Maximum Entropy Distributions

Joshua Goings

February 2021

1 Discrete uniform distribution

For the discrete case, consider N different possibilities, e.g. $i \in X = \{1, 2, \ldots, N\}$, and we have no other information other than the constraint that the probabilities p_i sum to 1. In this case, we maximize

$$H(p_i) - \sum_{i=1}^{N} p_i \ln p_i \tag{1}$$

subject to

$$\sum_{i=1}^{N} p_i = 1 \tag{2}$$

by the method of Lagrange multipliers. So we take the derivative with respect to p_i of the Lagrangian

$$\frac{\partial J(p_i, \lambda_0)}{\partial p_i} = \frac{\partial}{\partial p_i}\left(-\sum_{i=1}^{N} p_i \ln p_i + \lambda_0 \left(\sum_{i=1}^{N} p_i - 1\right)\right) \tag{3}$$

$$= -\ln p_i - 1 + \lambda_0 = 0 \tag{4}$$

and we set the result equal to zero, as this is a maximization. Note that the second derivative is negative, indicating we are at a maximum, e.g. $\frac{\partial^2 J(p,\lambda_0)}{\partial p_i^2} = -1/p_i$, and p_i is always positive. Therefore, we find that

$$p_i = \exp(\lambda_0 - 1) \tag{5}$$

and plugging into our normalization expression eq. (2) yields

$$\sum_{i=1}^{N} p_i = 1 \tag{6}$$

$$\sum_{i=1}^{N} \exp(\lambda_0 - 1) = 1 \tag{7}$$

$$N \exp(\lambda_0 - 1) = 1 \tag{8}$$

$$\exp(\lambda_0 - 1) = 1/N \tag{9}$$

which yields the discrete uniform probability distribution,

$$\boxed{p_i = 1/N} \tag{10}$$

This is the maximum entropy distribution for the case with N possible outcomes with no other information given (other than our probabilities are normalized).

2 Continuous uniform distribution

This is similar, to the discrete case we just saw, but now assume that the random variable x can take any value in $[a, b]$. Then we want to maximize

$$H(p(x)) = -\int_a^b p(x) \ln p(x) dx \tag{11}$$

subject to

$$\int_a^b p(x) dx = 1 \tag{12}$$

which gives us our Lagrangian

$$J(p(x), \lambda_0) = -\int_a^b p(x) \ln p(x) dx + \lambda_0 \left(\int_a^b p(x) dx - 1 \right) \tag{13}$$

differentiating the above with respect to $p(x)$ and setting to zero gives

$$\frac{\partial J(p(x), \lambda_0)}{\partial p(x)} = \frac{\partial}{\partial p(x)} \left(-\int_a^b p(x) \ln p(x) dx + \lambda_0 \left(\int_a^b p(x) dx - 1 \right) \right) \tag{14}$$

$$= -\ln p(x) - 1 + \lambda_0 = 0. \tag{15}$$

which gives us an expression for $p(x)$ as

$$p(x) = \exp(\lambda_0 - 1) \tag{16}$$

Like before, we can solve for λ_0 by plugging the result back in our normalization expression to get

$$\int_a^b p(x) dx = 1 \tag{17}$$

$$\int_a^b \exp(\lambda_0 - 1) \, dx = 1 \tag{18}$$

$$\exp(\lambda_0 - 1) \int_a^b dx = 1 \tag{19}$$

$$\exp(\lambda_0 - 1)(b - a) = 1 \tag{20}$$

$$\exp(\lambda_0 - 1) = \frac{1}{(b - a)} \tag{21}$$

yielding

$$\boxed{p(x) = \frac{1}{(b - a)}} \tag{22}$$

which is the continuous uniform distribution over $[a, b]$.

3 Cauchy distribution

The Cauchy distribution can be obtained in a similar way to the continuous uniform distribution, but in a particular geometric configuration. Consider the relationship between and angle θ_k and a point x_k on a line some distance away, as illustrated in fig. 1. In this case we want to consider the case where our random variable θ is an angle is on $[-\pi/2, \pi/2]$ and we don't know anything else about the underlying distribution other than it is normalized. So we maximize

$$H(p(\theta)) = -\int_{-\pi/2}^{\pi/2} p(\theta) \ln p(\theta) d\theta \tag{23}$$

subject to

$$\int_{-\pi/2}^{\pi/2} p(\theta) d\theta = 1 \tag{24}$$

Now, from the previous section we know that the MaxEnt procedure would result in $p(\theta) = 1/\pi$, which is obviously not the Cauchy distribution. To get the Cauchy distribution, we actually want to consider this case but in terms of a distribution over x, where the relationship between x and θ is given by $h \tan\theta = x - x_0$, where h and x_0 are arbitrary parameters.

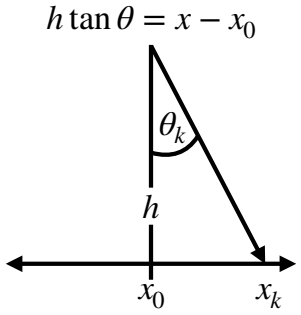

Figure 1: Relationship between θ and x. Here we want to consider the case where we don't know anything about the underlying probability distribution of θ other than that it is supported on $[-\pi/2, \pi/2]$ and that is is normalized, and we want this result as a distribution on x. h and x_0 are arbitrary parameters.

Since we have the relationship between θ and x, and the relationship is defined over $\theta \in [-\pi/2, \pi/2]$, we can do a change of variables to get $p(x)$ instead of $p(\theta)$ and then carry out the maximization of entropy. To do the change of variables, we need

$$h \tan\theta = x - x_0 \tag{25}$$

$$h \sec^2\theta \; d\theta = dx \tag{26}$$

$$h(\tan^2\theta + 1) \; d\theta = dx \tag{27}$$

$$h\left[\left(\frac{x - x_0}{h}\right)^2 + 1\right] \; d\theta = dx \tag{28}$$

$$\frac{1}{h}\left[(x - x_0)^2 + h^2\right] \; d\theta = dx \tag{29}$$

$$d\theta = \frac{h}{\left((x - x_0)^2 + h^2\right)} dx \tag{30}$$

so that

$$p(\theta) = p(x)\left|\frac{dx}{d\theta}\right| = p(x)\frac{1}{h}\left[(x - x_0)^2 + h^2\right] \tag{31}$$

For the limits on integration, we can easily see that $\tan(\pi/2) \to \infty$ and $\tan(-\pi/2) \to -\infty$, so $x \in (-\infty, \infty)$.

All together, we now want to maximize

$$H(p(\theta)) = -\int_{-\pi/2}^{\pi/2} p(\theta)\ln p(\theta)d\theta \tag{32}$$

$$H(p(x)) = -\int_{-\infty}^{\infty} \frac{p(x)\left((x-x_0)^2+h^2\right)}{h}\ln\left[\frac{p(x)\left((x-x_0)^2+h^2\right)}{h}\right]\frac{h}{\left((x-x_0)^2+h^2\right)}dx \tag{33}$$

$$H(p(x)) = -\int_{-\infty}^{\infty} p(x)\ln\left[\frac{p(x)\left((x-x_0)^2+h^2\right)}{h}\right]dx \tag{34}$$

$$H(p(x)) = -\int_{-\infty}^{\infty} p(x)\ln p(x)dx - \int_{-\infty}^{\infty} p(x)\ln\left[\frac{\left((x-x_0)^2+h^2\right)}{h}\right]dx \tag{35}$$

$$\tag{36}$$

subject to

$$\int_{-\infty}^{\infty} p(x)dx = 1 \tag{37}$$

therefore, our Lagrangian is

$$J(p(x),\lambda_0) = -\int_{-\infty}^{\infty} p(x)\ln p(x)dx \tag{38}$$

$$-\int_{-\infty}^{\infty} p(x)\ln\left[\frac{\left((x-x_0)^2+h^2\right)}{h}\right]dx \tag{39}$$

$$+\lambda_0\left(\int_{-\infty}^{\infty} p(x)dx - 1\right) \tag{40}$$

and carrying out the maximization yields

$$\frac{\partial J(p(x),\lambda_0)}{\partial p(x)} = -1 - \ln p(x) - \ln\left[\frac{\left((x-x_0)^2+h^2\right)}{h}\right] + \lambda_0 = 0 \tag{41}$$

which gives

$$p(x) = \exp\left(\lambda_0 - 1 - \ln\left[\frac{\left((x-x_0)^2+h^2\right)}{h}\right]\right) \tag{42}$$

$$= \frac{h}{\left((x-x_0)^2+h^2\right)}\exp(\lambda_0 - 1) \tag{43}$$

substituting this into the normalization condition allows us to eliminate the λ_0

$$\int_{-\infty}^{\infty} \frac{h}{\left((x-x_0)^2+h^2\right)}\exp(\lambda_0-1)dx = 1 \tag{44}$$

$$\exp(\lambda_0-1)\int_{-\infty}^{\infty} \frac{h}{\left((x-x_0)^2+h^2\right)}dx = 1 \tag{45}$$

$$\exp(\lambda_0-1)\pi = 1 \tag{46}$$

$$\exp(\lambda_0-1) = 1/\pi \tag{47}$$

$$\tag{48}$$

therefore, we get

$$\boxed{p(x) = \frac{h}{\pi\left((x - x_0)^2 + h^2\right)}} \tag{49}$$

So for sampling random angles and not knowing anything about the underlying distribution (e.g. θ is continuous uniform), we see that the resulting distribution over x is a Cauchy distribution, when x and θ have the relationship illustrated in fig. 1.

4 Exponential distribution

Now extend the continuous distribution to the case where we have a known expected value of $x = \mu$. We will limit ourselves to the support $[0, \infty)$. As before we maximize

$$H(p(x)) = -\int_0^\infty p(x) \ln p(x) dx \tag{50}$$

but now subject to

$$\int_0^\infty p(x) dx = 1, \qquad \int_0^\infty x p(x) dx = \mu \tag{51}$$

which gives us our Lagrangian

$$J(p(x), \lambda_0, \lambda_1) = -\int_0^\infty p(x) \ln p(x) dx + \lambda_0 \left(\int_0^\infty p(x) dx - 1\right) + \lambda_1 \left(\int_0^\infty x p(x) dx - \mu\right) \tag{52}$$

differentiating the above with respect to $p(x)$ and setting to zero gives

$$\frac{\partial J(p(x), \lambda_0, \lambda_1)}{\partial p(x)} = \frac{\partial}{\partial p(x)} \left(-\int_0^\infty p(x) \ln p(x) dx + \lambda_0 \left(\int_0^\infty p(x) dx - 1\right) + \lambda_1 \left(\int_0^\infty x p(x) dx - \mu\right)\right) \tag{53}$$

$$= -\ln p(x) - 1 + \lambda_0 + \lambda_1 x = 0. \tag{54}$$

which gives us an expression for $p(x)$ as

$$p(x) = \exp(\lambda_0 - 1) \exp(\lambda_1 x) \tag{55}$$

this results in the following expressions for our constraints:

$$\int_0^\infty p(x) dx = \exp(\lambda_0 - 1) \int_0^\infty \exp(\lambda_1 x) dx = \exp(\lambda_0 - 1) \left(\frac{-1}{\lambda_1}\right) = 1 \tag{56}$$

where $\frac{-1}{\lambda_1} < 0$ in order for the integral to converge. For the second constraint we have

$$\int_0^\infty x p(x) dx = \exp(\lambda_0 - 1) \int_0^\infty x \exp(\lambda_1 x) dx = \exp(\lambda_0 - 1) \left(\frac{1}{\lambda_1^2}\right) = \mu \tag{57}$$

Dividing the two constraints by each other allows us to solve for λ_1:

$$\frac{\exp(\lambda_0 - 1)}{\exp(\lambda_0 - 1)} \frac{-1/\lambda_1}{1/\lambda_1^2} = -\lambda_1 = 1/\mu \tag{58}$$

and then from the normalization we get

$$\exp(\lambda_0 - 1) \int_0^\infty \exp(-x/\mu) dx = \exp(\lambda_0 - 1)\, \mu = 1 \tag{59}$$

which means

$$\exp(\lambda_0 - 1) = 1/\mu \tag{60}$$

and so from knowing a fixed mean, we obtain the exponential distribution over $[0, \infty)$

$$\boxed{p(x) = \frac{1}{\mu} \exp\left(-\frac{x}{\mu}\right)} \tag{61}$$

5 Gaussian distribution

Now consider the case where we have a known mean μ and variance σ^2. We will consider $x \in (-\infty, \infty)$. Because variance is the expectation of the squared deviation of a random variable x from its mean, it suffices to introduce the constraint

$$\int_{-\infty}^{\infty} (x-\mu)^2 p(x) dx = \sigma^2 \tag{62}$$

which we will consider along with normalization when maximizing the entropy. This leads to the Lagrangian

$$J(p(x), \lambda_0, \lambda_1) = -\int_{-\infty}^{\infty} p(x) \ln p(x) dx + \lambda_0 \left(\int_{-\infty}^{\infty} p(x) dx - 1 \right) + \lambda_1 \left(\int_{-\infty}^{\infty} (x-\mu)^2 p(x) dx - \sigma^2 \right) \tag{63}$$

which, when differentiated with respect to $p(x)$ and set equal to zero, yields

$$\frac{\partial J(p(x), \lambda_0, \lambda_1)}{\partial p(x)} = -\ln p(x) - 1 + \lambda_0 + \lambda_1 (x-\mu)^2 = 0. \tag{64}$$

which gives us an expression for $p(x)$ as

$$p(x) = \exp(\lambda_0 - 1) \exp(\lambda_1 (x-\mu)^2) \tag{65}$$

For the first constraint, we find that

$$\int_{-\infty}^{\infty} p(x) dx = 1 \tag{66}$$

$$\int_{-\infty}^{\infty} \exp(\lambda_0 - 1) \exp(\lambda_1 (x-\mu)^2) dx = 1 \tag{67}$$

$$\exp(\lambda_0 - 1) \int_{-\infty}^{\infty} \exp(\lambda_1 (x-\mu)^2) dx = 1 \tag{68}$$

$$\exp(\lambda_0 - 1) \sqrt{\frac{\pi}{-\lambda_1}} = 1 \tag{69}$$

$$\exp(\lambda_0 - 1) = \sqrt{\frac{-\lambda_1}{\pi}}. \tag{70}$$

And for the second constraint we find

$$\int_{-\infty}^{\infty} (x-\mu)^2 p(x) dx = \sigma^2 \tag{71}$$

$$\int_{-\infty}^{\infty} \sqrt{\frac{-\lambda_1}{\pi}} (x-\mu)^2 \exp(\lambda_1 (x-\mu)^2) dx = \sigma^2 \tag{72}$$

$$\sqrt{\frac{-\lambda_1}{\pi}} \int_{-\infty}^{\infty} (x-\mu)^2 \exp(\lambda_1 (x-\mu)^2) dx = \sigma^2 \tag{73}$$

$$\sqrt{\frac{-\lambda_1}{\pi}} \cdot \frac{1}{2} \sqrt{\frac{\pi}{-\lambda_1^3}} = \sigma^2 \tag{74}$$

$$\lambda_1 = -\frac{1}{2\sigma^2}. \tag{75}$$

Which allows us to say that

$$\exp(\lambda_0 - 1) = \frac{1}{\sqrt{2\pi\sigma^2}} \tag{76}$$

Putting this all together yields the Gaussian, or normal distribution

$$\boxed{p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right)} \tag{77}$$

6 Bernoulli distribution

Moving back to the discrete case, let's consider when a random variable k is either 0 or 1, i.e. $k \in \{0, 1\}$ and the expected value of k is $\bar{p}$. This results in the Lagrangian

$$J(p_k, \lambda_0, \lambda_1) = -\sum_k p_k \ln p_k + \lambda_0 \left(\sum_k p_k - 1 \right) + \lambda_1 \left(\sum_k k p_k - \bar{p} \right) \tag{78}$$

Maximizing this Lagrangian gives

$$\frac{\partial J(p_k, \lambda_0)}{\partial p_k} = -1 - \ln p_k + \lambda_0 + \lambda_1 k = 0 \tag{79}$$

which yields the probability mass function p_k

$$p_k = \exp(\lambda_0 - 1) \exp(\lambda_1 k) \tag{80}$$

Taking care of the first constraint,

$$\sum_k \exp(\lambda_0 - 1) \exp(\lambda_1 k) = 1 \tag{81}$$

$$\exp(\lambda_0 - 1) \sum_k \exp(\lambda_1 k) = 1 \tag{82}$$

$$\exp(\lambda_0 - 1) = \frac{1}{\sum_k \exp(\lambda_1 k)} \tag{83}$$

$$\exp(\lambda_0 - 1) = \frac{1}{\exp(\lambda_1 \cdot 0) + \exp(\lambda_1 \cdot 1)} \tag{84}$$

$$\exp(\lambda_0 - 1) = \frac{1}{1 + \exp(\lambda_1)} \tag{85}$$

since k is either 0 or 1. Taking care of the second constraint,

$$\sum_k k \cdot \exp(\lambda_0 - 1) \exp(\lambda_1 k) = \bar{p} \tag{86}$$

$$\exp(\lambda_0 - 1) \sum_k k \cdot \exp(\lambda_1 k) = \bar{p} \tag{87}$$

$$\frac{\exp(\lambda_1)}{1 + \exp(\lambda_1)} = \bar{p} \tag{88}$$

again, since k is either 0 or 1. Then we can solve for λ_1,

$$\lambda_1 = \ln \left(\frac{\bar{p}}{1 - \bar{p}} \right) \tag{89}$$

which means that

$$\exp(\lambda_0 - 1) = \frac{1}{1 + \frac{\bar{p}}{1 - \bar{p}}} = (1 - \bar{p}) \tag{90}$$

Putting this all together we have

$$p_k = (1 - \bar{p}) \exp \left(k \cdot \ln \left(\frac{\bar{p}}{1 - \bar{p}} \right) \right) \tag{91}$$

$$= (1 - \bar{p}) \left(\frac{\bar{p}}{1 - \bar{p}} \right)^k \tag{92}$$

$$= \bar{p}^k (1 - \bar{p})^{k-1} \tag{93}$$

Which is the Bernoulli distribution

$$\boxed{p_k = \bar{p}^k (1 - \bar{p})^{k-1}} \tag{94}$$

for when $0 \leq \bar{p} \leq 1$ and k is either 0 or 1.

7 Binomial distribution

This is the case where we compute the probability for having N successes in M trials. The constraint here is the expected number of successes $\langle N \rangle$. Note that this will depend on the number of trials, and since we only care abut the number of successes, and not the order in which they were taken, we need to use the generalized form of entropy

$$H(p) = -\sum_N^M p_N \ln \frac{p_N}{m(N)} \tag{95}$$

where $m(N)$ is the Lebesgue measure, which accounts for the fact that we need to account for the fact that we are indifferent to the number of ways N can be accomplished. This is essentially the prior probability we assign to the different outcomes. For example, in the uniform distribution we had no reason to favor one proposition p_i over another, thus the principle of indifference led us to assign $m(i) = 1$ for all i, and the result led to each outcome being equally likely. But this is not always the case in combinatoric problems, for example, since (without replacement) there are 4 ways to pick a unique object out of a set of 4 unique objects, but 6 ways to pick out 2 objects out of the same set. So we would not expect the probabilities for 2 objects to be on the same scale as picking out 1 object; our prior information leads us to favor 4 choose 2 over 4 choose 1 – there are more ways it could happen. The measure $m(N)$ allows us to account for that.

Now, moving on, we are in addition to the normal entropy and normalization, constrained by the information

$$\sum_N^M N p_N = \langle N \rangle = \mu \tag{96}$$

Therefore, our Lagrangian reads

$$J(p_N, \ldots) = -\sum_N^M p_N \ln \frac{p_N}{m(N)} + \lambda_0 \left(\sum_N^M p_N - 1 \right) + \lambda_1 \left(\sum_N^M N p_N - \mu \right) \tag{97}$$

which leads to the maximization

$$\frac{\partial J}{\partial p_N} = 0 = -1 - \ln \frac{p_N}{m(N)} + \lambda_0 + \lambda_1 N \tag{98}$$

so that

$$p_N = m(N) \cdot \exp(\lambda_0 - 1) \exp(\lambda_1)^N \tag{99}$$

$$= \frac{M!}{N!(M-N)!} \exp(\lambda_0 - 1) \exp(\lambda_1)^N \tag{100}$$

and we chose the combinatoric measure because for each possible number of successes N, there are $M!/N!(M-N)!$ different ways of achieving this given M trials.

Solving for the first constraint:

$$1 = \sum_N^M p_N = \exp(\lambda_0 - 1) \sum_N^M \frac{M!}{N!(M-N)!} \exp(\lambda_1)^N \cdot (1)^{M-N} \tag{101}$$

$$= \exp(\lambda_0 - 1) \exp(\lambda_1 + 1)^M \tag{102}$$

$$\implies \exp(\lambda_0 - 1) = \exp(\lambda_1 + 1)^{-M} \tag{103}$$

where we multiplied by 1 to the power of $M - N$, which just equals one, in the first line in order to make use of the binomial formula and eliminate the sum. The inversion in the last line is made possible because the exponential is always greater than zero.

For the next constraint,

$$\mu = \sum_{N}^{M} N \exp(\lambda_1 + 1)^{-M} \frac{M!}{N!(M-N)!} \exp(\lambda_1)^N \tag{104}$$

$$= \exp(\lambda_1 + 1)^{-M} \sum_{N}^{M} N \frac{M!}{N!(M-N)!} \exp(\lambda_1)^N \tag{105}$$

$$= \exp(\lambda_1 + 1)^{-M} \cdot \exp(\lambda_1) \cdot M \cdot (\exp(\lambda_1) + 1)^{M-1} \tag{106}$$

$$= M \cdot \frac{\exp(\lambda_1)}{\exp(\lambda_1) + 1} \tag{107}$$

In fairness, I used WolframAlpha to finally eliminate the sum after the second line. If we let $p \coloneqq \mu/M$, then we can finally see that

$$\exp(\lambda_1) = \frac{p}{1-p} \tag{108}$$

which we can obtain because $p > 0$.

Okay. Putting it all together now:

$$\exp(\lambda_0 - 1) = \left(\frac{p}{1-p} + 1\right)^{-M} = \left(\frac{1}{1-p}\right)^{-M} = (1-p)^M \tag{109}$$

and

$$\exp(\lambda_1)^N = \left(\frac{p}{1-p}\right)^N = \left(\frac{1-p}{p}\right)^{-N} = \left(\frac{1}{p} - 1\right)^{-N} \tag{110}$$

which lets us finally show that

$$p_N = \frac{M!}{N!(M-N)!} \cdot (1-p)^M \cdot \left(\frac{1}{p} - 1\right)^{-N} \tag{111}$$

or, simply,

$$\boxed{p_N = \frac{M!}{N!(M-N)!} (p)^N (1-p)^{M-N}} \tag{112}$$

which is the binomial distribution.