Maximum Entropy Beyond Selecting Probability Distributions

Thach N. Nguyen, Olga Kosheleva, and Vladik Kreinovich

Abstract Traditionally, the Maximum Entropy technique is used to select a probability distribution in situations when several different probability distributions are consistent with our knowledge. In this paper, we show that this technique can be extended beyond selecting probability distributions, to explain facts, numerical values, and even types of functional dependence.

1 How Maximum Entropy Technique Is Currently Used

Need to select a distribution: formulation of a problem. Many data processing techniques assume that we know the probability distribution – e.g., the probability distributions of measurement errors, and/or probability distributions of the signals; see, e.g., [6, 7].

Often, however, we have only partial information about a probability distribution. In such cases, there are several different probability distributions which are consistent with the available knowledge. To apply to this situation a data processing algorithm which is based on the assumption that the probability distribution is known, we must select a single probability distribution out of all distributions which are consistent with our knowledge. How can we select such a distribution?

Main idea. By selecting a single distribution out of several, we inevitably decrease uncertainty. It is reasonable to select a distribution for which this decrease in uncertainty is as small as possible.

Thach N. Nguyen
Banking University of Ho Chi Minh City, 56 Hoang Dieu 2, Quan Thu Duc, Thu Duc
Ho Ch Minh City, Vietnam, e-mail: Thachnn@buh.edu.vn

Olga Kosheleva and Vladik Kreinovich
University of Texas at El Paso, 500 W. University,
El Paso, Texas 79968, USA, e-mail: olgak@utep.edu, vladik@utep.edu

How to describe this idea as a precise optimization problem. A natural way to measure uncertainty is by the average number of binary ("yes"-"no") questions that we need to ask to uniquely determine the corresponding random value (or, in the case of continuous variables, to determine the random value with a given accuracy ε).

One can show that for a probability distribution with a given probability density function $\rho(x)$, this average number of binary questions is asymptotically (when $\varepsilon \to 0$) proportional to the *entropy* $S(\rho) \stackrel{\text{def}}{=} -\int \rho(x) \cdot \ln(\rho(x))\,dx$ of this probability distribution; see, e.g., [5] and references therein.

For a class F of distributions, the average number of binary question is asymptotically proportional to $\max\limits_{\rho \in F} S(\rho)$. We want select a single distribution ρ_0 from the class F for which the decrease in uncertainty is the smallest possible, i.e., for which the difference $\max\limits_{\rho \in F} S(\rho) - S(\rho_0)$ is the smallest possible.

How to solve the corresponding optimization problem: enter maximum Entropy technique. There is a natural solution to this optimization problem: select a distribution ρ_0 for which the entropy is the largest possible, i.e., for which $S(\rho_0) = \max\limits_{\rho \in F} S(\rho)$. In this case. the desired difference is 0 – and so the decrease in uncertainty is asymptotically negligible.

This is the main idea behind the *Maximum Entropy techniques:* when we need to select a single distribution for the class of all possible distributions, we select the distribution ρ for which the entropy $S(\rho)$ attains the largest possible value.

Simple examples of using the Maximum Entropy techniques. In some cases, all we know is that the random variable is located somewhere on a given interval $[a,b]$, but we have no information about the probability of it being in different parts of this interval. Which probability distribution would we then select to represent this situation?

If we use the Maximum Entropy approach, then we need to maximize the expression $-\int_a^b \rho(x) \cdot \ln(\rho(x))\,dx$ under the condition that the function $\rho(x) \geq 0$ is a probability density function, i.e., that $\int_a^b \rho(x)\,dx = 1$.

Thus, we get a *constraint optimization problem:* optimize the entropy under the constraint $\int_a^b \rho(x)\,dx = 1$. To solve this constraint optimization problem, we can use the Lagrange multiplier method and reduce to the following unconstrained optimization problem of maximizing the following expression:

$$-\int_a^b \rho(x) \cdot \ln(\rho(x))\,dx + \lambda \cdot \left(\int_a^b \rho(x)\,dx - 1\right),$$

where λ is the *Lagrange multiplier* – a constant that needs to be determined so that the original constraint will be satisfied.

We want to find the function ρ, i.e., we want to find the values $\rho(x)$ corresponding to different inputs x. Thus, the unknowns in this optimization problem are the values $\rho(x)$ corresponding to different inputs x. To solve the resulting unconstrained optimization problem, we can simply differentiate the above expression by each of

the unknowns $\rho(x)$ and equate the resulting derivative to 0. As a result, we conclude that $-\ln(\rho(x)) - 1 + \lambda = 0$, hence $\ln(\rho(x))$ is a constant not depending on x (and equal to $\lambda - 1$). Therefore, the probability density function $\rho(x)$ itself is a constant. Thus, in this case, the Maximum Entropy technique leads to a *uniform* distribution on the interval $[a,b]$.

This conclusion makes perfect sense: if we have no information about which values from the interval $[a,b]$ are more probable and which are less probable, it is reasonable to conclude that all these values are equally probable, i.e., that $\rho(x) =$ const. (This idea goes back to Laplace and is known as the *Laplace Indeterminacy Principle*.)

In other situations, the only information that we have about the probability distribution on a real line is its first two moments $\int x \cdot \rho(x)\,dx = \mu$ and

$$\int (x-\mu)^2 \cdot \rho(x)\,dx = \sigma^2.$$

In this case, the Maximum Entropy technique means selecting a distribution for which the entropy is the largest under the above two constraints and the constraint that $\int \rho(x)\,dx = 1$. For this problem, the Lagrange multiplier methods leads to the following unconstrained optimization problem, in which λ_i are Lagrange multipliers:

$$\text{Maximize } -\int \rho(x) \cdot \ln(\rho(x))\,dx + \lambda_1 \cdot \left(\int x \cdot \rho(x)\,dx - \mu\right) +$$

$$\lambda_2 \cdot \left(\int (x-\mu)^2 \cdot \rho(x)\,dx - \sigma^2\right) + \lambda_3 \cdot \left(\int_a^b \rho(x)\,dx - 1\right).$$

Differentiating the maximized expression with respect to each unknown $\rho(x)$ and equating the resulting derivative to 0, we conclude that

$$-\ln(\rho(x)) - 1 + \lambda_1 \cdot x + \lambda_2 \cdot (x-\mu)^2 + \lambda_3 = 0,$$

i.e., we conclude that $\ln(\rho(x))$ is a quadratic function of x and thus, that $\rho(x) = \exp(\ln(\rho(x)))$ is a Gaussian distribution.

This conclusion is also in good accordance with common sense. Indeed:

- in many case, e.g., the measurement error results from many independent small effects and,
- according to the Central Limit Theorem, the distribution of the sum of a large number of independent small random variables is close to Gaussian.

There are many other examples of a successful use of the Maximum Entropy technique; see, e.g., [4].

A natural question. Since the Maximum Entropy technique works so well for selecting a distribution, can we extend it solving other problems – e.g., explaining a fact, finding the unknown value of a quantity, or finding the formula for a functional dependence?

What we do in this paper. In this paper, we show, on several examples, that such an extension is indeed possible. We will show it on case studies that cover all three types of possible problems: explaining a fact, finding the number, and finding the functional dependence.

2 First Case Study: How Maximum Entropy Techniques Can Be Used to Explain a Fact

Fact to be explained. This fact comes from a recent study [1], and it is related to the uncertainty of expert estimates.

Experts' estimates are imprecise – just like measuring instruments are imprecise. Moreover, when we ask the same expert after some time to estimate the same quantity, he/she will, in general, give a slightly different estimate – just like when we repeatedly measure the same quantity with the same measuring instrument, we, in general, get slightly different results. We can describe the expert's estimates x_i of a quantity x as $x_i = x + \Delta x_i$, where $\Delta x_i \stackrel{\text{def}}{=} x_i - x$ is the estimation error.

A reasonable way to gauge the expert's accuracy is to compute the mean square value of the expert's estimation error, i.e., the value $\sigma_x \stackrel{\text{def}}{=} \sqrt{\frac{1}{N} \cdot \sum_{i=1}^{n} (\Delta x_i)^2}$, where N is the overall number of estimates performed by this expert. This quantity describes the *intra-expert* variation of the expert estimate.

We can also compare the estimates $x_i = x + \Delta x_i$ and $y_i = x + \Delta y_i$ of two (or more) different experts and compute the standard deviation

$$\sigma_{xy} \stackrel{\text{def}}{=} \sqrt{\frac{1}{N} \cdot \sum_{i=1}^{n} (x_i - y_i)^2} = \sqrt{\frac{1}{N} \cdot \sum_{i=1}^{n} (\Delta x_i - \Delta y_i)^2}$$

that describes the *inter-expert variation* of expert estimates.

An interesting empirical fact is that in many situations, the intra-expert and inter-expert variations are practically equal: the difference between the two variations is about 3% [1].

Why does this fact need explanation. At first glance, it may seem that the above fact is very natural and does not need any sophisticated explanation. However, as we show, a deeper analysis makes this fact truly puzzling.

Indeed, the above estimates seem to be informally based on a simple probabilistic model, in which the differences Δx_i are instances of a random variable Δx with 0 mean. The above expression for the intra-expert variance is simply a sample-based estimation of this random variable's standard deviation: $\sigma_x \approx \sigma[\Delta x]$ and thus, $\sigma_x^2 \approx \sigma^2[\Delta x] = E[(\Delta x)^2]$, where, as usual, $E[\eta]$ denotes the expected value of a random quantity η, and $\sigma[\eta]$ denotes its standard deviation.

Similarly, the inter-expert variation is approximately equal to the standard deviation of the difference $\Delta x - \Delta y$ between the random variables Δx and Δy corresponding to two experts: $\sigma_{xy} \approx \sigma[\Delta x - \Delta y]$, i.e., $\sigma_{xy}^2 \approx E[(\Delta x - \Delta y)^2]$.

Thus, the fact that the intra-expert and the inter-expert variations coincide means that $E[(\Delta x - \Delta y)^2] \approx E[(\Delta x)^2] \approx E[(\Delta y)^2]$.

If experts were fully independent, then we would have $E[(\Delta x - \Delta y)^2] = E[(\Delta x)^2] + E[(\Delta y)^2]$, so we would have $\sigma_{xy}^2 \approx 2\sigma_x^2$ and $\sigma_{xy} \approx \sqrt{2} \cdot \sigma_x$, and the inter-expert variation would be at least 40% larger than the intra-expert one.

This we do not observe. It means that there *is* a correlation between the experts. If there was the perfect correlation, we would have $\Delta x_i = \Delta y_i$, and the inter-expert variation would be exactly 0.

In situations of *partial* correlation, we would get all possible values of σ_{xy} ranging from 0 to $\sqrt{2} \cdot \sigma_x$. So why, out of all possible values from interval $[0, \sqrt{2} \cdot \sigma_x]$, the value σ_x corresponds to the average inter-expert variation?

Maximum Entropy technique can help us explain this fact. To provide our explanation, let us express the inter-expert variation in terms of the (Pearson) correlation coefficient $r \stackrel{\text{def}}{=} \dfrac{E[\Delta x \cdot \Delta y]}{\sigma[\Delta x] \cdot \sigma[\Delta y]}$.

By definition of the inter-expert correlation, we have

$$\sigma_{xy}^2 = E[(\Delta x - \Delta y)^2] = E[(\Delta x)^2] + E((\Delta y)^2] - 2E(\Delta x \cdot \Delta y].$$

Here, $E(\Delta x)^2] = E(\Delta y)^2] = \sigma_x^2$, and, by definition of the correlation coefficient, $E[\Delta x \cdot \Delta y] = r \cdot \sigma[\Delta x] \cdot \sigma[\Delta y] = r \cdot \sigma_x^2$. Thus, the above formula for the inter-expert variation takes the form

$$\sigma_{xy}^2 = 2\sigma_x^2 - 2r \cdot \sigma_x^2 = 2 \cdot (1 - r) \cdot \sigma_x^2.$$

In general, the correlation r can take any value from -1 to 1, but in this case, since we assume that all experts are indeed experts, it is reasonable to assume that their estimates are non-negatively correlated, i.e., that $r \geq 0$. Thus, in this example, the set of possible value of the correlation r is the interval $[0, 1]$.

In different situations, we may have different values of the correlation coefficient: some experts may be independent, other pairs of experts may have the same background and thus, have strongly correlated estimates. So, in real life, there will be some probability distribution on the set $[0, 1]$ of all possible values of the correlation coefficient that reflects the frequency of different pairs of experts. We would like to estimate the average value $E[r]$ of r over this distribution. Then, by averaging over r, we will get the desired relation between the intra- and inter-expert variations:

$$\sigma_{xy}^2 = 2 \cdot (1 - E[r]) \cdot \sigma_x^2.$$

We do not have any information about which values r are more probable (i.e., more frequent) and which values r are less probable. In other words, in principle, all probability distributions on the interval $[0, 1]$ are possible. To perform the above estimation, we need to select a single distribution form this class.

It is reasonable to apply the Maximum Entropy technique to select such a distribution. As we have mentioned, in this case, the Maximum Entropy technique selects a uniform distribution on the interval $[0,1]$. For the uniform distribution on the interval $[0,1]$, the probability density is equal to 1, and the mean value is 0.5:

$$E[r] = \int_0^1 x \cdot \rho(x)\,dx = \int_0^1 x\,dx = \frac{x_2}{2}\Big|_0^1 = \frac{1^2}{2} - \frac{0^2}{2} = 0.5.$$

Substituting the value $E[r] = 0.5$ into the above formula $\sigma_{xy}^2 = 2 \cdot (1 - E[r]) \cdot \sigma_x^2$, we conclude that $\sigma_{xy}^2 = \sigma_x^2$, which is exactly the fact that we try to explain.

3 Second Case Study: How Maximum Entropy Techniques Can Be Used to Find a Numerical Value

Empirical fact. It has been observed that when people make crude estimates, their estimates differ by half-order of magnitude; see, e.g., [2]. For example, when people estimate the size of a crowd, they normally give answers like 100, 300, 1000, but it is much more difficult for them to distinguish, e.g., between 100 and 200. Similarly, when describing income, people talk about low six figures, high six figures, etc, – which is exactly half-orders of magnitude.

So, what is so special about the ratio 3 corresponding to half-order of magnitude? Why not 2 or 4?

There are explanations for this fact, but can we have a simpler one? There are explanations for the above fact; see, e.g., [3]. However, these explanations are somewhat complicated.

For a simple fact about commonsense reasoning, it is desirable to have a simpler, more intuitive explanation.

What we do in this section. In this section, we show that the Maximum Entropy technique can be used to provide a simpler explanation for this empirical fact.

Let us formulate this problem in precise terms. Let us assume that we have two quantities a and b, and a is smaller than b. For example, a and b are the salaries of two employees on the two layers of the company's hierarchy. If all we know is that $a < b$, what can we conclude about the relation between a and b?

Applying Maximum Entropy technique: first attempt. Let us try to apply the Maximum Entropy techniques to answer this question. For this purpose, it may sound reasonable to come up with some probability distribution on the set of all possible values of a and on the set of possible values of b. Here, we do not have any bound on a and b. In this case, similar to the case of interval bounds, the Maximum Entropy technique implies that $\rho(x) = \text{const}$ for all possible real numbers x – and thus, since we want $\rho(x) > 0$, we get $\int_0^\infty \rho(x)\,dx = \infty > 1$.

Applying Maximum Entropy technique: second attempt and the resulting explanation. To be able to meaningfully apply the Maximum Entropy idea, we need to consider *bounded* quantities. One such possibility is to consider, instead of the original salary a, the *fraction* of the overall salary $a+b$ that goes to a, i.e., the ratio

$$r \stackrel{\text{def}}{=} \frac{a}{a+b}.$$

We know that $a < b$, so this ratio takes all possible values from 0 to 0.5, where 0.5 corresponds to the ideal case when the salaries a and b are equal. By using the Maximum Entropy technique, we can conclude that the variable r is uniformly distributed on the interval $[0, 0.5)$. Thus, the average value of this variable is at the midpoint of this interval, when $r = 0.25$. So, on average, the salary a of the first person takes 1/4 of the overall amount $a+b$, and thus, the average salary b of the second person is equal to the remaining amount $1 - 1/4 = 3/4$. Thus, the ratio of the two salaries is exactly $\frac{b}{a} = \frac{3/4}{1/4} = 3$.

This corresponds exactly to the half-order of magnitude ratio that we are trying to explain. Thus, the Maximum Entropy technique indeed explains this empirical ratio.

4 Third Case Study: How Maximum Entropy Techniques Can Be Used to Find a Functional Dependence

Often, we need to find a functional dependence. In many practical situations, we know that the value of a quantity x uniquely determines the values of the quantity y, i.e., that $y = f(x)$ for some function $f(x)$.

- In some practical situations, this dependence is known, but
- in other situations, we need to find this dependence.

How the Maximum Entropy technique can help: the main idea. For each physical quantity, we usually know its bounds. Thus, we can safely assume that we know that:

- all possible values of the quantity x are in a known interval $[\underline{x}, \overline{x}]$, and
- all possible values of the quantity y are in a known interval $[\underline{y}, \overline{y}]$.

If we apply the Maximum Entropy technique to the quantity x, we conclude that x is uniformly distributed on the interval $[\underline{x}, \overline{x}]$. Similarly, if we apply the Maximum Entropy technique to the quantity y, we conclude that x is uniformly distributed on the interval $[\underline{y}, \overline{y}]$.

It is therefore reasonable to select a function $f(x)$ for which,

- when x is uniformly distributed on the interval $[\underline{x}, \overline{x}]$,
- the quantity $y = f(x)$ is uniformly distributed on the interval $[\underline{y}, \overline{y}]$.

What are the resulting functional dependencies? For a uniform distribution, the probability to be in an interval is proportional to its length. In particular, for a small interval $[x, x+\Delta]$ of width Δx, the probability to be in this interval is equal to $\rho_x \cdot \Delta x$.

The corresponding y-interval $[f(x), f(x+\Delta x)]$ has width

$$\Delta y = |f(x+\Delta x) - f(x)|.$$

For small Δx, we have

$$\frac{f(x+\Delta x)-f(x)}{\Delta x} \approx \lim_{h\to 0} \frac{f(x+h)-f(x)}{h} = f'(x).$$

Thus, for small Δx, we have $f(x+\Delta x) - f(x) \approx f'(x) \cdot \Delta x$ and therefore, $\Delta y \approx |f'(x)| \cdot \Delta x$. Since the variable y is also uniformly distributed, the probability for y to be in this interval is equal to $\rho_y \cdot \Delta y = \rho_y \cdot |f'(x)| \cdot \Delta x$.

Comparing this expression with the original formula $\rho_x \cdot \Delta x$ for the same probability, we conclude that $\rho_y \cdot |f'(x)| \cdot \Delta x = \rho_x \cdot \Delta x$, hence $|f'(x)| = \dfrac{\rho_x}{\rho_y}$, i.e., $|f'(x)| =$ const. So, we conclude that *the function $f(x)$ should be linear*.

What is our result and why it is interesting. Our conclusion is that if we have no information about the functional dependence, it is reasonable to assume that this dependence is linear.

This fits well with the usual engineering practice, where indeed the first idea is usually to try a linear dependence. However, the usual motivation for using a linear dependence first is that such a dependence is the easiest to analyze – and why would nature care which dependencies are easier for us to analyze? The Maximum Entropy argument seems more convincing, since it relies on the general ideas about uncertainty itself – and not on our ability to deal with this uncertainty.

Need for nonlinear dependencies. That we came up with an explanation for a linear dependence may be nice, but in practice, linear dependence is usually only the first approximation to the true non-linear dependence. Once we know that the a linear dependence is only an approximation, we would like to find a more adequate nonlinear model.

The Maximum Entropy technique can help beyond linear dependencies. It turns out that the Maximum Entropy technique can also help in finding such a nonlinear dependence – just like for probability distributions:

- once we have an additional information which is not consistent with the assumption that the actual distribution is uniform,
- we can add this information to the corresponding Maximum Entropy problem and get a non-uniform distribution consistent with this information.

We will actually describe two alternative ideas on in which the Maximum Entropy technique can help.

The Maximum Entropy technique can help beyond linear dependencies: first idea. The first, more direct, idea is to take into account that often, not only the

quantity y, but also its derivative $z \stackrel{\text{def}}{=} \dfrac{dy}{dx}$ (and sometimes, its second derivative as well) is also an observable quantity. For example, when y is a distance and x is time, then the first derivative $v \stackrel{\text{def}}{=} \dfrac{dy}{dx}$ is velocity and the second derivative $a \stackrel{\text{def}}{=} \dfrac{dv}{dx} = \dfrac{d^2y}{dx^2}$ is acceleration – both perfectly observable quantities.

If we apply the Maximum Entropy techniques to the dependence of velocity v on time x, we conclude that the velocity linearly depends on time – in which case, by integrating this dependence, we conclude that the distance is a quadratic function of time. Similarly, if we apply the Maximum Entropy technique to the dependence of acceleration a on time, then we conclude that the velocity is a quadratic function of time, and thus, that the distance is a cubic function of time.

The Maximum Entropy technique can help beyond linear dependencies: second idea. The second, less direct idea, is to take into account that when the dependence $y = f(x)$ is non-linear, then, even when the probability distribution for x is uniform, with density $\rho_x(x) = \rho_x = \text{const}$, the corresponding probability distribution $\rho_y(y)$ for the quantity y is, in general, *not* uniform.

How can we describe the dependence $\rho_y(y)$ of the probability density on y? To describe this auxiliary dependence, we can use the Maximum Entropy technique and conclude that this dependence is linear, i.e., that $\rho_y(y) = a + b \cdot y$. Now that we know the distributions for x and y, we can look for functions $f(x)$ for which:

- once x is uniformly distributed,
- the quantity $y = f(x)$ is distributed with the probability density $\rho_y(y) = a + b \cdot y$.

Similarly to the above case when both x- and y-distributions were uniform, the probability of being in the x-interval of width Δx is equal to $\rho_x \cdot \Delta x$, and on the other hand, it is equal to $\rho_y(y) \cdot |f'(x)| \cdot \Delta x = (a + b \cdot f(x)) \cdot |f'(x)| \cdot \Delta x$. By comparing these two expressions for the same probability, we conclude that

$$|f'(x)| \cdot (a + b \cdot f(x)) = \text{const},$$

i.e., that $\dfrac{df}{dx} \cdot (a + b \cdot f) = \text{const}$. By moving all the terms containing f to one side and all the terms containing x to another sides, we conclude that $\dfrac{df}{a + b \cdot f} = \text{const} \cdot x$. So, for $g \stackrel{\text{def}}{=} f + \dfrac{a}{b}$, we get $\dfrac{dg}{g} = c \cdot dx$. Integration leads to $\ln(g) = c \cdot x + C$ for some integration constant C, thus, $g = A \cdot \exp(c\dot{x})$, and $f = A \cdot \exp(c \cdot x) + \text{const}$.

By assuming that y is uniformly distributed, we get the inverse (logarithmic) dependence. By assuming that the dependence $\rho_y(y)$ on y is not linear but is described by one of these nonlinear formulas, we can get an even more complex dependence.

Thus, we can indeed use the Maximum Entropy technique to describe nonlinear dependencies as well.

Acknowledgments

This work was supported in part by the National Science Foundation grant HRD-1242122 (Cyber-ShARE Center of Excellence).

References

1. J. Garibaldi, "Type-2 Beyond the Centroid", *Proceedings of the IEEE International Confreence on Fuzzy Systems FUZZ-IEEE'2017*, Naples, Italy, July 8–12, 2017.
2. J. R. Hobbs, "Half orders of magnitude", In: L. Obrst and I. Mani (eds.), *Proceeding of the Workshop on Semantic Approximation, Granularity, and Vagueness, A Workshop of the Seventh International Conference on Principles of Knowledge Representation and Reasoning KR'2000*, Breckenridge, Colorado, April 11, 2000, pp. 28–38.
3. J. Hobbs and V. Kreinovich, "Optimal Choice of Granularity In Commonsense Estimation: Why Half-Orders of Magnitude", *International Journal of Intelligent Systems*, 2006, Vol. 21, No. 8, pp. 843–855.
4. E. T. Jaynes and G. L. Bretthorst, *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, UK, 2003.
5. H. T. Nguyen, V. Kreinovich, B. Wu, and G. Xiang, *Computing Statistics under Interval and Fuzzy Uncertainty*, Springer Verlag, Berlin, Heidelberg, 2012.
6. S. G. Rabinovich, *Measurement Errors and Uncertainty: Theory and Practice*, Springer Verlag, Berlin, 2005.
7. D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman and Hall/CRC, Boca Raton, Florida, 2011.