Maximum entropy principle with a die

An article posted on bathtubphysics.de

Jan Bergner

2022-10-07

On wikipedia, the principle of maximum entropy is defined as follows:

> The principle of maximum entropy states that the probability distribution which best represents the current state of knowledge about a system is the one with largest entropy, in the context of precisely stated prior data (such as a proposition that expresses testable information).

This means, that for a given probabilistic system, we should be able to find the best probability distribution for all possible states by writing entropy as function of the probability distribution and find the one that maximizes entropy.

When doing this, we will have to consider possible constraints.

A die as probabilistic system

An easy example would be a regular cubic die.

When it is cast, there are six possible results $\{1, 2, 3, 4, 5, 6\}$, whose probabilities can be denoted as $\{p_1, p_2, p_3, p_4, p_5, p_6\}$.

Using the Shannon-entropy-definition,

$$H = -\sum_{j=1}^{6} p_j \cdot \ln p_j.$$

Notice, we have chosen the example of a die, because there are six discrete options, i. e. we do not search a probability density function, but want to find proper probabilities for a set of possible results.

This will allow us to find the maximum entropy configuration with simple derivatives instead of requiring variational calculus.

Mathematically, H is a multi-dimensional function depending on the six distinct probabilities:

$$H = H(p_1, \cdots, p_6)$$

The maximum value can be found by calculating all six partial derivatives $\frac{\partial H}{\partial p_i}$ and setting them to zero. (It should be noted, however, that this is only a required condition, not a sufficient one.)

However, in this case, we also have to satisfy the normalization constraint. The sum of all probabilities has to be 1:

$$\sum_{j=1}^{6} p_j = 1$$

The Lagrange multiplier method

Now, we could rearrange the constraint equation to eliminate one of the variables. For example, we could rewrite to

$$p_6 = 1 - \sum_{k=1}^{5} p_k$$

and insert this into our entropy function. This would lead to a more complicated expression to maximize, but there is an easier way. Introducing a **Lagrange multiplier** λ.

I won't describe the Lagrange multiplier method in too much detail, here. Suffice it to say, we can define a new function H^* and maximize this one, instead:

$$H^*(p_1, \cdots, p_6, \lambda) = H(p_1, \cdots, p_6) - \lambda \left[\sum_{j=1}^{6} p_j - 1 \right]$$

Notice, H^* will additionally depend on λ, but this dependence will only ensure our normalization constraint, as the partial derivative $\frac{\partial H^*}{\partial \lambda}$ will simply yield the constraint equation.

If you do not believe, that the maximization of this new function will also maximize the old one, you can just read the literature on this. But as a short motivation, consider that the second summand will always be zero for a real-world-solution satisfying the normalization constraint. With this perspective, H and H^* are the same function for every possible real-world configuration of probabilities.

Now, let us calculate the derivative with respect to the i-th probability p_i step by step:

Calculating the probabilities from the maximum-entropy-principle

$$\begin{aligned}
\frac{\partial}{\partial p_i} H^*(p_1, \cdots, p_6, \lambda) =& \frac{\partial}{\partial p_i} \left[H(p_1, \cdots, p_6) - \lambda \left(\sum_{j=1}^{6} p_j \ - 1 \right) \right] \\
=& \frac{\partial}{\partial p_i} \left[- \sum_{j=1}^{6} p_j \cdot \ln p_j - \lambda \left(\sum_{j=1}^{6} p_j \ - 1 \right) \right] \\
=& - \frac{\partial}{\partial p_i} \left[\sum_{j=1}^{6} p_j \cdot \ln p_j + \lambda \sum_{j=1}^{6} p_j \ - \lambda \right] \\
=& - \frac{\partial}{\partial p_i} \left[\sum_{j=1}^{6} (p_j \cdot \ln p_j + \lambda p_j) \ - \lambda \right] \\
=& - \frac{\partial}{\partial p_i} \left[\sum_{j=1}^{6} (p_j \cdot \ln p_j + \lambda p_j) \right] + \frac{\partial}{\partial p_i} \lambda
\end{aligned}$$

Now, let us start with the Lagrange multiplier term. The way it is constructed, it is independent of the probabilities and thus, this derivative will just vanish.

As for the derivative of the sum, we can do every summand individually. Since the j-th summand will only depend on p_j and we do the derivative with respect to p_i, only the i-th summand will survive. With that, the derivative becomes straightforward:

$$\begin{aligned}
\frac{\partial}{\partial p_i} H^*(p_1, \cdots, p_6, \lambda) =& - \frac{\partial}{\partial p_i} \left[\sum_{j=1}^{6} (p_j \cdot \ln p_j + \lambda p_j) \right] + \frac{\partial}{\partial p_i} \lambda \\
=& - \frac{\partial}{\partial p_i} (p_i \cdot \ln p_i + \lambda p_i) \\
=& - (\ln p_i + 1 + \lambda) \\
=& - \ln p_i - 1 - \lambda
\end{aligned}$$

Notice, that λ is a constant that is not yet fixed to a concrete value.

This allows us to shorten our equation a bit by substituting

$$\lambda^* = 1 + \lambda.$$

With this,

$$\boxed{\frac{\partial}{\partial p_i} H^* = -\ln p_i - \lambda^*.}$$

We are looking for the entropy's maximum, i. e. all the derivatives with respect to any p_i must be equal to 0. We obtain:

$$\begin{aligned} \frac{\partial}{\partial p_i} H^* = 0 = &-\ln p_i - \lambda^* \\ \Leftrightarrow \quad \ln p_i = &-\lambda^* \\ p_i = &e^{-\lambda^*} \end{aligned}$$

This is already telling us a lot. The first and foremost result is:

All probabilities are equal.

(λ^* is a not-yet-defined constant and every one of the probabilities equals $p_i = e^{-\lambda^*}$.)

Now, we can use the normalization constraint. Here,

$$p_i = \frac{1}{6}.$$

If we really want to, we can also calculate λ^* or λ, respectively:

$$\begin{aligned} -\lambda^* &= \ln p_i \\ &= \ln\left(\frac{1}{6}\right) \\ &= -\ln 6 \\ \Leftrightarrow \quad \lambda^* &= \ln 6 \\ \Rightarrow \quad \lambda &= \ln 6 - 1 \end{aligned}$$

Equal probabilities

From the maximum-entropy-principle, we found all probabilities to be equal, i. e. every result is equally likely, when throwing the die.

However, the maths we used could not only represent the die but was quite general.

In fact, for any random system that can end up in one of six discrete states, the calculations would be the same.

Generalization to N probabilities

If we do not deal with six but N possible states, we would simply replace the six in the upper sum limit by N.

From here on, let us combine the distinct probabilities into a probability “vector”:

$$\underline{\mathbf{p}_N} := \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_N \end{pmatrix}$$

This will help to shorten the equations. We can furthermore introduce the “one-vector”

$$\underline{\mathbf{1}_N} := \underbrace{\begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix}^T}_{N \text{ times } 1},$$

that can be used to shorten the normalization constraint equation:

$$\underline{\mathbf{1}_N}^T \cdot \underline{\mathbf{p}_N} = 1$$

Now, let us also define an “entropy-contribution-vector”:

$$\underline{\mathbf{h}_N} := \begin{pmatrix} -p_1 \cdot \ln p_1 \\ -p_2 \cdot \ln p_2 \\ \vdots \\ -p_N \cdot \ln p_N \end{pmatrix}$$

With these shortcuts, we can write

$$H(\underline{\mathbf{p}_N}) = \underline{\mathbf{1}_N}^T \cdot \underline{\mathbf{h}_N}$$

and

$$\begin{aligned} H^*(\underline{\mathbf{p}_N}, \lambda) =& \underline{\mathbf{1}_N}^T \cdot \underline{\mathbf{h}_N} - \lambda \left(\underline{\mathbf{1}_N}^T \cdot \underline{\mathbf{p}_N} - 1 \right) \\ =& \underline{\mathbf{1}_N}^T \cdot \left[\underline{\mathbf{h}_N} - \lambda \left(\underline{\mathbf{p}_N} - \frac{1}{N} \underline{\mathbf{1}_N} \right) \right]. \end{aligned}$$

This is one way to write the entropy function and the helper function H^* with the Lagrange multiplier in a generalized form for an N-state-system.

All the derivatives being 0 means the gradient vector

$$\nabla_{N;\lambda} = \begin{pmatrix} \frac{\partial}{\partial p_1} \\ \frac{\partial}{\partial p_2} \\ \vdots \\ \frac{\partial}{\partial p_N} \\ \frac{\partial}{\partial \lambda} \end{pmatrix},$$

applied to H^* must be the zero-vector $\underline{\mathbf{0}_{N+1}}$:

$$\begin{aligned} \nabla_{N;\lambda} H^*(\underline{\mathbf{p}_N}, \lambda) =& \underline{\mathbf{0}_{N+1}} \\ =& \nabla_{N;\lambda} \left[\underline{\mathbf{1}_N}^T \cdot \left[\underline{\mathbf{h}_N} - \lambda \left(\underline{\mathbf{p}_N} - \frac{1}{N} \underline{\mathbf{1}_N} \right) \right] \right] \quad \mid \lambda^* = \lambda + 1 \\ =& \begin{pmatrix} -\ln p_1 - \lambda^* \\ -\ln p_2 - \lambda^* \\ \vdots \\ -\ln p_N - \lambda^* \\ 1 - \sum_{j=1}^{N} p_j \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} \\ \Rightarrow \quad p_1 = p_2 = \cdots = p_N =& e^{-\lambda^*} = \frac{1}{N} \end{aligned}$$

This implies, that all probabilities being equal is always the configuration that will maximize entropy. (In absence of further constraints, that is.)

Actually, this is a fundamental assumption in physics, when deriving macroscopic equations of ther-

modynamics from a microscopic description based on quantum mechanics. Using this assumption will eventually lead back to entropy being maximized.

Additional constraints

So far, we have dealt with probabilistic systems with no additional constraints, but the maximum-entropy-principle is not limited to that.

Let us just toy around with a simple example that goes back to our die. But in this case, it shall be an unfair die.

Consider the die not having equal density everywhere, but instead, it shall be heavier close to the *six*-face and lighter close to the *one*-face. For the sake of simplicity, let us assume that this has the effect of making a result of *one* twice as likely as a result of *six* when throwing the die:

$$p_1 = 2p_6$$

The other faces' probabilities shall not be skewed.

For this system, we can consider the new constraint by introducing a second Lagrange multiplier μ:

$$H^*(\underline{\mathbf{p}_6}, \lambda, \mu) = -\sum_{j=1}^{6} p_j \cdot \ln p_j - \lambda \left(\sum_{j=1}^{6} p_j \ - 1 \right) - \mu \left(p_1 - 2p_6 \right)$$

The derivative with respect to μ must be added to the gradient $\nabla_{6;\lambda,\mu}$, which implies:

$$\nabla_{6;\lambda,\mu} H^* = \underline{\mathbf{0}_8} = \begin{pmatrix} \frac{\partial}{\partial p_1} H^* \\ \frac{\partial}{\partial p_2} H^* \\ \frac{\partial}{\partial p_3} H^* \\ \frac{\partial}{\partial p_4} H^* \\ \frac{\partial}{\partial p_5} H^* \\ \frac{\partial}{\partial p_6} H^* \\ \frac{\partial}{\partial \lambda} H^* \\ \frac{\partial}{\partial \mu} H^* \end{pmatrix} = \begin{pmatrix} -\ln p_1 - \lambda^* - \mu \\ -\ln p_2 - \lambda^* \\ -\ln p_3 - \lambda^* \\ -\ln p_4 - \lambda^* \\ -\ln p_5 - \lambda^* \\ -\ln p_6 - \lambda^* + 2\mu \\ 1 - \sum\limits_{j=1}^{6} p_j \\ 2p_6 - p_1 \end{pmatrix}$$

Now, first of all, the non-tampered faces still are all equally likely:

$$p_2 = p_3 = p_4 = p_5 = e^{-\lambda^*}$$

For p_1 and p_6, it is a little bit different, as they also depend on the new Lagrange multiplier μ. To find the answer here, we need to consider the following four equations. (Notice, that p_3 to p_5 are rewritten as p_2 in the third equation.)

$$\begin{aligned} 0 =& -\ln p_1 - \lambda^* - \mu \\ 0 =& -\ln p_6 - \lambda^* + 2\mu \\ 1 =& p_1 + p_6 + 4p_2 \\ p_1 =& 2p_6 \end{aligned}$$

Let us replace $p_1 = 2p_6$:

$$\begin{aligned} 0 =& -\ln(2p_6) - \lambda^* - \mu \\ 0 =& -\ln p_6 - \lambda^* + 2\mu \\ 1 =& 3 \cdot p_6 + 4 \cdot p_2 \end{aligned}$$

We now multiply the first equation by two and add the second:

$$\begin{aligned} (\text{I}) \quad 0 =& -2\ln(2p_6) - 2\lambda^* - 2\mu \\ (\text{II}) \quad 0 =& -\ln p_6 - \lambda^* + 2\mu \\ (\Rightarrow) \quad 0 =& -2\ln(2p_6) - \ln p_6 - 3\lambda^* \end{aligned}$$

This one can be simplified:

$$\begin{aligned} 0 &= -2\ln(2p_6) - \ln p_6 - 3\lambda^* \\ &= -2(\ln 2 + \ln p_6) - \ln p_6 - 3\lambda^* \\ 2\ln 2 &= -3\ln p_6 - 3\lambda^* \\ \frac{2}{3}\ln 2 &= -\ln p_6 - \lambda^* \end{aligned}$$

This can be inserted into equation (II), to find μ. Notice, that at this point, this is just a low-hanging fruit. We do not really need it, as we are only interested into the actual probabilities:

$$\begin{aligned}
\frac{2}{3}\ln 2 &= -\ln p_6 - \lambda^* \\
0 &= -\ln p_6 - \lambda^* + 2\mu \\
(\Rightarrow)\ 0 &= \frac{2}{3}\ln 2 + 2\mu \\
\Leftrightarrow\ \mu &= -\frac{1}{3}\ln 2
\end{aligned}$$

On the other hand, we can now express p_6 in terms of λ^*:

$$\begin{aligned}
\frac{2}{3}\ln 2 &= -\ln p_6 - \lambda^* \\
\ln p_6 &= -\frac{2}{3}\ln 2 - \lambda^* \\
p_6 &= e^{-\frac{2}{3}\ln 2 - \lambda^*} \\
&= e^{-\frac{2}{3}\ln 2} \cdot \underbrace{e^{-\lambda^*}}_{=p_2} \\
&= e^{-\frac{2}{3}\ln 2} \cdot p_2 \\
&= 2^{-\frac{2}{3}} \cdot p_2 \\
\Leftrightarrow\ p_6 &= \frac{1}{\sqrt[3]{4}} \cdot p_2
\end{aligned}$$

Thus:

$$\begin{aligned}
p_1 &= 2p_6 \\
&= \frac{2}{\sqrt[3]{4}} \cdot p_2 \\
&= \sqrt[3]{2} \cdot p_2
\end{aligned}$$

Let us write this out:

$$\begin{aligned}
p_1 &= \sqrt[3]{2} \cdot p_2 \\
p_3 = p_4 = p_5 &= p_2 \\
p_6 &= \frac{1}{\sqrt[3]{4}} \cdot p_2
\end{aligned}$$

By normalization, we can finally put numbers on our probabilities:

$$\begin{aligned} 1 =& 3p_6 + 4p_2 \\ 1 =& \left(\frac{3}{\sqrt[3]{4}} + 4\right) \cdot p_2 \\ p_2 =& \frac{1}{\frac{3}{\sqrt[3]{4}} + 4} \\ \approx& 0.170 \end{aligned}$$

With a simple evaluation, we also find:

$$\begin{aligned} p_1 =& \frac{\sqrt[3]{2}}{\frac{3}{\sqrt[3]{4}} + 4} = \frac{2}{3 + 4\sqrt[3]{4}} \approx 0.214 \\ p_6 =& \frac{1}{\sqrt[3]{4}} \cdot \frac{1}{\frac{3}{\sqrt[3]{4}} + 4} = \frac{1}{3 + 4\sqrt[3]{4}} \approx 0.107 \end{aligned}$$

Indeed, throwing a six becomes by far the least likely result and throwing a one is more likely than two, three, four or five.

Unfortunately, we cannot really do a lot with this result.

After all, we usually don't have a die like this at our disposal to test, whether the calculated probabilites really work in real life.

What's more, it is not even clear, how to construct a die that will exactly fit the constraint we introduced. (I. e. $p_1 = 2p_6$)

So, let us check by a tedious independent calculation.

Re-calculating without Lagrange multipliers

Here we go…

Let us re-iterate our constraints, first:

$$\begin{aligned} \sum_{j=1}^{6} p_j =& 1 \\ p_1 =& 2p_6 \end{aligned}$$

Next, recall the original entropy

$$H = -\sum_{j=1}^{6} p_j \cdot \ln p_j.$$

Using constraint two, we can get rid of p_1 in the first constraint and the entropy equation. Notice, that I replaced the summation index, since the following sums will only go from 2 to 5. I think, this can help following along.

$$\begin{aligned} 1 &= \sum_{k=2}^{5} p_k + 3p_6 \\ \Leftrightarrow \; p_6 &= \frac{1}{3}\left(1 - \sum_{k=2}^{5} p_k\right) \end{aligned}$$

$$\begin{aligned} H &= -\sum_{k=2}^{5} p_k \cdot \ln p_k - 2p_6 \cdot \ln\left(2p_6\right) - p_6 \cdot \ln p_6 \\ &= -\sum_{k=2}^{5} p_k \cdot \ln p_k - 2p_6 \cdot \left(\ln 2 + \ln p_6\right) - p_6 \cdot \ln p_6 \\ &= -\sum_{k=2}^{5} p_k \cdot \ln p_k - 2\ln 2p_6 - 3p_6 \cdot \ln p_6 \end{aligned}$$

Now, we get rid of p_6:

$$\begin{aligned} H &= -\sum_{k=2}^{5} p_k \cdot \ln p_k - 2\ln 2p_6 - 3p_6 \cdot \ln p_6 \\ &= -\sum_{k=2}^{5} p_k \cdot \ln p_k - 2\ln 2\left[\frac{1}{3}\left(1 - \sum_{k=2}^{5} p_k\right)\right] - 3\left[\frac{1}{3}\left(1 - \sum_{k=2}^{5} p_k\right)\right] \cdot \ln\left[\frac{1}{3}\left(1 - \sum_{k=2}^{5} p_k\right)\right] \\ &= -\sum_{k=2}^{5} p_k \cdot \ln p_k - \frac{2}{3}\ln 2\left(1 - \sum_{k=2}^{5} p_k\right) - \left(1 - \sum_{k=2}^{5} p_k\right) \cdot \left[-\ln 3 + \ln\left(1 - \sum_{k=2}^{5} p_k\right)\right] \end{aligned}$$

Let the fun begin.

First of all, we can see, that we can exchange any pair of probabilities $\{p_2, p_3, p_4, p_5\}$ and the entropy function will not change.

The entropy is symmetrical in the respective probabilities and thus, all the derivatives should be identical and in turn, these probabilities will assume the same value in order to maximize the entropy.

In any case, we can already confirm one of the results from before, which is a good thing.

However, we will not just follow the symmetry argument but instead do our calculations to re-discover this result. We will compute the derivative with respect to probability p_i, where

$$i \in \{2, 3, 4, 5\}.$$

Let's calculate derivatives!

$$\begin{aligned}
0 = \frac{\partial H}{\partial p_i} =& \frac{\partial}{\partial p_i}\left[-\sum_{k=2}^{5} p_k \cdot \ln p_k - \frac{2}{3}\ln 2\left(1-\sum_{k=2}^{5} p_k\right) - \left(1-\sum_{k=2}^{5} p_k\right)\cdot\left[-\ln 3 + \ln\left(1-\sum_{k=2}^{5} p_k\right)\right]\right] \\
=& -\ln p_i - 1 + \frac{2}{3}\ln 2 + \left[-\ln 3 + \ln\left(1-\sum_{k=2}^{5} p_k\right)\right] \underbrace{- \left(1-\sum_{k=2}^{5} p_k\right)\cdot\frac{-1}{1-\sum_{k=2}^{5} p_k}}_{+1} \\
=& -\ln p_i + \frac{2}{3}\ln 2 + \left[-\ln 3 + \ln\left(1-\sum_{k=2}^{5} p_k\right)\right]
\end{aligned}$$

Recall and replace with

$$1 - \sum_{k=2}^{5} p_k = 3p_6 :$$

$$\begin{aligned}
0 =& -\ln p_i + \frac{2}{3}\ln 2 + \left[-\ln 3 + \ln\left(1-\sum_{k=2}^{5} p_k\right)\right] \\
=& -\ln p_i + \frac{2}{3}\ln 2 + [-\ln 3 + \ln(3p_6)] \\
=& -\ln p_i + \frac{2}{3}\ln 2 + [-\ln 3 + \ln 3 + \ln p_6] \\
\Leftrightarrow\ \ln p_i =& \frac{2}{3}\ln 2 + \ln p_6 \quad \mid e^{\cdots} \\
\Rightarrow\ p_i =& e^{\frac{2}{3}\ln 2}\cdot e^{\ln p_6} \\
=& \sqrt[3]{4}p_6 \\
\Leftrightarrow\ p_6 =& \frac{1}{\sqrt[3]{4}}p_i
\end{aligned}$$

for $p_i = p_2 = p_3 = p_4 = p_5$.

There we are, again. We could have chosen i from 2 to 5 and we would have always gotten the same result. (In terms of p_6.)

Furthermore, the relation between these probabilities is the same, that we originally derived with the Lagrange multiplier approach and thus, we get identical results.

It seems satisfactory to me to see at least one example, where the Lagrange multiplier formalism worked, properly.

However, I may at some point do a general derivation of why this is the case.

For now, we take what we already got.

Conclusions

- We have introduced the Shannon-entropy-definition.
- From that definition and the normalization constraint, we have demonstrated, that without further constraints, all possible results will have the same probability, if we require the entropy to be maximized. (Actually, we have not proven, that equal probabilities lead to a maximum instead of, say, a minimum.)
- We introduced the approach of Lagrange multipliers to deal with constraints.
- For one example, we demonstrated, that the Lagrange multiplier approach lead to the same result as using the constraints to reduce the entropy's parameters.

I know, this article feels a bit "stand-alone" and we do not really gain insight into an intriguing physical problem.

But it sets the stage for the incredibly useful Lagrange multiplier formalism and shows the deep link between entropy maximization and the equal distribution of probabilities.

And for what it's worth, I consider all of this quite beautiful math.