On the Principle of Maximum Entropy

Andreas Orthey

February 2024

The principle of maximum entropy [Jaynes, 2003] is a method to choose a probability distribution for a random variable amid partial information. The principle states that we should choose a distribution that maximizes the amount of uncertainty and gives us the least biased guess given all the information we have. In simple terms, we want a distribution that does not favor any event over another—unless we have concrete evidence to the contrary.

Let a random variable X be given with events $A = \{a_1, \dots, a_N\}$ and corresponding, unknown probabilities $P = \{p_1, \dots, p_N\}$. The entropy of this random variable is defined as

$$H(p_1, \dots, p_N) = -\sum p_i \ln(p_i), \tag{1}$$

subject to $p_i \geq 0$ and $\sum p_i = 1$, where summation is implied over $i = 1$ to N.

If you encounter this equation for the first time, you may need some time to process it. It helps to ask and answer some elementary questions:

- How can we optimize Eq. (1), especially when other information, like the mean, is available? (⇒ Sec. 1)
- What is a toy example to help us get a better feel for the problem? (⇒ Sec. 2)
- Why is entropy defined as it is, and how does it reflect our ignorance? Couldn't we just assign $\frac{1}{N}$ or something else to every event? (⇒ Sec. 3)

I will try to answer these questions as best I can in the following sections. Please feel free to skip to whichever question you find most interesting.

1 Solution to the Maximum Entropy Problem

To solve Eq. (1), we need to optimize under the constraint that $\sum p_i = 1$. One compelling approach is the Lagrange multiplier method [Bertsekas, 2014], which takes constraints directly into account using the Lagrangian function. For our case, we define the Lagrangian as

$$L(p_1, \dots, p_N) = -\sum p_i \ln(p_i) - \lambda \left(\sum p_i - 1 \right), \tag{2}$$

where λ is the Lagrange multiplier. Eq. (2) can be solved by taking the derivatives with respect to the probabilities and the Lagrange multiplier and setting them to zero, since the derivatives must vanish at the maximum point. This gives

$$\frac{dL}{dp_i} = 0, \quad \frac{dL}{d\lambda} = 0. \tag{3}$$

Suppose we want to add additional information to this function, for example, a known mean $\mathbb{E}[X] = m$ of the distribution. We can do this by adding another term as

$$L(p_1, \ldots, p_N) = -\sum_{i=1}^{N} p_i \ln(p_i) - \lambda \left(\sum p_i - 1 \right) - \mu \left(\mathbb{E}[X] - m \right). \tag{4}$$

2 Example with a (Loaded) Die

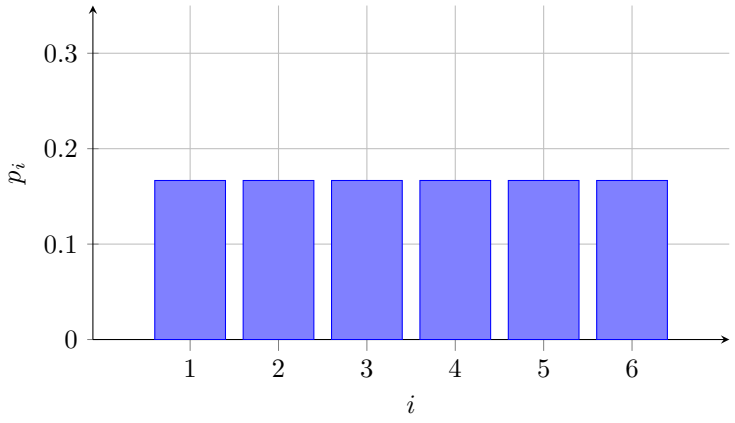
Let us consider an example where our random variable is a six-sided die. First, we consider the case where no other information is available. To find the solution to Eq. (2), we first compute the derivative with respect to p_i as

$$\frac{dL}{dp_i} = -(\ln(p_i) + 1) - \lambda. \qquad \text{(By the product rule)}$$

This derivative must evaluate to zero at the maximum point. This gives us $\ln(p_i) = -(\lambda + 1)$, for which the solution is $p_i = e^{-(\lambda+1)}$. Putting this into the derivative for λ, we get

$$\frac{dL}{d\lambda} = \sum p_i - 1 = N e^{-(\lambda+1)} - 1. \tag{5}$$

Setting this to zero, we obtain a solution for p_i as $e^{-(\lambda+1)} = \frac{1}{N} = p_i$. We stop at this point, since we have found our distribution. Earlier works have denoted this result as the principle of indifference [Keynes, 1921, Jaynes, 2003], since no event is favored. The histogram below shows the values of this distribution.



2.1 An Unexpected Mean

With no constraints, we assign $p_i = \frac{1}{N}$ to the probabilities of a die. However, once constraints are added, things become more interesting.

Suppose we know from experiments that our die has an unexpected mean value of $\mathbb{E}[X] = \sum i \cdot p_i = 4$ (note that a fair six-sided die would have a mean value of 3.5). Let us add this as an additional constraint. Our Lagrangian in this case is

$$L(p_1, \ldots, p_N) = -\sum_{i=1}^{N} p_i \ln(p_i) - \lambda \left(\sum p_i - 1 \right) - \mu \left(\sum i \cdot p_i - 4 \right).$$

To find our solution, we first evaluate the derivatives as

$$\begin{aligned} \frac{dL}{dp_i} &= -(\ln(p_i) + 1) - \lambda - i\mu, \\ \frac{dL}{d\lambda} &= \sum p_i - 1, \\ \frac{dL}{d\mu} &= \sum i \cdot p_i - 4. \end{aligned}$$

We can then evaluate p_i as a function of λ and μ as

$$p_i = e^{-(1+\lambda+i\mu)} = \frac{e^{-i\mu}}{e^{1+\lambda}}.$$

Using this, we substitute p_i into the two constraints to yield

$$\frac{1}{e^{1+\lambda}} \left(e^{-\mu} + e^{-2\mu} + e^{-3\mu} + e^{-4\mu} + e^{-5\mu} + e^{-6\mu} \right) = 1, \tag{6}$$

$$\frac{1}{e^{1+\lambda}} \left(e^{-\mu} + 2e^{-2\mu} + 3e^{-3\mu} + 4e^{-4\mu} + 5e^{-5\mu} + 6e^{-6\mu} \right) = 4. \tag{7}$$

These are two equations with two unknowns. Let us simplify by setting $x = e^{-\mu}$ to get

$$\begin{aligned} \frac{1}{e^{1+\lambda}} \left(x + x^2 + x^3 + x^4 + x^5 + x^6 \right) &= 1, \\ \frac{1}{e^{1+\lambda}} \left(x + 2x^2 + 3x^3 + 4x^4 + 5x^5 + 6x^6 \right) &= 4. \end{aligned}$$

We can equate and rearrange these equations to yield

$$(x + 2x^2 + 3x^3 + 4x^4 + 5x^5 + 6x^6) - 4(x + x^2 + x^3 + x^4 + x^5 + x^6) = 0.$$

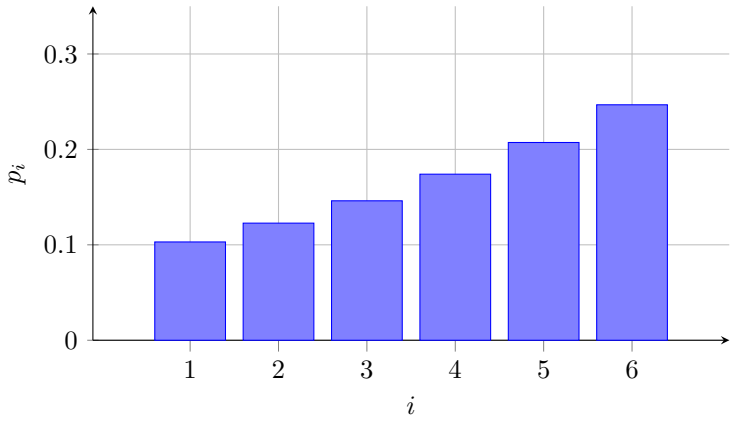
After further simplification, this yields

$$-3x - 2x^2 - x^3 + x^5 + 2x^6 = 0.$$

Since $x > 0$, numerical evaluation yields one real root at $x \simeq 1.19$[1]. This directly yields $\mu = -\ln(1.19) \simeq -0.174625$. Substituting into Eq. (6), we get

$$\lambda = \ln\big(x + x^2 + x^3 + x^4 + x^5 + x^6\big) - 1,$$

which gives $\lambda \simeq 1.44701$. This yields the desired values p_i, which we can plot to obtain the following distribution:



We see that this distribution is skewed toward higher values, which reflects the higher mean value. Note that we obtained this distribution solely from the given mean value together with the principle of maximum entropy.

3 Why is Entropy Defined as It Is?

If you encounter Eq. (1) for the first time, it might look strange, and you might wonder why entropy is not defined differently. To be fair, there are many possibilities. For example, we could minimize the distance between probabilities as $\|p_i - p_j\|$ or maximize the squared values p_i^2, or do something else. Why do we insist on the structure of Eq. (1), and where does it come from?

It turns out that Eq. (1) is forced upon us once we clearly define the conditions we want our function to have. Let us assume that we want to find a function $H(p_1, \ldots, p_N)$ that is maximal for a distribution that is maximally non-committal and favors no event over another [Jaynes, 2003]. What mathematical properties should such a function possess?

- **Additivity Property**. If two events p and q are independent, then H should be an additive function of them, i.e., $H(p, q) = H(p) + H(q)$.
- **Continuity Property**. H should be a continuous function of probabilities p_i, since small changes in p_i should correspond to small changes in H; otherwise, we could get large jumps in H.

[1] Finding roots of a polynomial is a subject in itself, and many algorithms exist to find all roots, such as the Durand-Kerner method Durand [1960], Kerner [1966]. To obtain this result, we queried Wolfram Alpha `https://www.wolframalpha.com/`.

- **Maximum on Uniformity**. Entropy should reach its maximum when all probabilities are equally likely.
- **Symmetry Property**. If the probabilities are permuted in any way, the result should remain the same; i.e., if we relabel but retain the same values, we should expect the same output.

These are relatively mild, sensible properties on which we can all agree. However, Eq. (1) then appears as a logical consequence [Jaynes, 2003].

3.1 Derivation of the Entropy

We are thus searching for a function $H(p_1, \ldots, p_N)$ that fulfills the above-mentioned properties. First, let us clarify what H should represent. Given a set of events, H should measure the uncertainty of the events. Consider the six-sided die. If all outcomes are equally likely, H should attain its maximum. If the die always shows 6 on every throw, the uncertainty is zero. However, every event has its own uncertainty. To bring them together, the only reasonable way to formulate H is as an average uncertainty over all events, where each event contributes uncertainty based on the probability of the event occurring. Since H is supposed to be continuous and additive, we arrive at the form

$$H(p_1, \ldots, p_N) = \sum_{i=1}^{N} p_i h(p_i), \tag{8}$$

where h is a continuous function of p_i, representing the uncertainty of a specific event. Let us investigate the structure of this h function.

For that, assume we have two independent events with probabilities p and q. When we compute the uncertainty $h(pq)$, it should be additive, i.e., $h(pq) = h(p) + h(q)$, since the events are independent. What form could h have to fulfill this?

To see this, let us substitute $p = e^x$ and $q = e^y$ to yield $h(e^{x+y}) = h(e^x) + h(e^y)$. Define a new function $f(x) = h(e^x)$, which gives

$$f(x + y) = f(x) + f(y). \tag{9}$$

This is Cauchy's functional equation [Jurkat, 1965], which has a unique solution $f(x) = kx$ with k being a constant (for f continuous). Thus, we have $h(p) = h(e^x) = f(x) = kx$. Since $h(e^x) = kx$, h must be the inverse of the exponential function, which is the logarithm. Thus, we arrive at $h(p) = k \ln(p)$.

Note that both $h(p) = k \ln(p)$ and $h(p) = -k \ln(p)$ are valid solutions. To resolve this, consider the Maximum on Uniformity property. If we use the positive term, we would reach a minimum of the function when all probabilities are equivalent. The negative term gives the maximum, as desired. Thus, $h(p) = -k \ln(p)$ fulfills this property. Any logarithm will suffice, so we choose the

natural logarithm and set $k = 1$. We then arrive at the full equation for the entropy as

$$H(p_1, \ldots, p_N) = -\sum_{i=1}^{N} p_i \ln(p_i). \tag{10}$$

Consider the six-sided die again. If all events are equally likely, we get $H = -\sum_{i=1}^{N} \frac{1}{6} \ln\left(\frac{1}{6}\right) \simeq 1.79$. However, if one event is certain (e.g., we always throw a six, i.e., $p_6 = 1$), we get $H = -1 \cdot \ln(1) = 0$, as desired.

References

Dimitri P Bertsekas. *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014.

E. Durand. Equations du type F(x) = 0: Racines d'un polynome. In Masson et al., editors, *Solutions Numériques des Equations Algébriques*, volume 1. 1960.

Edwin T Jaynes. *Probability theory: The logic of science*. Cambridge university press, 2003.

Wolfgang B Jurkat. On cauchy's functional equation. *Proceedings of the American Mathematical Society*, 16(4):683–686, 1965.

Immo O. Kerner. Ein gesamtschrittverfahren zur berechnung der nullstellen von polynomen. *Numerische Mathematik*, 8(3):290–294, 1966. doi: 10.1007/BF02162564.

John Maynard Keynes. *A Treatise on Probability*. Macmillan and Co., London, 1921.