Journal of Innovative Technology and Education, Vol. 3, 2016, no. 1, 1 - 7
HIKARI Ltd, www.m-hikari.com
http://dx.doi.org/10.12988/jite.2016.51017

Maximum Entropy Approach is not Arbitrary as it May Seem at First Glance

Olga Kosheleva and Vladik Kreinovich

University of Texas at El Paso
500 W. University
El Paso, TX 79968, USA

Abstract

When we only have partial information about the probability distribution, i.e., when several different probability distributions are consistent with our knowledge, then it makes sense to select a distribution with the largest entropy. In particular, when we only know that the quantity is located within a certain interval – and we have no information about the probability of different values within this intervals – then it is reasonable to assume that all these values are equally probable, i.e., that we have a uniform distribution on this interval. The problem with this idea is that if we apply it to the same quantity after a non-linear rescaling, we get a different (non-uniform) distribution in the original scale. In other words, it seems that the results of applying the Maximum Entropy approach are rather arbitrary: they depend on what exactly scale we apply them to. In this paper, we show how to overcome this subjectivity: namely, we propose to take into account that, due to measurement inaccuracy, we always have finitely many possible measurement results, and this finiteness makes the results of applying the Maximum Entropy approach uniquely determined.

Mathematics Subject Classification: 62P99, 94A17, 91B16

Keywords: maximum entropy approach, re-scaling

1 Maximum Entropy Approach and Its Limitations

Need to describe probabilities. One of the main objectives of science is to predict future events based on the available information. In many practical situations, it is not possible to uniquely predict the future events: there are many factors which are difficult to take into account. For example, while we can predict tomorrow's weather reasonably well, these predictions are not exact. In such situations, when we know that for the same future quantity, several different values are possible, it is desirable to describe the frequency of different possible values, i.e., to describe the probability distribution on the set of all possible values.

Predictions are based on the known current and past values of different quantities. These values come from measurement and are also usually only approximate. We usually know the range of possible values of the measurement error, and we would like to know the probability of different possible values of this error; see, e.g., [5].

Often, we only have partial information about probabilities. In many practical situations, we only have partial information about the probabilities. For example, we may only know the range of possible values, but we have no information about the probabilities of different values within this range. Sometimes, we know the first moments of the probability distribution, but we do not know its shape, etc.

In all these cases, there are several different probability distributions which are consistent with our knowledge.

It is often important to select a single probability distribution. According to the decision theory, a rational (consistent) decision makes should select a decision with the largest value of the expected utility. If we know the probabilities of different outcomes, then this expected utility is with respect to these probabilities; if we do not have full information about these probabilities, this means that the decision of the rational decision maker corresponds to some "subjective probability" distribution [1, 3, 4, 6].

So, to make a rational decision, out of all probability distributions which are consistent with our knowledge, we must select a single one that will be used for decision making. How can we select such a distribution?

How to select a single probability distribution: from Laplace's Indeterminacy Principle to Maximum Entropy. To decide how to select a single probability distribution let us start with the simplest case when we have finitely many (n) alternatives, and we have no information about the probabilities of different alternatives.

This situation does not change if we swap two alternatives i and j. Thus, it makes to require that the selected probabilities p_i and p_j should also not change after this swap, i.e., that we should have $p_i = p_j$. So, it is reasonable to select probabilities for which $p_i = p_j$ for all i and j, i.e., in which all n alternatives have the exact same probability. Since these probabilities should add up to 1, they must be equal to $p_i = \dfrac{1}{n}$.

This natural idea – that if we have no reason to assume that one of the alternatives is more probable, then we assume that they are all equally probable – is known as *Laplace's Indeterminacy Principle*.

In the continuous case, a similar idea has been formalized as the *Maximum Entropy approach*, according to which, out of all possible probability distributions with different probability density functions $\rho(x)$, we should select the one with the largest value of the entropy $S \stackrel{\text{def}}{=} -\int \rho(x) \cdot \ln(\rho(x))\, dx$ [2].

Maximum entropy approach: successes and limitations. The Maximum Entropy approach has been successfully applied to many application areas; see, e.g., [2].

However, this approach has a serious limitation. Let us illustrate it on the example of velocity v. Let us assume that the only information that we know about the velocity is that it is somewhere within the interval $[\underline{v}, \overline{v}]$. In this case, the maximum entropy selects a uniform probability distribution on this interval, with the probability density $\rho(v) = \dfrac{1}{\overline{v} - \underline{v}}$. So far, so good.

However, let us consider the same situation from a different viewpoint. Suppose that we are interested in the kinetic energy $E = \dfrac{1}{2} \cdot m \cdot v^2$. The fact that v is between $\underline{v}$ and $\overline{v}$ means that the kinetic energy belongs to the interval $[\underline{E}, \overline{E}]$, where $\underline{E} = \dfrac{1}{2} \cdot m \cdot (\underline{v})^2$ and $\overline{E} = \dfrac{1}{2} \cdot m \cdot (\overline{v})^2$. Since this is the only information that we have about energy, it is reasonable to apply the Maximum Entropy approach and to conclude that energy E is uniformly distributed on this interval, with probability density $\rho(E) = \dfrac{1}{\overline{E} - \underline{E}}$.

This also sounds reasonable, but the problem is that if the velocity v is uniformly distributed, then the corresponding kinetic energy $E = \dfrac{1}{2} \cdot m \cdot v^2$ is *not* uniformly distributed. Indeed, for every E, the corresponding velocity is equal to $v = \sqrt{\dfrac{2E}{m}}$. Thus, the probability $\rho(E) \cdot \Delta E$ to have the energy between E and $E + \Delta E$ is equal to the probability to have velocity between $\sqrt{\dfrac{2E}{m}}$ and $\sqrt{\dfrac{2E + \Delta E}{m}}$. Since the probability distribution on velocities is uniform, this probability is proportional to the width $\sqrt{\dfrac{2E + \Delta E}{m}} - \sqrt{\dfrac{2E}{m}}$ of this

velocity interval. For small E, this width is proportional to the derivative of the function $v(E)$, i.e., is equal to $\sqrt{\frac{2}{m}} \cdot \frac{1}{2\sqrt{E}} \cdot \Delta E$. Thus, this probability – which is equal to $\rho(E) \cdot \Delta E$ – is equal to $\frac{1}{\overline{v} - \underline{v}} \cdot \sqrt{\frac{1}{2m}} \cdot \frac{1}{\sqrt{E}} \cdot \Delta E$ and thus, $\rho(E) = \frac{1}{\overline{v} - \underline{v}} \cdot \sqrt{\frac{1}{2m}} \cdot \frac{1}{\sqrt{E}}$. One can easily see that this is *not* a uniform distribution.

So, if we apply the Maximum Entropy approach to velocities, we get a uniform distribution on velocities, but *not* for kinetic energy. Similarly, if we apply the Maximum Entropy approach to kinetic energy, we get a uniform distribution on energy, but not on velocities. It start sounding as if Maximum Entropy approach is *subjective*: its result changes depending on which of the related quantities we apply it to.

What we do in this paper. In this paper, we explain how to overcome the above subjectivity.

2 How to Eliminate Arbitrariness When Applying the Maximum Entropy Approach

Our main idea. The problem with the Maximum Entropy approach comes from the fact that we assume that the corresponding quantity can take infinitely many possible values x – namely, all real numbers x from the corresponding interval $[\underline{x}, \overline{x}]$ can be possible values of this quantity. In the case when we only have finitely many alternatives, there is no arbitrariness, the probabilities are uniquely determined.

Good news is that the above assumption about infinitely many possible values is an approximation to the real-life situation – an approximation that is intended to make our analysis easier. In real life, we do not observe infinitely many different possible values of the quantity x: the only information that we get comes from measurements; measurement results in a finite string of symbols, there are only finitely many such strings, so there are only finitely many measurement results.

We will show that if we take this finiteness into account, then the above arbitrariness disappears. Let us illustrate this idea on two examples.

First example: measurements with the same absolute measurement error. Let us first consider the case when all the measurements have the same absolute measurement error Δ. In this case, once we know the measurement result $\widetilde{x}$, we can conclude that the actual (unknown) value of the corresponding quantity x is within the interval $[\widetilde{x} - \Delta, \widetilde{x} + \Delta]$.

If for two different measurement results $\widetilde{x}_1$ and $\widetilde{x}_2$, the corresponding intervals $[\widetilde{x}_1-\Delta,\widetilde{x}_1+\Delta]$ and $[\widetilde{x}_2-\Delta,\widetilde{x}_2+\Delta]$, this means that the measurement results $\widetilde{x}_1$ and $\widetilde{x}_2$ may describe the same actual value of the corresponding quantity. Once the value $\widetilde{x}_1$ is fixed, the smallest value $\widetilde{x}_2>\widetilde{x}_1$ which is guaranteed to describe a different actual value x is the smallest value $\widetilde{x}_2>\widetilde{x}_1$ for which the corresponding intervals do not have a non-point intersection, i.e., the value for which $\widetilde{x}_2-\Delta=\widetilde{x}_1+\Delta$ and thus, for which $\widetilde{x}_2=\widetilde{x}_1+2\Delta$. Similarly, we have $\widetilde{x}_3=\widetilde{x}_2+2\Delta=\widetilde{x}_1+2\cdot(2\Delta)$, and, in general, $\widetilde{x}_k=\widetilde{x}_1+(k-1)\cdot(2\Delta)$.

In accordance with Laplace's Indeterminacy Principle, these values $\widetilde{x}_k$ are equally probable. In the limit $\Delta\to 0$, we thus get a uniform distribution on the original interval $[\underline{x},\overline{x}]$.

How does this solve our problem? At first glance, we get the exact same result – the uniform distribution. So how does this help us? It does help, because this result is based on the assumption that all the measurement have the same absolute measurement accuracy.

If the measurements of velocity have the same absolute measurement accuracy, then we get the uniform distribution for velocities. However, when the widths of all the intervals $[\widetilde{v}-\Delta,\widetilde{v}+\Delta]$ are the same, for the energy $E=\dfrac{1}{2}\cdot m\cdot v^2$, the corresponding intervals $\left[\dfrac{1}{2}\cdot m\cdot(\widetilde{v}-\Delta)^2,\dfrac{1}{2}\cdot m\cdot(\widetilde{v}+\Delta)^2\right]$ have *different* widths – meaning that the corresponding indirect measurements of energy have different absolute errors. Thus, the above argument justifying the unform distribution for velocity cannot be applied to energy, and we do not make a confusing conclusion that the energy distribution should also be uniform.

Vice versa, it is possible that all the measurements of kinetic energy have the same absolute accuracy. In this case, the energy distribution is uniform, but, in this case, the corresponding velocity interval widths are all different – and thus, we can no longer conclude that the velocity distribution is uniform.

Second example: measurements with the same relative measurement error. Let us now consider another realistic case, when all the measurements have the same relative measurement error δ. In this case, once we know the measurement result $\widetilde{x}$, we can conclude that the actual (unknown) value of the corresponding quantity x is within the interval $[\widetilde{x}\cdot(1-\delta),\widetilde{x}\cdot(1+\delta)]$.

If for two different measurement results $\widetilde{x}_1$ and $\widetilde{x}_2$, the corresponding intervals $[\widetilde{x}_1\cdot(1-\delta),\widetilde{x}_1\cdot(1+\delta)]$ and $[\widetilde{x}_2\cdot(1-\delta),\widetilde{x}_2\cdot(1+\delta)]$, this means that the measurement results $\widetilde{x}_1$ and $\widetilde{x}_2$ may describe the same actual value of the corresponding quantity. Once the value $\widetilde{x}_1$ is fixed, the smallest value $\widetilde{x}_2>\widetilde{x}_1$ which describes a different actual value is the smallest value for which these intervals do not have a non-point intersection, i.e., the value for which

$\widetilde{x}_2 \cdot (1-\delta) = \widetilde{x}_1 \cdot (1+\delta)$ and thus, for which $\widetilde{x}_2 = \widetilde{x}_1 \cdot q$, where $q \stackrel{\text{def}}{=} \dfrac{1+\delta}{1-\delta}$. Similarly, we have $\widetilde{x}_3 = q \cdot \widetilde{x}_2 = q^2 \cdot \widetilde{x}_1$, and, in general, $\widetilde{x}_k = q^{k-1} \cdot \widetilde{x}_1$.

Here, the values $X_k \stackrel{\text{def}}{=} \ln(\widetilde{x}_k)$ satisfy the condition $X_i = X_k + (k-1) \cdot \ln(q)$. In accordance with Laplace's Indeterminacy Principle, the values $\widetilde{x}_k$ (and thus, the values X_k) are equally probable. In the limit $\Delta \to 0$, we thus get a uniform distribution in the logarithm scale, on the interval $[\underline{X}, \overline{X}]$, where $\underline{X} = \ln(\underline{x})$ and $\overline{X} = \ln(\overline{x})$ — and thus, *not* a uniform distribution in the original x-scale.

Conclusion. We have just given two examples. Similar computations can be repeated for more complex cases, e.g., when the measurement error consists of an absolute and a relative error, i.e., when the corresponding interval of possible values of x has the form $[\widetilde{x} \cdot (1-\delta) - \Delta, \widetilde{x} \cdot (1+\delta) + \Delta]$. In all these cases, our discrete analysis leads to a unique Maximum Entropy-motivated distribution, with no arbitrariness.

We can therefore conclude that, if we take the actual discreteness into account, then the Maximum Entropy approach is indeed not as arbitrary as it may seem at first glance.

Acknowledgments

This work was supported in part by the US National Science Foundation grants HRD-0734825, HRD-1242122, and DUE-0926721.

The authors are thankful to Scott Ferson for valuable discussions.

References

[1] P. C. Fishburn, *Utility Theory for Decision Making*, John Wiley and Sons, New York, 1969.

[2] E. T. Jaynes and G. L. Bretthorst, *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, UK, 2003.

[3] D. R. Luce and H. Raiffa, *Games and Decisions, Introduction and Critical Survey*, John Wiley and Sons, New York, 1957.

[4] R. B. Myerson, *Game Theory: Analysis of Conflict*, Harvard University Press, Cambridge, Massachusetts, 1991.

[5] S. G. Rabinovich, *Measurement Errors and Uncertainty, Theory and Practice*, Springer Verlag, Berlin, 2005.
http://dx.doi.org/10.1007/0-387-29143-1

[6] P. Suppes, D. M. Krantz, R. D. Luce, and A. Tversky, *Foundations of Measurement Vol. II: Geometrical, Threshold, and Probabilistic Representations*, Academic Press, San Diego, California, 1989.

Received: November 15, 2015; Published: January 2, 2016