# 6. Principle of maximum entropy

## Contents

*(All variables are real and one-dimensional unless otherwise specified.)*

## 1. Meanings of entropy

Entropy is deduced from probability, so the former is as philosophically profound as the latter. Entropy may be interpreted differently depending on the context. It may be understood as, somewhat paradoxically, both **information** and **uncertainty**, i.e. a lack of information.

### 1.1 Whose entropy is it?

Consider a fair coin and a fair dice. Their values of entropy are respectively $\langle \log_2 2 \rangle = 1$ bit and $\langle \log_2 6 \rangle \approx 2.58$ bits. The dice has a higher entropy than the coin, but what do these numbers actually convey?

**Information.**

- The dice bears a higher entropy because it **intrinsically** has more possible outcomes than the coin and is thus more random.
- More entropy gives you more **information**, i.e. some sort of **power** for you to "alter" the universe. On one hand, if you toss the coin and get a head, you eliminate one possible event "you get a tail" from the universe. On the other hand, if you throw the dice and get a one, you eliminate the other five possible events. You **eliminate more possibilities** from the universe with the dice, so it has a higher entropy.

**Uncertainty.**

- The dice bears a higher entropy because **you the observer** knows fewer about it than about the coin.
- More entropy gives you more **uncertainty**, i.e. some sort of **resistance** that you have to overcome. On one hand, if I have tossed the coin and ask you to guess my outcome, you choose from two possible values. On the other hand, if I have thrown the dice and ask you to

guess my outcome, you choose from six possible values. The dice gives you **more choices**, so it has a more entropy.

The former perspective views entropy as the coin's and the dice's properties, which objectively exists without you, whereas the latter perspective views entropy as a matter of the observer—once I tell you the dice is in fact fully biased and you trust me, its entropy in your mind drops to zero although I may be lying. Here, the dice does not change; you change. After all, such an oxymoron arises due to the sharply distinct **frequentist** and **Bayesian** interpretations towards probability.

Shannon developed his theory of communication from the frequentist perspective because his entropy measures **a message's information**, i.e. the **frequency** of every possible symbol, then he tried to maximize the mutual information between a message's transmitter and receiver. Jaynes, conversely, developed his principle of maximum entropy from the Bayesian perspective, so his entropy measures **a scientist's uncertainty** on his statements; he argued that the an **objective** statement should be as uncertain as possible.

# 2. Entropy as uncertainty

An objective scientist should be honest to his observations, so he should not assume anything that he does not know. Consequently, he should make himself **as uncertain as possible** by maximizing **"his"** (his brain's, his mind's, his consciousness's, whatever) entropy subject to the given information.

## 2.1 Lagrangian multiplier

This is a typical problem of **constrained optimization**, which **Lagrangian multiplier** is devoted to deal with. In the simplest case, we optimize a two-variable function $f(x, y)$ subject to an equation $g(x, y) = c$. We first let $h(x, y) \equiv g(x, y) - c$ and define a **Lagrangian function**

$$\mathcal{L}(x, y, r) = f(x, y) - rh(x, y)\,,$$

then we solve

$$\nabla\mathcal{L} = \left(\frac{\partial\mathcal{L}}{\partial x}, \frac{\partial\mathcal{L}}{\partial y}, \frac{\partial\mathcal{L}}{\partial r}\right) = 0$$

to obtain $\left(x^*, y^*, r^*\right)$. The constrained optima of $f(x, y)$ are then given by $f\left(x^*, y^*\right)$. More generally, we optimize an $n$-variable function $f(\mathbf{x})$ for $\mathbf{x} \equiv \left(x_1, x_2, \ldots, x_n\right)$ subject to $m$ equations $g_j(\mathbf{x}) = c_j$. We similarly define a Lagrangian function

$$\mathcal{L}(\mathbf{x}, \mathbf{r}) = f(\mathbf{x}) - \mathbf{r} \cdot \mathbf{h}(\mathbf{x})$$

with $\mathbf{r} \equiv (r_1, r_2, \ldots, r_m)$ and $\mathbf{h}(\mathbf{x}) \equiv [g_1(\mathbf{x}) - c_1, g_2(\mathbf{x}) - c_2, \ldots, g_n(\mathbf{x}) - c_n]$, then we solve

$$\nabla \mathcal{L} = \left( \frac{\partial \mathcal{L}}{\partial x_1}, \frac{\partial \mathcal{L}}{\partial x_2}, \ldots, \frac{\partial \mathcal{L}}{\partial x_n}, \frac{\partial \mathcal{L}}{\partial r_1}, \frac{\partial \mathcal{L}}{\partial r_2}, \ldots, \frac{\partial \mathcal{L}}{\partial r_m} \right) = 0$$

to obtain $(\mathbf{x}^*, \mathbf{r}^*)$, which yield the constrained optima $f(\mathbf{x}^*)$.

In principle, we should also check whether the constrained optima obtained are maxima, minima, or saddle points; however, we definitely obtain a maximum finally as entropy is a concave function with one and only one local maximum.

The coming examples define entropy with natural logarithm because it simplifies calculus.

## 2.2 A silly example: a mysterious dice

Suppose that we, a group of objective scientists, get a mysterious discrete random number generator, i.e. a dice. We know nothing about the dice except that its outcome is discrete. What can we say about the dice's probability distribution?

**Solution.** To be the most objective, we should not assume the dice has six faces. A general dice has $n$ faces and lands on the $i$th face with probability $p_i \geq 0$, which satisfies the normalization condition $\sum_{i=1}^{n} p_i = 1$. Then we need to be **maximally uncertain** about the distribution, so we conditionally maximize the dice's entropy $H = -\sum_{i=1}^{n} p_i \ln p_i$. The corresponding Lagrangian function is

$$\mathcal{L}(\mathbf{p}, r) = -\sum_{i=1}^{n} p_i \ln p_i - r \left( \sum_{i=1}^{n} p_i - 1 \right) .$$

To solve $\nabla \mathcal{L} = 0$, we specifically consider the $j$th face:

$$\frac{\partial \mathcal{L}}{\partial p_j} = -\ln p_j - 1 - r = 0 .$$

This yields $p_j = e^{-1-r}$, which turns out to be constant for all faces. Combining this result with the normalization condition, we conclude that $p_i = 1/n$ for all $i$.

## 2.3 A sensible example: a less mysterious dice

We, the objective scientists, get another dice. This time, we know more about the dice.

- It has six faces marked from one to six.
- The mean of its outcome is $\mu = 3.75$.

What can we say about the dice's probability distribution now?

**Solution.** Denote the dice's outcome by $X \in \{1,2,3,4,5,6\}$, then let $p_i \equiv P(X=i)$ be subject to $p_i \geq 0$ and $\sum_{i=1}^{6} p_i = 1$. The second given constraint corresponds to $\sum_{i=1}^{6} ip_i = \mu$ , so we have the following Lagrangian function:

$$\mathcal{L}(\mathbf{p}, r_0, r_1) = -\sum_{i=1}^{6} p_i \ln p_i - r_0 \left(\sum_{i=1}^{6} p_i - 1\right) - r_1 \left(\sum_{i=1}^{6} ip_i - \mu\right) .$$

Upon differentiation, it becomes

$$\frac{\partial \mathcal{L}}{\partial p_j} = -\ln p_j - 1 - r_0 - r_1 j = 0 \Rightarrow p_j = e^{-1-r_0-r_1 j} .$$

Substituting the result back into the normalization constraint, we get

$$e^{-1-r_0} = \frac{1}{\sum_{i=1}^{6} e^{-r_1 i}} \quad \Rightarrow \quad p_j = \frac{e^{-r_1 j}}{\sum_{i=1}^{6} e^{-r_1 i}} .$$

Finally, we substitute the refined probability into the unexploited constraint of mean to get

$$\frac{\sum_{j=1}^{6} je^{-r_1 j}}{\sum_{i=1}^{6} e^{-r_1 i}} = \mu .$$

We can only solve it numerically: $\mu = 3.75$ gives $r_1 \approx -0.0861$ and $\mathbf{p} \approx (0.133, 0.145, 0.158, 0.172, 0.188, 0.204)$.

## 2.4 A tricky example: an even less mysterious dice

The procedure of entropy maximization stays the same when we are given more information, but we must formulate our Langrangian function skillfully: some makes our lives easier.

Consider last subsection's dice again. What is this dice's new probability distribution if we further know that the variance of its outcome is $\sigma^2 = 0.6$?

**Solution.** The constraint of variance means $\sum_{i=1}^{6} (i-\mu)^2 p_i = \sigma^2$. You may be tempted to define the Lagrangian function as

$$\begin{aligned}\hat{\mathcal{L}}(\mathbf{p}, r_0, r_1, r_2) = &-\sum_{i=1}^{6} p_i \ln p_i - r_0 \left(\sum_{i=1}^{6} p_i - 1\right) \\ &- r_1 \left(\sum_{i=1}^{6} ip_i - \mu\right) - r_2 \left[\sum_{i=1}^{6} (i-\mu)^2 p_i - \sigma^2\right] ,\end{aligned}$$

with which you will obtain

$$\hat{p}_j = \frac{e^{-r_1 j - r_2 (j-\mu)^2}}{\sum_{i=1}^{6} e^{-r_1 i - r_2 (i-\mu)^2}} .$$

Although this is not wrong—you can solve two unknowns $r_2$ and $r_3$ with the two constraints of mean and variance—this is **unnecessarily complicated**. In fact, the constraint of variance entails the constraint of mean: once you employ $\sum_{i=1}^{6} (i-\mu)^2 p_i = \sigma^2$, you have implicitly required $\mu$ to be the mean. Therefore, we can drop the redundant constraint of mean and simplify the Lagrangian function as

$$\mathcal{L}(\mathbf{p}, r_0, r_2) = -\sum_{i=1}^{6} p_i \ln p_i - r_0 \left(\sum_{i=1}^{6} p_i - 1\right) - r_2 \left[\sum_{i=1}^{6} (i-\mu)^2 p_i - \sigma^2\right] ,$$

which yields

$$p_j = \frac{e^{-r_2 (j-\mu)^2}}{\sum_{i=1}^{6} e^{-r_2 (i-\mu)^2}} ,$$

having one unknown fewer. For $\begin{cases} \mu = 3.75 \\ \sigma^2 = 0.6 \end{cases}$, we get $r_2 \approx 0.832$ and $\mathbf{p} \approx (0.001, 0.040, 0.322, 0.489, 0.140, 0.008)$.

# 3. Continuous distributions

Let us now discuss the principle of maximum entropy in terms of **continuous** random variables. Although its idea keeps intact, the definition of entropy for continuous variables becomes a bit nontrivial, and the consequent operation of Lagrangian multiplier requires extra attention.

## 3.1 Differential entropy

For a continuous random variable $X$ that follow a distribution $f(x)$, Shannon defined its entropy, aka **differential entropy**, as

$$H(X) \overset{\text{Shannon}}{=} -\int f(x) \ln f(x) \mathrm{d}x$$

by somewhat naively replacing summation with integration. This definition **usually works**, but it creates absurdity when you transform $X$ to $Y = g(X)$ with a function $g$. It fails because of **dimensional inconsistency**: on one hand, the integral requires $[f(x)] = [1/\mathrm{d}x]$; on the other hand, logarithm requires a dimensionless argument $[f(x)] = [1] \neq [1/\mathrm{d}x]$.

To fix this, Jaynes introduced a function $m(x)$ and redefined differential entropy as

$$H(X) \overset{\text{Jaynes}}{=} -\int f(x)\ln\left[\frac{f(x)}{m(x)}\right]\mathrm{d}x$$

so that $f(x)/m(x)$ is dimensionless and safe for logarithm—but what is $m(x)$?

### 3.2 Example: nothing but a range

Suppose we have a continuous random variable $X$. If we know nothing other than $X \in [a, b]$, what can we say about its distribution $f(x)$?

Consider Jaynes's definition of differential entropy for the time being although we do not know the nature of $m(x)$ yet.

**Solution.** We have to maximize the entropy $H(X) = -\int_a^b f(x)\ln\frac{f(x)}{m(x)}\mathrm{d}x$ subject to the normalization condition $\int_a^b f(x)\mathrm{d}x = 1$. Following the intuition learned from previous discrete cases, we can write down

$$J(f, r) = -\int_a^b f(x)\ln\frac{f(x)}{m(x)}\mathrm{d}x - r\left[\int_a^b f(x)\mathrm{d}x - 1\right],$$

but the result looks undesirably bizarre. While $J$ seems to be some kind of Lagrangian function, how can we differentiate it with respect to $f$, which is also a function?

**Mathematics.** Mathematicians call the strange object $J$ a **functional**, which maps a function to a number. To differentiate a functional with respect to a function, we need the **calculus of variations**. If $J(f) = \int L[f(x)]\mathrm{d}x$ up to some additive constants, the derivative of $J$ with respect to $f$ is

$$\begin{aligned}\frac{\delta J}{\delta f} &\equiv \frac{\partial L}{\partial f} - \frac{\mathrm{d}}{\mathrm{d}x}\frac{\partial L}{\partial f'} + \frac{\mathrm{d}^2}{\mathrm{d}x^2}\frac{\partial L}{\partial f''} - \frac{\mathrm{d}^3}{\mathrm{d}x^3}\frac{\partial L}{\partial f'''} + \cdots \\ &= \sum_{n=0}^{\infty}(-1)^n\frac{\mathrm{d}^n}{\mathrm{d}x^n}\frac{\partial L}{\partial f^{(n)}},\end{aligned}$$

in which $f$ and its $n$th derivative $f^{(n)}$ are treated as usual numerical variables on the right-hand side. The equation is called the **Euler-Lagrange equation**, and the function $L$ is, somewhat thanks to Lagrange's wisdom, also called a Lagrangian function.

With the slightly confusing terminology, the Lagrangian function of the functional in our example is

$$L(f) = -f \ln \frac{f}{m} - rf = -f \ln f - rf + f \ln m$$

and yields a functional derivative

$$\frac{\delta J}{\delta f} = \frac{\mathrm{d}L}{\mathrm{d}f} = -\ln f - 1 - r + \ln m\,.$$

In order to maximize $J$, we solve $\frac{\delta J}{\delta f} = 0$ and get $f(x) = m(x)e^{-1-r}$; finally, by substituting it into the normalization condition $\int_a^b f(x)\mathrm{d}x = 1$, we can eliminate the multiplier $r$ and get

$$e^{-1-r} = \frac{1}{\int_a^b m(x)\mathrm{d}x} \quad \Rightarrow \quad f(x) = \frac{m(x)}{\int_a^b m(x)\mathrm{d}x}\,.$$

This answer has revealed the nature of $m(x)$.

**Moral.** In fact, $\frac{m(x)}{\int_a^b m(x)\mathrm{d}x}$ can be regarded as the **prior distribution** of $X$, whereas $f(x)$ is the **posterior distribution** that one believes in after considering some constraints; consequently, $H(X)$ means the gain in uncertainty after one updates his belief. When nothing is known about a continuous variable but its range a priori, the posterior should be just identical to the prior. Since we usually choose to believe that $X$ is distributed in the range **uniformly**, we may simply set $m(x) = 1$ so that

$$\frac{m(x)}{\int_a^b m(x)\mathrm{d}x} = \frac{1}{\int_a^b 1\mathrm{d}x} = \frac{1}{b-a}$$

represents a uniform prior. This choice also results in numerical equivalence between Jaynes's and Shannon's definitions.

**Trivia.** Jaynes's differential entropy turns out to be formally equivalent to the **Kullback-Leibler divergence** (aka **relative entropy**)

$$D(f\|m) = \int f(x) \ln\left[\frac{f(x)}{m(x)}\right]\mathrm{d}x$$

from $m(x)$ to $f(x)$ despite their independent attempts to refine Shannon's theory. (Jaynes published his works among physicists since late 1950s, whereas Kullback and Leibler published theirs among mathematicians in 1951.) Still, the leading negative sign in Jaynes's definition shows their slightly different mindsets: Jaynes wanted to maximize his uncertainty, while Kullback and Leibler wanted to minimize their information gain.

## 3.3 Example: nothing but a range and a mean

Having known $X \in [a, b]$, how should we update the variable's distribution $f(x)$ after we know its mean is $\mu$?

**Solution.** This question is conceptually identical to Example 2.3 although the computation here is more tedious due to the calculus of variations. We can first define a functional

$$J(f, r_0, r_1) = -\int_a^b f(x) \ln f(x) \mathrm{d}x - r_0 \left[\int_a^b f(x) \mathrm{d}x - 1\right] - r_1 \left[\int_a^b x f(x) \mathrm{d}x - \mu\right]$$

and extract its corresponding Lagrangian function

$$L(f) = -f \ln f - r_0 f - r_1 x f\,,$$

which yields

$$\frac{\delta J}{\delta f} = \frac{\mathrm{d}L}{\mathrm{d}f} = -\ln f - 1 - r_0 - r_1 x\,.$$

Therefore, $\dfrac{\delta J}{\delta f} = 0$ implies $f(x) = e^{-1-r_0-r_1 x}$. We can then employ the normalization condition $\displaystyle\int_a^b f(x) \mathrm{d}x = 1$ to get

$$f(x) = \frac{e^{-r_1 x}}{\int_a^b e^{-r_1 x} \mathrm{d}x} = \frac{-r_1 e^{-r_1 x}}{e^{-r_1 b} - e^{-r_1 a}}\,,$$

while $r_1$ remains to be solved using the constraint of mean $\displaystyle\int_a^b x f(x) \mathrm{d}x = \mu$. If $a = 0$ and $b \to \infty$, we can reduce the distribution to $f(x) = r_1 e^{-r_1 x}$ and obtain $r_1 = \dfrac{1}{\mu}$ analytically, resulting in the **exponential distribution** $f(x) = \dfrac{1}{\mu} e^{-x/\mu}$.

## 3.4 Example: nothing but a range, a mean, and a variance

What happens to $f(x)$ if we even know the variance of the previous variable $X$ is $\sigma^2$?

**Solution.** In general,

$$f(x) = \frac{e^{-r_2 (x-\mu)^2}}{\int_a^b e^{-r_2 (x-\mu)^2} \mathrm{d}x}$$

for some $r_2$ that satisfies the constraint of variance $\displaystyle\int_a^b (x-\mu)^2 f(x)\mathrm{d}x = \sigma^2$. If $a \to -\infty$ and $b \to \infty$, we can get $r_2 = \dfrac{1}{2\sigma^2}$ and reduce $f(x)$ to the **normal distribution**.

Exponential distribution and normal distribution are thus referred to as **maximum entropy distributions** (https://en.wikipedia.org/wiki/Maximum_entropy_probability_distribution), which are arguably the most rational guess when we know little about a variable.