Chapter 10

Principle of Maximum Entropy

In Chapter 9, we discussed the technique of estimating input probabilities of a process that is consistent with known constraints expressed in terms of averages, or expected values, of one or more quantities, but is otherwise as unbiased as possible. This technique was described there for the simple case of one constraint and three input events, in which case the technique can be carried out analytically. Now it is described here more generally.

10.1 Problem Setup

Before the Principle of Maximum Entropy can be used the problem domain needs to be set up. In cases involving physical systems, this means that the various states in which the system can exist need to be identified, and all the parameters involved in the constraints known. For example, the energy, electric charge, and other quantities associated with each of the quantum states is assumed known. It is not assumed in this step which particular state the system is actually in (which state is occupied) indeed it is assumed that we do not know and cannot know this with certainty, and so instead we deal with the probability of each of the states being occupied. In applications to nonphysical systems, again the various possible events have to be enumerated and the properties of each state known, particularly the values associated with each of the constraints. In these notes we will apply the general mathematical derivation to two examples, one a crude business model, and the other a crude model of a physical system.

10.1.1 Berger's Burgers

This example was used in previous chapters of these notes dealing with inference (Chapter 8) and the simple form of the Principle of Maximum Entropy (Chapter 9). A fast-food restaurant offers three meals: burger, chicken, and fish. Now we assume that the menu has been extended to include a gourmet low-fat tofu meal. The price, Calorie count, and probability of each meal being delivered cold are as listed in Table 10.1.

10.1.2 Magnetic Dipole Model

An array of magnetic dipoles (think of them as tiny magnets) are subjected to an externally applied magnetic field H and therefore the energy of the system depends on their orientations and on the applied field. Our system contains two such dipoles, but it will from time to time be able to interchange information and energy with either of two environments, which are much larger collections of dipoles. Each dipole, both in the system and in its two environments, can be either "up" or "down," so there are four possible states for

Author: Paul Penfield, Jr.

Item	Entree	Cost	Calories	Probability of arriving hot	Probability of arriving cold
Value Meal 1	Burger	\$1.00	1000	0.5	0.5
Value Meal 2	Chicken	\$2.00	600	0.8	0.2
Value Meal 3	Fish	\$3.00	400	0.9	0.1
Value Meal 4	Tofu	\$8.00	200	0.6	0.4

Table 10.1: Berger's Burgers

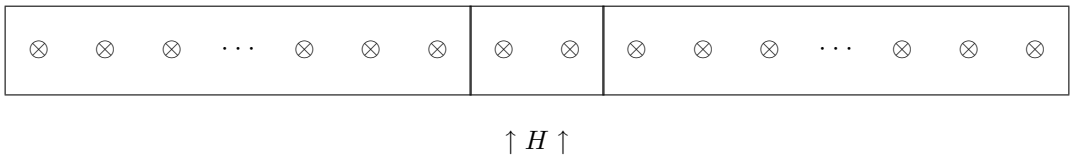

Figure 10.1: Dipole moment example.
(Each dipole can be either up or down.)

the system, "up-up," "up-down," "down-up," and "down-down." The energy of a dipole is proportional to the applied field and depends on its orientation, and the energy of each state is the sum of the energies of the two dipoles.

State	Alignment	Energy
A	up-up	$-m_d H$
B	up-down	0
C	down-up	0
D	down-down	$m_d H$

Table 10.2: Magnetic Dipole Moments

The constant m_d is expressed in Joules per Tesla, and its value depends on the physics of the particular dipole. If the dipoles are electron spins, then $m_d = 2\mu_B\mu_0$ where $\mu_0 = 4\pi \times 10^{-7}$ henries per meter (in rationalized MKS units) is the permeability of free space, $\mu_B = \hbar e/2m_e = 9.272 \times 10^{-24}$ Joules per Tesla is the Bohr magneton, and where $\hbar = h/2\pi$, $h = 6.626 \times 10^{-34}$ Joule-seconds is Plank's constant, $e = 1.602 \times 10^{-19}$ coulombs is the magnitude of the charge of an electron, and $m_e = 9.109 \times 10^{-31}$ kilograms is the rest mass of an electron.

In Figure 10.1, the system is shown between two environments, and there are barriers between the environments and the system (represented by the vertical bars) which prevent interaction (later we will remove the barriers to permit interaction). The dipoles, in both the system and the environments, are shown as $\otimes$ and may be either spin-up or spin-down. The magnetic field is shown applied to the system only, not to the environments.

The virtue of a model with only two dipoles is that it is simple enough that the calculations can be carried out easily. Such a model is, of course, hopelessly simplistic and cannot be expected to lead to numerically accurate results. A more realistic model would require so many dipoles and so many states that practical computations on the collection could never be done. For example, if our system is a chemical element with one mole of material (an amount with mass in grams equal to the atomic weight of the element) then there would be Avogadro's number $N_A = 6.02 \times 10^{23}$ of atoms, and a correspondingly large number of electron spins, so the number of possible states would be 2 raised to that power. Just how large this number is can be appreciated by noting that the earth contains no more than 2^{170} atoms, and the visible universe has roughly 2^{265} atoms; both of these numbers are way less than the number of states in our model. Even if we

are less ambitious and want to compute with a much smaller sample, say 200 spins, and want to represent in our computer the probability of each state (using only 8 bits per state), we would still need more bytes of memory than there are atoms in the earth. Clearly it is impossible to compute with so many states, so the techniques described in these notes cannot be carried through in detail. Nevertheless there are certain conclusions and general relationships we will be able to establish.

10.2 Probabilities

Although the problem has been set up, we do not know which actual state the system is in. To express what we do know despite this ignorance, or uncertainty, we assume that each of the possible states A_i has some probability of occupancy $p(A_i)$ where i is an index running over the possible states. A probability distribution $p(A_i)$ has the property that each of the probabilities is between 0 and 1 (possibly being equal to either 0 or 1), and (since the input events are mutually exclusive and exhaustive) the sum of all the probabilities is 1:

$$1 = \sum_i p(A_i) \tag{10.1}$$

As has been mentioned before, two observers may, because of their different knowledge, use different probability distributions. Therefore probability, and all quantities that are based on probabilities, are subjective, or observer-dependent.

10.3 Entropy

Our uncertainty is expressed quantitatively by the information which we do not have about the state occupied. This information is

$$S = \sum_i p(A_i) \log_2 \left(\frac{1}{p(A_i)} \right) \tag{10.2}$$

Information is measured in bits, as a consequence of the use of logarithms to base 2 in the Equation 10.2.

In dealing with real physical systems, with a huge number of states and therefore an entropy that is a very large number of bits, it is convenient to multiply the summation above by Boltzmann's constant $k_B = 1.381 \times 10^{-23}$ Joules per Kelvin, and also use natural logarithms rather than logarithms to base 2. Then S would be expressed in Joules per Kelvin:

$$S = k_B \sum_i p(A_i) \ln \left(\frac{1}{p(A_i)} \right) \tag{10.3}$$

In the context of both physical systems and communication systems the uncertainty is known as the entropy. Note that in general the entropy, because it is expressed in terms of probabilities, depends on the observer. One person may have different knowledge of the system from another, and therefore would calculate a different numerical value for entropy.

10.4 Constraints

The entropy has its maximum value when all probabilities are equal (we assume the number of possible states is finite), and the resulting value for entropy is the logarithm of the number of states, with a possible scale factor like k_B. If we have no additional information about the system, then such a result seems reasonable. However, if we have additional information then we ought to be able to find a probability distribution that is better in the sense that it has less uncertainty.

For simplicity we consider only one such constraint here. We assume that we know the expected value of some quantity (the Principle of Maximum Entropy can handle multiple constraints but the mathematical procedures and formulas become more complicated). The quantity in question is one for which each of the states of the system has its own amount, and the expected value is found by averaging the values corresponding to each of the states, taking into account the probabilities of those states. Thus if there is a quantity G for which each of the states has a value $g(A_i)$ then we want to consider only those probability distributions for which the expected value is G

$$G = \sum_i p(A_i)g(A_i) \tag{10.4}$$

Of course this constraint cannot be achieved if G is less than the smallest $g(A_i)$ or greater than the largest $g(A_i)$.

10.4.1 Examples

For our Berger's Burgers example, suppose we are told that the average price of a meal is \$2.50, and we want to estimate the separate probabilities of the various meals without making any other assumptions. Then our constraint would be

$$\$2.50 = \$1.00p(B) + \$2.00p(C) + \$3.00p(F) + \$8.00p(T) \tag{10.5}$$

Note that the probabilities are dimensionless and so both the expected value of the constraint and the individual values must be expressed in the same units, in this case dollars.

For our magnetic-dipole example, assume the energies for states A, B, C, and D are denoted $e(i)$ where i is one of A, B, C, or D and assume the expected value of the energy is known to be some value E. All these energies are expressed in Joules. Then

$$E = e(A)p(A) + e(B)p(B) + e(C)p(C) + e(D)p(D) \tag{10.6}$$

The energies $e(A)\ldots e(D)$ depend on the externally applied magnetic field H and this parameter, which will be carried through the derivation, will end up playing an important role. If the formulas for the $e(i)$ from Table 10.2 are used here,

$$E = m_d H(p(A) - p(D)) \tag{10.7}$$

10.5 Maximum Entropy, Analytic Form

The **Principle of Maximum Entropy** is based on the premise that when estimating the probability distribution, you should select that distribution which leaves you the largest remaining uncertainty (i.e., the maximum entropy) consistent with your constraints. That way you have not introduced any additional assumptions or biases into your calculations.

This principle was used in the last chapter in the simple case of three probabilities and one constraint. The entropy could be maximized analytically. Using the constraint and the fact that the probabilities add up to 1, we expressed two of the unknown probabilities in terms of the third.

Next, the possible range of values of the probabilities was determined using the fact each of the three lies between 0 and 1. Then, these expressions were substituted into the formula for entropy S so that it was expressed in terms of a single probability. Then any of several techniques could be used to find the value of that probability for which S is the largest.

10.6 Maximum Entropy, Lagrange Multiplier Form

This analytical technique of section 10.5 does not work when there are more than three possible states and only one constraint (this is because it is only practical when the constraint can be used to express the entropy in terms of a single variable). When there are, say, four unknowns and two equations, the entropy would be a function of two variables, rather than one. Although it would be possible to search in a plane for the maximum entropy, this would only work for four probabilities in the distribution. If there were five, searching in a three-dimensional space would be necessary. A more general procedure is clearly needed, and this is provided by the use of Lagrange Multipliers.

We will develop this procedure here for the case of a single constraint and an arbitrary (finite) number of probabilities, although multiple constraints can be handled also. We assume the average value of some quantity with values $g(A_i)$ associated with the various events A_i is G. Thus

$$1 = \sum_i p(A_i) \tag{10.8}$$

$$G = \sum_i p(A_i) g(A_i) \tag{10.9}$$

The entropy associated with this probability distribution is

$$S = \sum_i p(A_i) \log_2 \left(\frac{1}{p(A_i)} \right) \tag{10.10}$$

when expressed in bits. For physical systems the entropy is more conveniently expressed in Joules per Kelvin,

$$S = k_B \sum_i p(A_i) \ln \left(\frac{1}{p(A_i)} \right) \tag{10.11}$$

10.6.1 Lagrange Multipliers

The technique of Lagrange Multipliers is named after the French mathematician, Joseph-Louis Lagrange (1736 - 1813)[1]. Instead of using the constraint equations to reduce the number of unknowns, we increase the number of unknowns. We define the "Lagrange multipliers" α and β and then the "Lagrangian" function L:

$$L = S - (\alpha - \log_2 e) \left(\sum_i p(A_i) - 1 \right) - \beta \left(\sum_i g(A_i) p(A_i) - G \right) \tag{10.12}$$

where $e = 2.7183$ is the base of natural logarithms, so $\log_2 e = 1.4427$. The Lagrange multiplier α is measured in bits, just like entropy, and β is measured in bits per unit of G (for our examples, bits per dollar or bits per Joule). If S is expressed in Joules per Kelvin, and natural logarithms are used in the entropy definition, then the formula for L is slightly different:

$$L = S - k_B(\alpha - 1) \left(\sum_i p(A_i) - 1 \right) - k_B \beta \left(\sum_i g(A_i) p(A_i) - G \right) \tag{10.13}$$

and the units for α and β are no longer in bits: α is dimensionless and β is expressed in the inverse of the units for G.

The Principle of Maximum Entropy is carried out by finding α, β, and all $p(A_i)$ such that L is made the largest. These values of $p(A_i)$ also make S the largest it can be, subject to the constraints. By introducing two new variables, we have (surprisingly) simplified the problem so that all the quantities of interest can be expressed in terms of one of the variables, and a procedure can be followed to find that one.

[1] See a biography of him at http://www-groups.dcs.st-andrews.ac.uk/~history/Mathematicians/Lagrange.html

Since α only appears once in the expression for L, the quantity that multiplies it must be zero for the values that maximize L (otherwise a small change in a could increase L). Similarly, β only appears once in the expression for L. Thus in the general case the $p(A_i)$ that we are seeking must satisfy

$$0 = \sum_i p(A_i) - 1 \tag{10.14}$$

$$0 = \sum_i g(A_i)p(A_i) - G \tag{10.15}$$

Maximizing L with respect to each of the $p(A_i)$ is done by differentiating L with respect to one of the $p(A_i)$ while keeping α, β, and all other $p(A_i)$ constant. The result is

$$\log_2\left(\frac{1}{p(A_i)}\right) = \alpha + \beta g(A_i) \tag{10.16}$$

or

$$p(A_i) = 2^{-\alpha}2^{-\beta g(A_i)} \tag{10.17}$$

(a similar formula is obtained if S is measured in Joules per Kelvin.)

Once α and β are known, the probabilities $p(A_i)$ can be found and, if desired, the entropy S can be calculated. In fact, if S is needed, it can be calculated directly, without evaluating the $p(A_i)$ – this is necessary if there are dozens or more probabilities to deal with. This short-cut is found by taking the formula for $\log_2(1/p(A_i))$ just above, multiplying it by $p(A_i)$, and summing over i. The left-hand side is S and the right-hand side simplifies because α and β are independent of i. The result is

$$S = \alpha + \beta G \tag{10.18}$$

We still need to find α and β. In the general case, if the equation just above for $p(A_i)$ is summed over the probabilities, the result is, after a little algebra,

$$\alpha = \log_2\left(\sum_i 2^{-\beta g(A_i)}\right) \tag{10.19}$$

If β is known then α can be calculated using this equation. But we still need to find β. This is more difficult; in fact most of the computational difficulty associated with the Principle of Maximum Entropy lies in this step. If there are a modest number of states and only one constraint in addition to the equation involving the sum of the probabilities, this step is not hard, as we will see. If there are more constraints this step becomes increasingly complicated, and if there are a large number of states the calculations cannot be done using this method. In the case of more realistic models for physical systems, this summation is impossible to calculate, although the general relations among the quantities other than $p(A_i)$ remain valid.

To find β, take the formula for $p(A_i)$ just above, multiply it by $g(A_i)$ and by 2^{α}, and sum over the probabilities. The left hand side becomes $G2^{\alpha}$, because neither α nor G depends on i. We already have an expression for α in terms of β, so the left hand side becomes $\sum_i G2^{-\beta g(A_i)}$. The right hand side becomes $\sum_i g(A_i)2^{-\beta g(A_i)}$. Thus,

$$0 = \sum_i (g(A_i) - G)2^{-\beta g(A_i)} \tag{10.20}$$

If this equation is multiplied by $2^{\beta G}$, the result is

$$0 = f(b) \tag{10.21}$$

where the function $f(\beta)$ is

$$f(b) = \sum_i (g(A_i) - G)2^{-b(g(A_i)-G)} \tag{10.22}$$

This is the fundamental equation that is to be solved. The function $f(\beta)$ depends on the model of the problem (i.e., the various $g(A_i)$), and on the average value G, and that is all. It does not depend on α or the probabilities $p(A_i)$. For the given value of G, the value of β that maximizes L and therefore S is the value for which $f(\beta) = 0$. Even if there are so many states that it is not possible to represent all their probabilities, it still may be possible to find β by evaluating $f(\beta)$ without actually performing the summation.

How do we know that there is any value of β for which $f(\beta) = 0$? First, notice that since G lies between the smallest and the largest $g(A_i)$, there is at least one $g(A_i)$ for which $(g(Ai) - G)$ is positive and at least one for which it is negative. It is not difficult to show that $f(\beta)$ is a monotonic function of β, in the sense that if $\beta_2 > \beta_1$ then $f(\beta_2) < f(\beta_1)$. For large positive values of β, the dominant term in the sum is the one that has the smallest value of $g(A_i)$, and hence f is negative. Similarly, for large negative values of β, f is positive. It must therefore be zero for one and only one value of β (this reasoning relies on the fact that $f(\beta)$ is a continuous function.)

10.6.2 Examples

For the Berger's Burgers example, suppose that you are told the average meal price is \$2.50, and you want to estimate the probabilities $p(B)$, $p(C)$, $p(F)$, and $p(T)$.

$$1 = p(B) + p(C) + p(F) + p(T) \tag{10.23}$$

$$0 = \$1.00p(B) + \$2.00p(C) + \$3.00p(F) + \$8.00p(T) - \$2.50 \tag{10.24}$$

$$S = p(B)\log_2\left(\frac{1}{p(B)}\right) + p(C)\log_2\left(\frac{1}{p(C)}\right) + p(F)\log_2\left(\frac{1}{p(F)}\right) + p(T)\log_2\left(\frac{1}{p(T)}\right) \tag{10.25}$$

$$\begin{aligned} L = {} & S - (\alpha - \log_2 e)(p(B) + p(C) + p(F) + p(T) - 1) \\ & -\beta(\$1.00p(B) + \$2.00p(C) + \$3.00p(F) + \$8.00p(T) - \$2.50) \end{aligned} \tag{10.26}$$

The entropy is the largest, subject to the constraints, if

$$p(B) = 2^{-\alpha}2^{-\beta\$1.00} \tag{10.27}$$
$$p(C) = 2^{-\alpha}2^{-\beta\$2.00} \tag{10.28}$$
$$p(F) = 2^{-\alpha}2^{-\beta\$3.00} \tag{10.29}$$
$$p(T) = 2^{-\alpha}2^{-\beta\$8.00} \tag{10.30}$$

where

$$a = \log_2(2^{-\beta\$1.00} + 2^{-\beta\$2.00} + 2^{-\beta\$3.00} + 2^{-\beta\$8.00}) \tag{10.31}$$

and β is the value for which $f(\beta) = 0$ where

$$f(b) = \$0.50 \times 2^{-\$0.50\beta} + \$5.50 \times 2^{-\$5.50\beta} - \$1.50 \times 2^{\$1.50\beta} - \$0.50 \times 2^{\$0.50\beta} \tag{10.32}$$

A little trial and error (or use of a zero-finding program) gives $\beta = 0.2586$ bits/dollar, $\alpha = 1.2371$ bits, $p(B) = 0.3546$, $p(C) = 0.2964$, $p(F) = 0.2478$, $p(T) = 0.1011$, and $S = 1.8835$ bits. The entropy is smaller than the 2 bits which would be required to encode a single order of one of the four possible meals using a

fixed-length code. This is because knowledge of the average price reduces our uncertainty somewhat. If more information is known about the orders then a probability distribution that incorporates that information would have even lower entropy.

For the **magnetic dipole example**, we carry the derivation out with the magnetic field H set at some unknown value. The results all depend on H as well as E.

$$1 = p(A) + p(B) + p(C) + p(D) \tag{10.33}$$

$$\begin{aligned} E &= e(A)p(A) + e(B)p(B) + e(C)p(C) + e(D)p(D) \\ &= m_d H(p(A) - p(D)) \end{aligned} \tag{10.34}$$

$$S = p(A)\log_2\left(\frac{1}{p(A)}\right) + p(B)\log_2\left(\frac{1}{p(B)}\right) + p(C)\log_2\left(\frac{1}{p(C)}\right) + p(D)\log_2\left(\frac{1}{p(D)}\right) \tag{10.35}$$

$$L = S - (\alpha - \log_2 e)(p(A) + p(B) + p(C) + p(D) - 1) - \beta(m_d H p(A) - m_d H p(D) - E) \tag{10.36}$$

The entropy is the largest, for the energy E and magnetic field H, if

$$p(A) = 2^{-a}2^{-\beta m_d H} \tag{10.37}$$
$$p(B) = 2^{-\alpha} \tag{10.38}$$
$$p(C) = 2^{-\alpha} \tag{10.39}$$
$$p(D) = 2^{-a}2^{\beta m_d H} \tag{10.40}$$

where

$$\alpha = \log_2(2^{-\beta m_d H} + 2 + 2^{\beta m_d H}) \tag{10.41}$$

and β is the value for which $f(\beta) = 0$ where

$$f(\beta) = (m_d H - E)2^{-\beta(m_d H - E)} - 2E2^{\beta E} - (m_d H + E)2^{\beta(m_d H + E)} \tag{10.42}$$

If there were many more than four possible states, the procedure to calculate β would have been impractical. We therefore ask, in Chapter 12 of these notes, what we can tell about the various quantities even if we never actually calculate numerical values for them.