What Is a Natural Probability Distribution on the Class of All Continuous Functions: Maximum Entropy Approach Leads to Wiener Measure

Vladik Kreinovich and Saeid Tizpaz Niari

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

1. Need to select a probability distribution under uncertainty

- Usually, we do not have full knowledge of the situation, we only have a partial knowledge.
- So, several different states of the studied system are possible.
- Usual data processing techniques assume that:
  - we know the probabilities of different states,
  - i.e., in mathematical terms, that we know the probability distribution of the set of possible states.
- However, in practice, we often do not have full knowledge of these probabilities.
- So, there are several probability distributions consistent with our knowledge.

2. Need to select a probability distribution under uncertainty (cont-d)

- For example, often, for some quantities x:
  - we only know the interval of possible values $[\underline{x}, \overline{x}]$,
  - but we do not have any information about the probabilities of different values from this interval.
- To apply the usual data processing technique to such situations, we need to select one of the possible probability distributions.

3. How can we make such a selection: Maximum Entropy approach

- If all we know is the interval of possible values, then:
  - since there is no reason to assume that some values are more probable than others,
  - it is reasonable to select a distribution in which all these values are equally probable,
  - i.e., the uniform distribution on this interval.
- This idea goes back to Laplace and is therefore known as *Laplace Indeterminacy Principle*.
- More generally, a natural idea is not to introduce certainty where there is none.

4. How can we make such a selection: Maximum Entropy approach (cont-d)

- For example:
  - if all we know is that the actual value of the quantity x is located on the interval $[\underline{x}, \overline{x}]$,
  - then one of the possible probability distributions is when x is equal to $\underline{x}$ with probability 1.
- However:
  - if we select this distribution as a representation of the original interval uncertainty,
  - we would thus introduce certainty where there is none.
- To describe this idea in precise terms, we need to formally describe what we mean by certainty.

5. How can we make such a selection: Maximum Entropy approach (cont-d)

- A natural measure of uncertainty is the number of binary ("yes"-"no") questions that we need to ask:
  - to determine the exact value of the desired quantity,
  - or, alternatively, to determine this value with some given accuracy $\varepsilon > 0$.
- It is known that this number of questions can be gauged by Shannon's entropy $S = -\int \rho(x) \cdot \ln(\rho(x))\, dx$.
- Here, $\rho(x)$ is the probability density.

6. How can we make such a selection: Maximum Entropy approach (cont-d)

- In these terms, the requirement not to add certainty – i.e., equivalently, not to decrease uncertainty – means that:
  - we should not select a distribution
  - if there is another possible distribution with higher uncertainty.
- In other words, out of all possible distributions, we should select the one for which the entropy attains its maximal value.
- This idea is known as the *Maximum Entropy approach.*

7. Successes of the Maximum Entropy approach

- The Maximum Entropy approach has many successful applications; let us give a few examples.
- Suppose that all we know about a quantity x is that its value is located somewhere in the interval $[\underline{x}, \overline{x}]$.
- We have no information about the probabilities of different values within this interval.
- Then the Maximum Entropy approach leads to the uniform distribution on this interval.
- This is in full accordance with the Laplace Indeterminacy Principle.

8. Successes of the Maximum Entropy approach (cont-d)

- Suppose that we know the probability distributions of two quantities x and y.
- However, we do not have any information about the possible correlation between these two quantities.
- This correlation could be positive, could be negative.
- In this case, the Maximum Entropy approach leads to the conclusion that the variables x and y are independent.
- This is also in good accordance with common sense:
  - if we have no information about the dependence between the two quantities,
  - it makes sense to assume that they are independent.

9. Successes of the Maximum Entropy approach (cont-d)

- Finally, suppose that all we know is the mean and standard deviation of a random variable.
- Then, out of all possible distributions with these two values, the Maximum Entropy approach selects the Gaussian (normal) distribution.
- This selection also makes perfect sense.
- Indeed, in most practical situations, there are many different factors contributing to randomness.
- It is known that under reasonable conditions, the joint effect of many factors leads to the normal distribution.
- The corresponding mathematical result is known as the Central Limit Theorem.

10. Remaining challenge

- Usually, the Maximum Entropy approach is applied to situations when:
  - the state of the system is described by finitely many quantities $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$, and
  - the question is to select a probability distribution on the set of all possible tuples $x = \left(x^{(1)}, x^{(2)}, \ldots, x^{(n)}\right)$.
- In some practical situations, however, what is not fully known is also a function.
- E.g., the function that describes how the system's output is related to the system's input.
- In most cases, this dependence is continuous – small changes in x cause small changes in $y = f(x)$.

11. Remaining challenge (cont-d)

- In such situations, often:
  - we know a class of possible functions $f(x)$,
  - but we do not have a full information about the probabilities of different functions from this class,
  - and in many cases, we have no information at all about these probabilities.
- In such cases, several different probability distributions on the class of all continuous functions are consistent with our knowledge.
- It is desirable to select a single probability distribution from this class.
- How can we do it?

12. What we do in this talk

- In this talk, we show that the Maximum Entropy approach can help to solve this challenge.
- Namely, this approach leads to the selection of the Wiener measure, a probability distribution that describes the Brownian motion.
- In this distribution:
  - the joint distribution of any combination of values $f(x)$, $f(y)$, $\ldots$, is Gaussian, and
  - for some constant $C > 0$, we have

$$E[f(x)] = 0 \text{ and } E[(f(x) - f(y))^2] = C \cdot |x - y| \text{ for all } x \text{ and } y.$$

- Here $E[\cdot]$, as usual, means the expected value.
- Let us show how this result can be derived.

13. From the practical viewpoint, all the values are discrete

- From the purely mathematical viewpoint:
  - there are infinitely many real numbers, and
  - thus, there are infinitely many possible values of the input quantity x.
- However, from the practical viewpoint, at each moment of time, there is a limit $\varepsilon > 0$ on the accuracy with which we can measures x.
- As a result, we have a discrete set of distinguishable values of the quantity x.
- E.g., the values $x_0 = 0$, $x_1 = x_0 + \varepsilon = \varepsilon$, $x_2 = x_1 + \varepsilon = 2\varepsilon$, and, in general, $x_k = k \cdot \varepsilon$.
- In this description, to describe a function $y = f(x)$ means to describe its values $y_k = f(x_k)$ on these values x_k.

14. In these terms, what does continuity mean?

- In the discretized description, the continuity of the dependence $y = f(x)$ means, in effect, that:
  - the value y_{k-1} and y_k corresponding to close values of x
  - cannot differ by more than some small constant δ:

$$|\Delta_k| \le \delta, \text{ where } \Delta_k \stackrel{\text{def}}{=} y_k - y_{k-1}.$$

- Thus:
  - to describe a probability distribution on the class of all continuous functions
  - means describing a probability distribution on the set of all the tuples $y = (y_0, y_1, \ldots)$.

15. Now, we are ready to apply the Maximum Entropy approach

- Once know the value y_0 – e.g., we know that $y_0 = 0$ – and we know the differences Δ_k, we can easily reconstruct the values y_k:

$$y_k = y_0 + \Delta_1 + \Delta_2 + \ldots + \Delta_k.$$

- So:
  - to describe the probability distribution on the set of all possible tuples,
  - it is sufficient to describe the distribution on the set of all possible tuples of differences $\Delta = (\Delta_1, \Delta_2, \ldots)$.
- To find this distribution, let us apply the Maximum Entropy approach.
- In our description, the only thing that we know about the values Δ_k is that each of these values is located somewhere in the interval $[-\delta, \delta]$.

16. Now, we are ready to apply the Maximum Entropy approach (cont-d)

- Thus, as we have mentioned earlier, the Maximum Entropy approach implies:
  - that each difference Δ_k is uniformly distributed on this interval, and
  - that the differences Δ_k and Δ_ℓ corresponding to $k \neq \ell$ are independent.

17. This indeed leads to the Wiener measure

- For each variable Δ_k, its mean is 0, and its variance is equal to $\delta^2/3$.
- For the sum of independent random variables:
  - the mean is equal to the sum of the means, and
  - the variance is equal to the sum of the variances.
- Thus, we conclude that for each quantity y_k, its mean $E[y_k]$ and variance $V[y_k]$ are equal to:

$$E[y_k] = 0, \quad V[y_k] = k \cdot \frac{\delta^2}{3}.$$

- Let us reformulate this expression in terms of the values $f(x)$.
- The value y_k denotes $f(x)$ for $x = k \cdot \varepsilon$, so $k = \dfrac{x}{\varepsilon}$.
- Thus, the above formula implies that

$$E[f(x)] = 0 \text{ and } V[f(x)] = C \cdot x, \text{ where } C \stackrel{\text{def}}{=} \frac{\delta^2}{3\varepsilon^2}.$$

18. This indeed leads to the Wiener measure (cont-d)

- Similarly, for the difference $y_k - y_\ell = \Delta_{k+1} + \Delta_{k+2} + \ldots + \Delta_\ell$, we conclude that $E[(y_k - y_\ell)^2] = |\ell - k| \cdot \frac{\delta^2}{3}$.
- Thus, for every two values x and y, we have

$$E[(f(x) - f(y))^2] = C \cdot |x - y|.$$

- These are exactly the formulas for the Wiener measure.
- To complete our conclusion, we need to show that:
  - the probability distribution of each value $f(x)$ is Gaussian, and
  - the joint distribution of values $f(x)$, $f(y)$, etc., is Gaussian.
- Indeed, according to the above formula, each value $f(x) = y_k$ is the sum of $k = x/\varepsilon$ independent random variables Δ_k.

19. This indeed leads to the Wiener measure (cont-d)

- Here:
  - As our measurements become more and more accurate, i.e., as ε tends to 0,
  - the number of such terms tends to infinity.
- Thus, according to the above-mentioned Central Limit Theorem, the distribution of $f(x)$ tends to Gaussian.
- Similarly, we can show that the joint distribution of several values $f(x)$, $f(y)$, $\ldots$, is also Gaussian.
- So, we can indeed conclude that the selected probability distribution on the set of all continuous functions is the Wiener measure.

20. Acknowledgments

- This work was supported in part by the National Science Foundation grants:
  - 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and
  - HRD-1834620 and HRD-2034030 (CAHSI Includes).
- It was also supported by the AT&T Fellowship in Information Technology.
- It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478.