Chapter 9

Principle of Maximum Entropy

Section 8.2 presented the technique of estimating input probabilities of a process that are unbiased but consistent with known constraints expressed in terms of averages, or expected values, of one or more quantities. This technique, the Principle of Maximum Entropy, was developed there for the simple case of one constraint and three input events, in which case the technique can be carried out analytically. It is described here for the more general case.

9.1 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 cannot ever know this with certainty, and so we deal instead with the probability of each of the states being occupied. In applications to nonphysical systems, the various possible events have to be enumerated and the properties of each determined, particularly the values associated with each of the constraints. In this Chapter we will apply the general mathematical derivation to two examples, one a business model, and the other a model of a physical system (both very simple and crude).

9.1.1 Berger's Burgers

This example was used in Chapter 8 to deal with inference and the analytic form of the Principle of Maximum Entropy. A fast-food restaurant offers three meals: burger, chicken, and fish. Now we suppose 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 listed in Table 9.1.

9.1.2 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. For simplicity our system contains only one such dipole, which from time to time is able to interchange information and energy with either of two environments, which are much larger collections of dipoles. Each

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

Table 9.1: Berger's Burgers

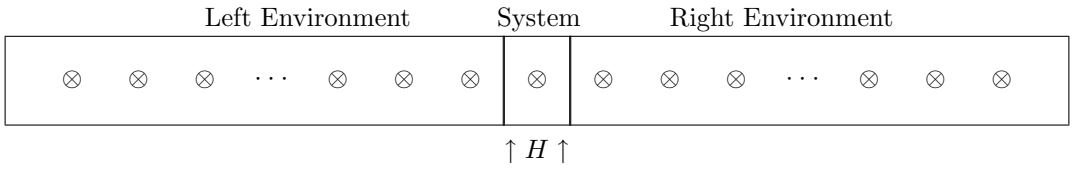

Figure 9.1: Dipole moment example.
(Each dipole can be either up or down.)

dipole, both in the system and in its two environments, can be either "up" or "down." The system has one dipole so it only has two states, corresponding to the two states for that dipole, "up" and "down" (if the system had n dipoles it would have 2^n states). The energy of each dipole is proportional to the applied field and depends on its orientation; the energy of the system is the sum of the energies of all the dipoles in the system, in our case only one such.

State	Alignment	Energy
U	up	$-m_d H$
D	down	$m_d H$

Table 9.2: Magnetic Dipole Moments

The constant m_d is expressed in Joules per Tesla, and its value depends on the physics of the particular dipole. For example, the dipoles might be electron spins, in which case $m_d = 2\mu_B\mu_0$ 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 9.1, the system is shown between two environments, and there are barriers between the environments and the system (represented by vertical lines) which prevent interaction (later we will remove the barriers to permit interaction). The dipoles, in both the system and the environments, are represented by the symbol $\otimes$ and may be either spin-up or spin-down. The magnetic field shown is applied to the system only, not to the environments.

The virtue of a model with only one dipole 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, a mole of a chemical element is a small amount by everyday standards, but it contains Avogadro's number $N_A = 6.02252 \times 10^{23}$ of atoms, and a correspondingly large number of electron spins; 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 about 2^{265} atoms; both of these numbers are way less than the number of states in that 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.

9.2 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:

$$1 = \sum_i p(A_i) \tag{9.1}$$

As has been mentioned before, two observers may, because of their different knowledge, use different probability distributions. In other words, probability, and all quantities that are based on probabilities, are subjective, or observer-dependent. The derivations below can be carried out for any observer.

9.3 Entropy

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

$$S = \sum_i p(A_i) \log_2 \left(\frac{1}{p(A_i)} \right) \tag{9.2}$$

Information is measured in bits, as a consequence of the use of logarithms to base 2 in the Equation 9.2.

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:

$$S = k_B \sum_i p(A_i) \ln \left(\frac{1}{p(A_i)} \right) \tag{9.3}$$

In the context of both physical systems and communication systems the uncertainty is known as the entropy. Note that because the entropy is expressed in terms of probabilities, it also depends on the observer, so two people with different knowledge of the system would calculate a different numerical value for entropy.

9.4 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 in the form of constraints then the assumption of equal probabilities would probably not be consistent with those constraints. Our objective is to find the probability distribution that has the greatest uncertainty, and hence is as unbiased as possible.

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 are 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 a known value $\widetilde{G}$

$$\widetilde{G} = \sum_i p(A_i)g(A_i) \tag{9.4}$$

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

9.4.1 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

$$\$2.50 = \$1.00p(B) + \$2.00p(C) + \$3.00p(F) + \$8.00p(T) \tag{9.5}$$

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

$$\widetilde{E} = e(U)p(U) + e(D)p(D) \tag{9.6}$$

The energies $e(U)$ and $e(D)$ depend on the externally applied magnetic field H. 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 9.2 are used here,

$$\widetilde{E} = m_d H[p(D) - p(U)] \tag{9.7}$$

9.5 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 Chapter 8 for 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 that 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.

This analytical technique does not extend to cases with more than three possible states and only one constraint. It is only practical because the constraint can be used to express the entropy in terms of a single variable. If there are, say, four unknowns and two equations, the entropy would be left as a function of two variables, rather than one. It would be necessary to search for its maximum in a plane. Perhaps this seems feasible, but what if there were five unknowns? (Or ten?) Searching in a space of three (or eight) dimensions would be necessary, and this is much more difficult.

A different approach is developed in the next section, one well suited for a single constraint and many probabilities.

9.6 Maximum Entropy, Single Constraint

Let us assume the average value of some quantity with values $g(A_i)$ associated with the various events A_i is known; call it $\widetilde{G}$ (this is the constraint). Thus there are two equations, one of which comes from the constraint and the other from the fact that the probabilities add up to 1:

$$1 = \sum_i p(A_i) \tag{9.8}$$

$$\widetilde{G} = \sum_i p(A_i) g(A_i) \tag{9.9}$$

where $\widetilde{G}$ cannot be smaller than the smallest $g(A_i)$ or larger than the largest $g(A_i)$.

The entropy associated with this probability distribution is

$$S = \sum_i p(A_i) \log_2 \left(\frac{1}{p(A_i)} \right) \tag{9.10}$$

when expressed in bits. In the derivation below this formula for entropy will be used. It works well for examples with a small number of states. In later chapters of these notes we will start using the more common expression for entropy in physical systems, expressed in Joules per Kelvin,

$$S = k_B \sum_i p(A_i) \ln \left(\frac{1}{p(A_i)} \right) \tag{9.11}$$

9.6.1 Dual Variable

Sometimes a problem is clarified by looking at a more general problem of which the original is a special case. In this case, rather than focusing on a specific value of G, let's look at all possible values of G, which means the range between the smallest and largest values of $g(A_i)$. Thus G becomes a variable rather than a known value (the known value will continue to be denoted $\widetilde{G}$ here). Then rather than express things in terms of G as an independent variable, we will introduce a new **dual variable**, which we will call β, and express all the quantities of interest, including G, in terms of it. Then the original problem reduces to finding the value of β which corresponds to the known, desired value $\widetilde{G}$, i.e., the value of β for which $G(\beta) = \widetilde{G}$.

The new variable β is known as a **Lagrange Multiplier**, named after the French mathematician Joseph-Louis Lagrange (1736–1813)[1]. Lagrange developed a general technique, using such variables, to perform constrained maximization, of which our current problem is a very simple case. We will not use the mathematical technique of Lagrange Multipliers—it is more powerful and more complicated than we need.

Here is what we will do instead. We will start with the answer, which others have derived using Lagrange Multipliers, and prove that it is correct. That is, we will give a formula for the probability distribution $p(A_i)$ in terms of the β and the $g(A_i)$ parameters, and then prove that the entropy calculated from this distribution, $S(\beta)$ is at least as large as the entropy of any probability distribution that has the same expected value for G, namely $G(\beta)$. Therefore the use of β automatically maximizes the entropy. Then we will show how to find the value of β, and therefore indirectly all the quantities of interest, for the particular value $\widetilde{G}$ of interest (this will be possible because $G(\beta)$ is a monotonic function of β so calculating its inverse can be done with zero-finding techniques).

9.6.2 Probability Formula

The probability distribution $p(A_i)$ we want has been derived by others. It is a function of the dual variable β:

$$p(A_i) = 2^{-\alpha} 2^{-\beta g(A_i)} \tag{9.12}$$

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

which implies

$$\log_2\left(\frac{1}{p(A_i)}\right) = \alpha + \beta g(A_i) \tag{9.13}$$

where α is a convenient abbreviation[2] for this function of β:

$$\alpha = \log_2\left(\sum_i 2^{-\beta g(A_i)}\right) \tag{9.14}$$

Note that this formula for α guarantees that the $p(A_i)$ from Equation 9.12 add up to 1 as required by Equation 9.8.

If β is known, the function α and the probabilities $p(A_i)$ can be found and, if desired, the entropy S and the constraint variable G. In fact, if S is needed, it can be calculated directly, without evaluating the $p(A_i)$—this is helpful if there are dozens or more probabilities to deal with. This short-cut is found by multiplying Equation 9.13 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

$$S = \alpha + \beta G \tag{9.15}$$

where S, α, and G are all functions of β.

9.6.3 The Maximum Entropy

It is easy to show that the entropy calculated from this probability distribution is at least as large as that for any probability distribution which leads to the same expected value of G.

Recall the Gibbs inequality, Equation 6.4, which will be rewritten here with $p(A_i)$ and $p'(A_i)$ interchanged (it is valid either way):

$$\sum_i p'(A_i)\log_2\left(\frac{1}{p'(A_i)}\right) \le \sum_i p'(A_i)\log_2\left(\frac{1}{p(A_i)}\right) \tag{9.16}$$

where $p'(A_i)$ is any probability distribution and $p(A_i)$ is any other probability distribution. The inequality is an equality if and only if the two probability distributions are the same.

The Gibbs inequality can be used to prove that the probability distribution of Equation 9.12 has the maximum entropy. Suppose there is another probability distribution $p'(A_i)$ that leads to an expected value G' and an entropy S', i.e.,

$$1 = \sum_i p'(A_i) \tag{9.17}$$

$$G' = \sum_i p'(A_i) g(A_i) \tag{9.18}$$

$$S' = \sum_i p'(A_i)\log_2\left(\frac{1}{p'(A_i)}\right) \tag{9.19}$$

Then it is easy to show that, for any value of β, if $G' = G(\beta)$ then $S' \le S(\beta)$:

[2]The function $\alpha(\beta)$ is related to the **partition function** $Z(\beta)$ of statistical physics: $Z = 2^\alpha$ or $\alpha = \log_2 Z$.

$$\begin{aligned} S' &= \sum_i p'(A_i) \log_2 \left(\frac{1}{p'(A_i)} \right) \\ &\leq \sum_i p'(A_i) \log_2 \left(\frac{1}{p(A_i)} \right) \\ &= \sum_i p'(A_i)[\alpha + \beta g(A_i)] \\ &= \alpha + \beta G' \\ &= S(\beta) + \beta[G' - G(\beta)] \end{aligned} \tag{9.20}$$

where Equations 9.16, 9.13, 9.17, 9.18, and 9.15 were used. Thus the entropy associated with any alternative proposed probability distribution that leads to the same value for the constraint variable cannot exceed the entropy for the distribution that uses β.

9.6.4 Evaluating the Dual Variable

So far we are considering the dual variable β to be an independent variable. If we start with a known value $\widetilde{G}$, we want to use G as an independent variable and calculate β in terms of it. In other words, we need to invert the function $G(\beta)$, or find β such that Equation 9.9 is satisfied.

This task is not trivial; 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. 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, start with Equation 9.12 for $p(A_i)$, multiply it by $g(A_i)$ and by 2^{α}, and sum over the probabilities. The left hand side becomes $G(\beta)2^{\alpha}$, because neither α nor $G(\beta)$ depend on i. We already have an expression for α in terms of β (Equation 9.14), so the left hand side becomes $\sum_i G(\beta)2^{-\beta g(A_i)}$. The right hand side becomes $\sum_i g(A_i)2^{-\beta g(A_i)}$. Thus,

$$0 = \sum_i [g(A_i) - G(\beta)]2^{-\beta g(A_i)} \tag{9.21}$$

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

$$0 = f(\beta) \tag{9.22}$$

where the function $f(\beta)$ is

$$f(\beta) = \sum_i [g(A_i) - G(\beta)]2^{-\beta[g(A_i) - G(\beta)]} \tag{9.23}$$

Equation 9.22 is the fundamental equation that is to be solved for particular values of $G(\beta)$, for example $\widetilde{G}$. The function $f(\beta)$ depends on the model of the problem (i.e., the various $g(A_i)$), and on $\widetilde{G}$, and that is all. It does not depend explicitly on α or the probabilities $p(A_i)$.

How do we know that there is any value of β for which $f(\beta) = 0$? First, notice that since $\widetilde{G}$ lies between the smallest and the largest $g(A_i)$, there is at least one i for which $(g(A_i) - \widetilde{G})$ 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.)

9.6.5 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 $p(T)$. Here is what you know:

$$1 = p(B) + p(C) + p(F) + p(T) \tag{9.24}$$

$$0 = \$1.00p(B) + \$2.00p(C) + \$3.00p(F) + \$8.00p(T) - \$2.50 \tag{9.25}$$

$$S = p(B)\log_2\left(\frac{1}{p(B)}\right) + p(C)\log_2\left(\frac{1}{p(C)}\right) + p(F)\log_2\left(\frac{1}{p(F)}\right) + p(T)\log_2\left(\frac{1}{p(T)}\right) \tag{9.26}$$

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

$$p(B) = 2^{-\alpha}2^{-\beta\$1.00} \tag{9.27}$$
$$p(C) = 2^{-\alpha}2^{-\beta\$2.00} \tag{9.28}$$
$$p(F) = 2^{-\alpha}2^{-\beta\$3.00} \tag{9.29}$$
$$p(T) = 2^{-\alpha}2^{-\beta\$8.00} \tag{9.30}$$

where

$$\alpha = \log_2(2^{-\beta\$1.00} + 2^{-\beta\$2.00} + 2^{-\beta\$3.00} + 2^{-\beta\$8.00}) \tag{9.31}$$

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

$$f(\beta) = \$0.50 \times 2^{-\$0.50\beta} + \$5.50 \times 2^{-\$5.50\beta} - \$1.50 \times 2^{\$1.50\beta} - \$0.50 \times 2^{\$0.50\beta} \tag{9.32}$$

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 unspecified value. The results all depend on H as well as E.

$$1 = p(U) + p(D) \tag{9.33}$$

$$\begin{aligned} \widetilde{E} &= e(U)p(U) + e(D)p(D) \\ &= m_d H[p(U) - p(D)] \end{aligned} \tag{9.34}$$

$$S = p(U)\log_2\left(\frac{1}{p(A)}\right) + p(D)\log_2\left(\frac{1}{p(D)}\right) \tag{9.35}$$

The entropy is the largest, for the energy $\widetilde{E}$ and magnetic field H, if

$$p(U) = 2^{-\alpha}2^{-\beta m_d H} \tag{9.36}$$
$$p(D) = 2^{-\alpha}2^{\beta m_d H} \tag{9.37}$$

where

$$\alpha = \log_2(2^{-\beta m_d H} + 2^{\beta m_d H}) \tag{9.38}$$

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

$$f(\beta) = (m_d H - \widetilde{E})2^{-\beta(m_d H - \widetilde{E})} - (m_d H + \widetilde{E})2^{\beta(m_d H + \widetilde{E})} \tag{9.39}$$

Note that this example with only one dipole, and therefore only two states, does not actually require the Principle of Maximum Entropy because there are two equations in two unknowns, $p(U)$ and $p(D)$ (you can solve Equation 9.39 for β using algebra). If there were two dipoles, there would be four states and algebra would not have been sufficient. If there were many more than four possible states, this procedure to calculate β would have been impractical or at least very difficult. We therefore ask, in Chapter 11 of these notes, what we can tell about the various quantities even if we cannot actually calculate numerical values for them using the summation over states.

6.050J / 2.110J Information and Entropy
Spring 2008