Chapter 9

Principle of Maximum Entropy: Simple Form

In the last chapter, we discussed one technique of estimating the input probabilities of a process given that the output event is known. This technique, which relies on the use of Bayes' Theorem, only works if the process is lossless (in which case the input can be identified with certainty) or the a priori input probability distribution is assumed (in which case the technique refines the initial probability distribution to take account of the known output).

The Principle of Maximum Entropy is a technique that can be used to estimate input probabilities more generally. The result is a probability distribution 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 principle is described first for the simple case of one constraint and three input events, in which case the technique can be carried out analytically. Then it is described more generally in Chapter 10.

This principle has applications in many domains, but was originally motivated by statistical physics, which attempts to relate macroscopic, measurable properties of physical systems to a description at the atomic or molecular level. It can be used to approach physical systems from the point of view of information theory, because the probability distributions can be derived by avoiding the assumption that the observer has more information than is actually available. Information theory, particularly the definition of information in terms of probability distributions, provides a quantitative measure of ignorance (or uncertainty, or entropy) that can be maximized mathematically to find the probability distribution that is maximally unbiased.

This approach to statistical physics was pioneered by Edwin T. Jaynes (1922 - 1998), a professor at Washington University in St. Louis, and previously Stanford University. The seminal publication was

- E. T. Jaynes, "Information Theory and Statistical Mechanics," Physical Review, vol. 106, no. 4, pp. 620-630; May 15, 1957.
  (http://bayes.wustl.edu/etj/articles/theory.1.pdf)

Other references of interest by Jaynes include:

- a continuation of this paper, E. T. Jaynes, "Information Theory and Statistical Mechanics. II," Physical Review, vol. 108, no. 2, pp. 171-190; October 15, 1957.
  (http://bayes.wustl.edu/etj/articles/theory.1.pdf)
- a review paper, including an example of estimating probabilities of an unfair die, E. T. Jaynes, "Information Theory and Statistical Mechanics," pp. 181-218 in "Statistical Physics," Brandeis Summer

Author: Paul Penfield, Jr.

Institute 1962, W. A. Benjamin, Inc., New York, NY; 1963.
(http://bayes.wustl.edu/etj/articles/brandeis.pdf)

- personal history of the approach, Edwin T. Jaynes, "Where Do We Stand on Maximum Entropy?," pp. 15-118, in "The Maximum Entropy Formalism," Raphael D. Levine and Myron Tribus, editors, The MIT Press, Cambridge, MA; 1979.
(http://bayes.wustl.edu/etj/articles/stand.on.entropy.pdf)

The philosophy of assuming maximum uncertainty as an approach to thermodynamics is discussed in

- Chapter 3 of M. Tribus, "Thermostatics and Thermodynamics," D. Van Nostrand Co, Inc., Princeton, NJ; 1961.

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 states is assumed known. Often quantum mechanics is needed for this task. It is not assumed in this step which particular state is actually occupied; indeed it is assumed that we do not know and cannot know this with certainty, and so we deal instead with the probability of each of the states being occupied. Thus we use probability as a means of coping with our lack of complete knowledge. Naturally we want to avoid inadvertently assuming more knowledge than we actually have, and the Principle of Maximum Entropy is the technique for doing this. In the application to nonphysical systems, again the various possible events have to be enumerated and the properties of each type associated with each of the possibilities known. In these notes we will derive a simple form of the Principle of Maximum Entropy and apply it to the restaurant example set up in the last chapter.

9.1.1 Berger's Burgers

This example was described in Chapter 8. A fast-food restaurant offers three meals: burger, chicken, and fish. The price, Calorie count, and probability of each meal being delivered cold are as as listed in Table 9.1

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

Table 9.1: Berger's Burgers

9.2 Probabilities

Even though now the problem has been set up, we do not know which state the system is actually 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. In the case of the restaurant model we have three such probabilities, which for simplicity we denote $p(B)$, $p(C)$, and $p(F)$ for the three meals.

A probability distribution $p(A_i)$ has the property that each of the probabilities is between or equal to 0 and 1, and, since the input events are mutually exclusive and exhaustive, the sum of all the probabilities is 1:

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

If any of the probabilities is equal to 1 then all the other probabilities are 0, and we then know exactly which state the system is in; in other words, we have no uncertainty and there is no need to resort to probabilities.

Since probabilities are used to cope with our lack of knowledge, and since one person may have more knowledge than another, it follows that two observers may, because of their different knowledge, use different probability distributions. In this sense probability, and all quantities that are based on probabilities, are subjective.

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 because we are using logarithms to base 2.

In the context of physical systems this uncertainty is known as the entropy. In communication systems the uncertainty regarding which actual message is to be transmitted is also known as the entropy of the source. 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. The Principle of Maximum Entropy is used to discover the probability distribution which leads to the highest value for this uncertainty, thereby assuring that no information is inadvertently assumed.

If one of the probabilities is equal to 1 then all the other probabilities are 0 and the entropy evaluates to 0 bits.

9.4 Constraints

It is a property of the entropy formula above that it has its maximum value when all probabilities are equal (we assume the number of possible states is finite). 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, namely 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

$$G = \sum_i p(A_i) g(A_i) \tag{9.3}$$

Note that this constraint cannot be achieved if G is less than the smallest $g(A_i)$ or larger than the largest $g(A_i)$.

For our Berger's Burgers example, suppose we are told that the average price of a meal is \$1.75, and we want to estimate the separate probabilities of the various meals without making any other assumptions. Then our constraint would be

$$\$1.75 = \$1.00p(B) + \$2.00p(C) + \$3.00p(F) \tag{9.4}$$

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.

9.5 Maximum Entropy, Analytic Form

Here we demonstrate the Principle of Maximum Entropy for a very simple case, one in which there is one constraint and three variables. It will be possible to go through all the steps analytically.

Suppose you have been hired by Carnivore Corporation, the parent company of Berger's Burgers, to analyze their worldwide sales. You visit Berger's Burgers restaurants all over the world, and determine that, on average, people are paying $1.75 for their meals. As part of Carnivore's commitment to global homogeneity, the price of each meal is exactly the same in every restaurant (after local currencies are converted to U.S. dollars). The prices are $1 for the burger meal, $2 for the chicken meal, and $3 for the fish meal.

After you return, your supervisors ask about the probabilities of a customer ordering each of the three value meals. You are horrified to realize that you forgot to gather that information, and there is no time to repeat your trip. You have to make the best estimate of the probabilities $p(B)$, $p(C)$, and $p(F)$ consistent with the two things you do know:

$$1 = p(B) + p(C) + p(F) \tag{9.5}$$

$$\$1.75 = \$1.00p(B) + \$2.00p(C) + \$3.00p(F) \tag{9.6}$$

Since you have three unknowns and only two equations, there is not enough information to solve for the unknowns. The amount of your uncertainty about the probability distribution is the entropy

$$S = p(B)\log_2\left(\frac{1}{p(B)}\right) + p(C)\log_2\left(\frac{1}{p(C)}\right) + p(F)\log_2\left(\frac{1}{p(F)}\right) \tag{9.7}$$

What should your strategy be? There are a range of values of the probabilities that are consistent with what you know. However, these leave you with different amounts of uncertainty S. If you choose one for which S is small, you are assuming something you do not know. For example, if your average had been $2.00 rather than $1.75, you could have met both of your constraints by assuming that everybody bought the chicken meal. Then your uncertainty would have been 0 bits. Or you could have assumed that half the orders were for burgers and half for fish, and the uncertainty would have been 1 bit. Neither of these assumptions seems particularly appropriate, because each goes beyond what you know. How can you find that probability distribution that uses no further assumptions beyond what you already know?

The **Principle of Maximum Entropy** states the rather obvious point that you should select that probability 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.

For the simple case of three probabilities and two constraints, this is easy to do analytically. Working with the two constraints, two of the unknown probabilities can be expressed in terms of the third. For our case we can multiply the first equation above by $1.00 and subtract it from the second, to eliminate $p(B)$. Then we can multiply the first by $2.00 and subtract it from the second, thereby eliminating $p(C)$:

$$p(C) = 0.75 - 2p(F) \tag{9.8}$$

$$p(B) = 0.25 + p(F) \tag{9.9}$$

Next, the possible range of values of the probabilities can be determined. Since each of the three lies between 0 and 1, it is easy to conclude from these results that

$$0 \leq p(F) \leq 0.375 \tag{9.10}$$
$$0 \leq p(C) \leq 0.75 \tag{9.11}$$
$$0.25 \leq p(B) \leq 0.625 \tag{9.12}$$

Next, these expressions can be substituted into the formula for entropy so that it is expressed in terms of a single probability. Thus

$$S = (0.25+p(F)) \log_2 \left(\frac{1}{(0.25 + p(F))} \right) + (0.75-2p(F)) \log_2 \left(\frac{1}{(0.75 - 2p(F))} \right) + p(F) \log_2 \left(\frac{1}{p(F)} \right) \tag{9.13}$$

Any of several techniques can now be used to find the value of $p(F)$ for which S is the largest. In this case the maximum occurs for $p(F) = 0.216$ and hence $p(B) = 0.466$, $p(C) = 0.318$, and $S = 1.517$ bits.

After estimating the input probability distribution, any averages over that distribution can be estimated. For example, in this case the average Calorie count can be calculated, or the expected number of meals served cold.

Let's remind ourselves what we have done. We have expressed our constraints in terms of the unknown probability distributions. One of these constraints is that the sum of the probabilities is 1. The other involves the average value of some quantity, in this case cost (or it could have been the average Calorie count). We used these constraints to eliminate two of the variables. We then expressed the entropy in terms of the remaining variable. Finally, we found the value of the remaining variable for which the entropy is the largest. The result is a probability distribution that is consistent with the constraints but which has the largest possible uncertainty. Thus we have not inadvertently introduced any biases into the probability estimation.

This technique requires that the model for the system be known at the outset; the only thing not known is the probability distribution. As carried out in this section, with a small number of unknowns and one more unknown than constraint, the derivation can be done analytically. For more complex situations a more general approach using Lagrange Multipliers is necessary. That is the topic of Chapter 10.