Maximum Entropy Beyond Selecting Probability Distributions

Thach N. Nguyen[1], Olga Kosheleva[2], and Vladik Kreinovich[2]

[1]Banking University of Ho Chi Minh City, Vietnam
Thachnn@buh.edu.vn
[2]University of Texas at El Paso, El Paso, Texas 79968, USA
vladik@utep.edu, olgak@utep.edu

1. 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, etc.
- Often, however, we have only partial information about a probability distribution.
- Then, several probability distributions are consistent with the available knowledge.
- We want to apply, to this situation:
  - a data processing algorithm
  - which is based on the assumption that the probability distribution is known.

2. Need to Select a Distribution (cont-d)

- We want to apply, to this situation:
  - a data processing algorithm
  - which is based on the assumption that the probability distribution is known.
- For this, we must select a single probability distribution out of all possible distributions.
- How can we select such a distribution?

3. Maximum Entropy Approach

- 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.
- 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.
- In the case of continuous variables, to determine the random value with a given accuracy ε.

4. Maximum Entropy Approach (cont-d)

- One can show that this average number is asymptotically (when $\varepsilon \to 0$) proportional to the *entropy*

$$S(\rho) \stackrel{\text{def}}{=} -\int \rho(x) \cdot \ln(\rho(x))\, dx.$$

- For a class F of distributions, the average number of binary question is asymptotically proportional to

$$\max_{\rho \in F} S(\rho).$$

- If we select a distribution, uncertainty decreases.
- We want to select a distribution ρ_0 for which the decrease in uncertainty is the smallest.
- We thus select a distribution ρ_0 for which the entropy is the largest possible: $S(\rho_0) = \max\limits_{\rho \in F} S(\rho)$.

5. 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 $[a, b]$.
- We then maximize $-\int_a^b \rho(x)\cdot\ln(\rho(x))\,dx$ under the condition 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.
- This method reduces our problem to the following unconstrained optimization problem:

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

- Here λ is the *Lagrange multiplier*.

6. Simple Examples of Using the Maximum Entropy Techniques (cont-d)

- The value λ 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.

7. Simple Examples of Using the Maximum Entropy Techniques (cont-d)

- Therefore, $\rho(x)$ 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;
  - it is thus reasonable to conclude that all these values are equally probable, i.e., that $\rho(x) = \text{const}$.
- This idea goes back to Laplace and is known as the *Laplace Indeterminacy Principle*.
- In other situations, the only information that we have about $\rho(x)$ is the first two moments

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

8. Simple Examples of Using the Maximum Entropy Techniques (cont-d)

- Then, we select $\rho(x)$ for which $S(\rho)$ is the largest under these two constraints and $\int \rho(x)\,dx = 1$.
- For this problem, the Lagrange multiplier methods leads to maximizing:

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

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

- Differentiating w.r.t. $\rho(x)$ and equating the derivative to 0, we conclude that

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

- So, $\ln(\rho(x))$ is a quadratic function of x and thus, $\rho(x) = \exp(\ln(\rho(x)))$ is a Gaussian distribution.

9. Simple Examples of Using the Maximum Entropy Techniques (final)

- This conclusion is also in good accordance with common sense; indeed:
  - in many cases, e.g., the measurement error results from many independent small effects and,
  - according to the Central Limit Theorem, the distribution of such sum is close to Gaussian.
- There are many other examples of a successful use of the Maximum Entropy technique.

10. A Natural Question

- Rhe Maximum Entropy technique works well for selecting a distribution.
- Can we extend it to solving other problems?
- In this talk, 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.

11. Fact to Explain

- Experts' estimates are imprecise – just like measuring instruments are imprecise.
- When we ask the expert after some time to estimate the same quantity, we get give a slightly different value.
- We can describe the expert's estimates x_i of 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 estimation error:

$$\sigma_x \stackrel{\text{def}}{=} \sqrt{\frac{1}{N} \cdot \sum_{i=1}^{n} (\Delta x_i)^2}.$$

- This quantity describes the *intra-expert* variation of the expert estimate.

12. Fact to Explain (cont-d)

- We can also compare the estimates $x_i = x + \Delta x_i$ and $y_i = x + \Delta y_i$ of two experts:

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

- This value 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%.
- Let us show that this fact is puzzling.

13. Fact to Explain (cont-d)

- Indeed, 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 $E[(\Delta x - \Delta y)^2] = E[(\Delta x)^2] + E[(\Delta y)^2]$ hence $\sigma_{xy}^2 \approx 2\sigma_x^2$.
- This we do not observe, so there *is* a correlation between the experts.
- If there was the perfect correlation, we would have $\Delta x_i = \Delta y_i$, and $\sigma_{xy} = 0$.
- 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_{xy} corresponds to σ_x?

14. Maximum Entropy Technique Can Explain This Fact

- 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 $\sigma_{xy}^2 = E[(\Delta x - \Delta y)^2] = E[(\Delta x)^2] + E((\Delta y)^2] - 2E(\Delta x \cdot \Delta y]$.
- $E(\Delta x)^2] = E(\Delta y)^2] = \sigma_x^2$, and, by definition of r:
$$E[\Delta x \cdot \Delta y] = r \cdot \sigma[\Delta x] \cdot \sigma[\Delta y] = r \cdot \sigma_x^2.$$
- Thus, $\sigma_{xy}^2 = 2\sigma_x^2 - 2r \cdot \sigma_x^2 = 2 \cdot (1 - r) \cdot \sigma_x^2$.
- In general, the correlation r can take any value from -1 to 1.
- We assumed that all experts are indeed experts.
- It is thus reasonable to assume that their estimates are non-negatively correlated, i.e., that $r \geq 0$.

15. Maximum Entropy Technique Can Explain This Fact (cont-d)

- 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 correlation.
- So, in real life, there will be some probability distribution on the set $[0, 1]$ of all possible values of r.
- We would like to estimate the average value $E[r]$ of r over this distribution.

16. Maximum Entropy Technique Can Explain This Fact (cont-d)

- 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).
- 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 earlier, in this case, the Maximum Entropy technique selects a uniform distribution on $[0, 1]$.

17. Maximum Entropy Technique Can Explain This Fact (final)

- *Reminder:* $\sigma_{xy}^2 = 2 \cdot (1 - E[r]) \cdot \sigma_x^2$.
- Maximum Entropy technique leads to a uniform distribution for r.
- For the uniform distribution on $[0, 1]$, the probability density is equal to 1, and the mean value is 0.5.
- Substituting the value $E[r] = 0.5$ into the above formula for σ_{xy}^2, we conclude that $\sigma_{xy}^2 = \sigma_x^2$.
- This is exactly the fact that we try to explain.

18. Explaining a Value: Empirical Fact

- When people make crude estimates, their estimates differ by half-order of magnitude.
- For example, when people estimate the size of a crowd, they normally give answers like 100, 300, 1000.
- 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.
- This 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 the above fact; however, these explanations are somewhat complicated.

19. Explaining a Value: Empirical Fact (cont-d)

- For a simple fact about commonsense reasoning, it is desirable to have a simpler, more intuitive explanation.
- 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?
- Let us try to apply the Maximum Entropy techniques to answer this question.
- It may sound reasonable to come up with some probability distributions on the sets of a's and b's.
- Here, we do not have any bound on a and b.

20. Explaining a Value: Empirical Fact (cont-d)

- In this case, the Maximum Entropy technique implies that $\rho(x) = \text{const}$ for all x.
- Thus, $\int_0^\infty \rho(x)\,dx = \infty > 1$.
- 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 $r \stackrel{\text{def}}{=} \dfrac{a}{a+b}$.
- We know that $a < b$, so this ratio takes all possible values from 0 to 0.5.
- Here, 0.5 corresponds to the ideal case when the salaries a and b are equal.

21. Explaining a Value: Empirical Fact (cont-d)

- By using the Maximum Entropy technique, we conclude that r is uniformly distributed on $[0, 0.5)$.
- 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 $a + b$.
- Thus, the average salary b of the second person is equal to the remaining amount $1 - 1/4 = 3/4$.
- So, the ratio of the two salaries is exactly $\dfrac{b}{a} = \dfrac{3/4}{1/4} = 3$.
- 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.

22. Explaining a Functional Dependence

- Often, we know that the value of a quantity x uniquely determines the values of the quantity y.
- So, $y = f(x)$ for some function $f(x)$.
- In some practical situations, this dependence is known.
- In other situations, we need to find this dependence.
- How can we find this dependence?
- 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 $[\underline{y}, \overline{y}]$.

23. Explaining a Functional Dependence (cont-d)

- If we apply the Maximum Entropy technique to x, we conclude that x is uniformly distributed on $[\underline{x}, \overline{x}]$.
- Similarly, we conclude that y is uniformly distributed on $[\underline{y}, \overline{y}]$.
- 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 $[\underline{y}, \overline{y}]$.
- For a uniform distribution, the probability to be in an interval is proportional to its length.
- For a small interval $[x, x + \Delta]$ of width Δx, the probability to be in this interval is equal to $\rho_x \cdot \Delta x$.

24. Explaining a Functional Dependence (cont-d)

- 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
$$\rho_y \cdot |f'(x)| \cdot \Delta x = \rho_x \cdot \Delta x, \text{ so } |f'(x)| = \frac{\rho_x}{\rho_y} = \text{const}.$$

25. Explaining a Functional Dependence (cont-d)

- So, we conclude that *the function* $f(x)$ *should be linear.*
- 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 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,
  - but why would nature care which dependencies are easier for us to analyze?

26. Explaining a Functional Dependence (final)

- 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.

27. Need for Nonlinear Dependencies

- 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.
- It turns out that the Maximum Entropy technique can also help in finding such a nonlinear dependence.
- The first, more direct, idea is to take into account that often,
  - not only the quantity y is observable, but also its derivative $z \stackrel{\text{def}}{=} \dfrac{dy}{dx}$ is an observable quantity.,
  - and sometimes, its second derivative as well.

28. Need for Nonlinear Dependencies (cont-d)

- For example, when y is a distance and x is time, then:
  - the first derivative $v \stackrel{\text{def}}{=} \frac{dy}{dx}$ is velocity and
  - the second derivative $a \stackrel{\text{def}}{=} \frac{dv}{dx} = \frac{d^2y}{dx^2}$ is acceleration,
  - both perfectly observable quantities.
- If we apply the Maximum Entropy techniques to the dependence of v on x, we get $v = a + b \cdot x$.
- In this 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,
  - we conclude that the velocity is a quadratic function of time, and
  - thus, that the distance is a cubic function of time.

29. The Maximum Entropy Technique Can Help Beyond Linear Dependencies: Second Idea

- The second 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?
- We can use the Maximum Entropy technique and conclude that this dependence is linear: $\rho_y(y) = a + b \cdot y$.

30. Beyond Linear Dependencies: Second Idea (cont-d)

- 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 $\rho_y(y) = a + b \cdot y$.
- The probability of being in the x-interval of width Δx is equal to $\rho_x \cdot \Delta x$.
- 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 $\dfrac{df}{dx} \cdot (a + b \cdot f) = \text{const}.$

31. Beyond Linear: 2nd Idea (cont-d)

- By moving all the terms containing f to one side and all the terms containing x to another sides, we get
$$\frac{df}{a+b\cdot f}=\text{const}\cdot x.$$
- So, for $g \stackrel{\text{def}}{=} f+\frac{a}{b}$, we get $\frac{dg}{g}=c\cdot dx$.
- Integration leads to $\ln(g)=c\cdot x+C$ for some 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.
- Assuming that $\rho_y(y)$ is described by one of these non-linear formulas, we can get an even more complex $f(x)$.
- So, Maximum Entropy can describe nonlinear $f(x)$.

32. Acknowledgments

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