sigma^2[q]=E[q^2]-(E[q])^2$. Thus, when the mean $E[q]$ is fixed, maximizing the standard deviation is equivalent to maximizing the second moment $E[q^2]$.

In our case, $q(x)\approx -\log_2(\rho(x))$, so the second moment has the form

$$E[q^2]=\int (\log_2(\rho(x))^2\cdot\rho(x)\,dx. \tag{52}$$

This is exactly our function S_2. Thus, we arrive at the following interpretation of the new function S_2.

Resulting interpretation. As we remember, entropy is the average number of binary questions that we need to ask to locate the corresponding random value x. Similarly, the new function S_2 is the expected value of the square of the number of binary questions.

When the entropy is fixed, maximizing this new function is equivalent to maximizing the standard deviation of the number of binary questions.

Thus, when we know the *average* number of binary questions need to locate x, among all possible distributions we select a one for which the *standard deviation* of the number of binary questions is the largest.

V. Conclusions and Future Work

Formulation of the problem. In many practical situations, we have incomplete information about the probabilities. In this case, among all possible probability distributions, it is desirable to select the most "typical" one. Traditionally, we select the distribution which has the largest possible value of the entropy S. For example, for a continuous distribution with probability density $\rho(x)$, entropy has the form

$$S=-\int \log_2(\rho(x))\cdot\rho(x)\,dx. \tag{53}$$

This Maximum Entropy approach has good justifications, and it can be applied (and has been successfully applied) in many practical situations.

However, there are situations in which this method cannot be directly applied. Specifically, in some case, one of the characteristics that we know about the distribution is its entropy S_0. In this case, all distributions which are consistent with this information have the exact same entropy value S_0. Thus, it is impossible to select one of them based on the value of its entropy – as the Maximum Entropy approach suggest.

Our main result. In this paper, we show that in such situations, we should maximize a special characteristic S_2. For example, for a continuous distribution with probability density $\rho(x)$, this characteristic has the form

$$S_2=\int \log_2^2(\rho(x))\cdot\rho(x)\,dx. \tag{54}$$

Remaining open questions. An interesting open question is: what if we know *both* the values of the entropy S and of the new characteristic S_2. In this case, neither the Maximum Entropy approach, not maximizing the new characteristic S_2 help. It is desirable to extend the Maximum Entropy approach to such situations, i.e., to come up with a new characteristic S_3 that should be maximized in this case.

Similarly, when we know the values of S, S_2, and S_3, we should be able to maximize yet another characteristic S_4, etc.

Acknowledgment

This work was also supported in part by the National Science Foundation grants HRD-0734825 and DUE-0926721, by Grant 1 T36 GM078000-01 from the National Institutes of Health, and by Grant 5015 from the Science and Technology Centre in Ukraine (STCU), funded by European Union.

The authors are thankful to Ron Yager for valuable discussions.

References

[1] E. T. Jaynes, *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, Massuchesetts, 2003.

[2] G. J. Klir, *Uncertainty and Information: Foundations of Generalized Information Theory*, IEEE Press and John Wiley, Hoboken, New Jersey, 2005.

[3] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman & Hall/CRC, Boca Raton, Florida, 2004.

[4] G. Xiang, *Fast Algorithms for Computing Statistics under Interval Uncertainty, with Applications to Computer Science and to Electrical and Computer Engineering*, PhD Dissertation, Department of Computer Science, University of Texas at El Paso, 2007.

[5] R. R. Yager, "Entropy and Specificity in a Mathematical Theory of Evidence", In: R. R. Yager and L. Liu (Eds.), *Classic Works of the Dempster-Shafer Theory of Belief Functions*, Springer Studies in Fuzziness and Soft Computing, Vol. 219, Springer Verlag, Berlin Heidelberg, 2008, pp. 291–310.