Chapter 8

Inference

In Chapter 7 the process model was introduced as a way to account for ow of information through processes that are discrete, nite, and memoryless, but which may be nondeterministic and lossy. Although motivated by the way many communication systems work, the model is more general.

Formulas were given for input information I, loss L, mutual information M, noise N, and output information J. Each of these is measured in bits, although in a setting in which many symbols are chosen, one after another, they may be multiplied by the rate of symbol selection and then expressed in bits per second. The information ow is shown in Figure 8.1. All these quantities depend on the input probability distribution $p(A_i)$.

If the input probabilities are known and a particular output outcome is observed, is it possible to infer which input event led to that outcome? Sometimes the input event can be identi ed with certainty, but more often the inferences must be based on calculated changes in the input probabilities. This is typically how communication systems work|the output is observed and the \most probable" input event is inferred. Inference in this context requires probability updating . This is the topic of Section 8.1.

However, if the input probabilities are not known, this approach does not work. We need a way to get the initial probability distribution. Then we can apply the techniques of Section 8.1.

A widely used technique for deriving initial probabilities is the Principle of Maximum Entropy, introduced in Section 8.2 and used later in these notes. This principle is based on the sensible notion that the initial probability distribution should be as unbiased as possible.

8.1 Probability Updating

It is often necessary to determine the input event when only the output event has been observed. This is the case for communication systems, in which the objective is to infer the symbol emitted by the source so that it can be reproduced at the output. It is also the case for memory systems, in which the objective is to recreate the original bit pattern without error.

In general, this inference requires that changes in the input probability distribution caused by our knowledge of the output event be calculated This is straightforward if the input probability distribution $p(A_i)$ and the conditional output probabilities, conditioned on the input events, $p(B_j \mid A_i) = c_{ji}$, are known. These \forward" conditional probabilities c_{ji} form a matrix with as many rows as there are output events, and as many columns as there are input events. They are a property of the process, and do not depend on the input probabilities $p(A_i)$.

The unconditional probability $p(B_j)$ of each output event B_j is

Author: Paul Pen eld, Jr.
This document: http://www.mtl.mit.edu/Courses/6.050/2013/notes/chapter8.pdf
Version 1.7, March 26, 2013. Copyright c 2013 Massachusetts Institute of Technology

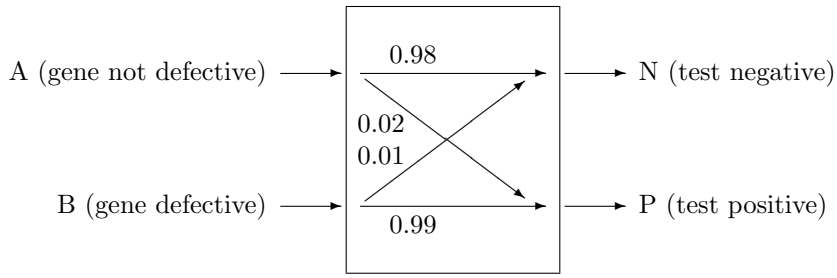

Figure 8.3: Huntington's Disease Test

model is whether people with a family history would take the test, or whether they would prefer to live not knowing what the future holds in store. The development of the test was funded by a group including Guthrie's widow and headed by a man named Milton Wexler (1908–2007), who was concerned about his daughters because his wife and her brothers all had the disease. Wexler's daughters, whose situation inspired the development of the test, decided not to take it.

Next, consider the application of this test to someone with an unknown family history, so that $p(A) = 0.9995$ and $p(B) = 0.0005$. Then, if the test is negative, the probability of that person carrying the defective gene $p(B \mid N)$ is

$$\frac{0.0005 \times 0.01}{0.0005 \times 0.01 + 0.9995 \times 0.98} = 0.000005105 \tag{8.10}$$

and the probability of that person carrying the normal gene $p(A \mid N)$ is

$$\frac{0.9995 \times 0.98)}{0.0005 \times 0.01 + 0.9995 \times 0.98} = 0.999994895 \tag{8.11}$$

On the other hand, if the test is positive, the probability the person has the defective gene $p(B \mid P)$ is

$$\frac{0.0005 \times 0.99}{0.0005 \times 0.99 + 0.9995 \times 0.02} = 0.02416 \tag{8.12}$$

and the probability of not having the defect $p(A \mid P)$ is

$$\frac{0.9995 \times 0.02}{0.0005 \times 0.99 + 0.9995 \times 0.02} = 0.97584 \tag{8.13}$$

The test does not distinguish the two possible inputs, since the overwhelming probability is that the person has a normal gene, regardless of the test results. In other words, a positive test result is more likely to have been caused by a testing error than a defective gene. There seems to be no point in testing people without a family history. (Of course in practice repeated tests could be done to reduce the false positive rate.)

An information analysis demonstrates the difference between these two cases. First, recall that probabilities are subjective, or observer-dependent. The lab technician performing the test presumably does not know whether there is a family history, and so would not be able to infer anything from the results. Only someone who knows the family history could make a useful inference. Second, it is instructive to calculate the information flow in the two cases. Recall that all five information measures (I, L, M, N, and J) depend on the input probabilities. A straightforward calculation of the two cases leads to the information quantities (in bits) in Table 8.1. Note how small the mutual information M is compared to the noise N if there is no known family history.

Clearly the test conveys information about the patients' status by reducing the uncertainty in the case where there is a family history of the disease. On the other hand, without a family history there is very little information that could possibly be conveyed because there is so little initial uncertainty.

	$p(A)$	$p(B)$	I	L	M	N	J
Family history	0.5	0.5	1.00000	0.11119	0.88881	0.11112	0.99993
Unknown Family history	0.9995	0.0005	0.00620	0.00346	0.00274	0.14141	0.14416

Table 8.1: Model for Huntington's Disease Test Process

8.1.3 Example: Berger's Burgers

A former 6.050J/2.110J student opened a fast-food restaurant, and named it in honor of the very excellent Undergraduate Assistant of the course. At Berger's Burgers, meals are prepared with state-of-the-art high-tech equipment using reversible computation for control. To help reduce global climate change there are no warming tables, but instead the entropy created by discarding information is used to keep the food warm. Because the rate at which information is discarded in a computation is unpredictable, the food does not always stay warm. There is a certain probability, different for the different menu items, of a meal being "COD" (cold on delivery).

The three original menu items are Burger Meals 1, 2, and 3. Burger Meal 1 (beef) costs \$1, delivers 1000 Calories, and has a probability 0.5 of arriving cold. Burger Meal 2 (chicken) costs \$2, has 600 Calories, and a probability 0.2 of arriving cold. Burger Meal 3 (fish) costs \$3, has 400 Calories, and has a 0.1 probability of being cold.

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

Table 8.2: Berger's Burgers

There are several inference questions that can be asked about Berger's Burgers. All require an initial assumption about the buying habits of the public, i.e., about the probability of each of the three meals being ordered $p(B)$, $p(C)$, and $p(F)$. Then, upon learning another fact, such as a particular customer's meal arriving cold, these probabilities can be updated to be consistent with the new information.

Suppose you arrive at Berger's Burgers with your friends and place your orders. Assume that money is plentiful so you and your friends are equally likely to order any of the three meals. Also assume that you do not happen to hear what your friends order or see how much they pay. Also assume that you do not know your friends' taste preferences and that the meals come in identical packages so you cannot tell what anyone else received by looking.

Before the meals are delivered, you have no knowledge of what your friends ordered and might assume equal probability of 1/3 for $p(B)$, $p(C)$, and $p(F)$. You can then calculate the average amount paid per meal (\$2.00), the average Calorie count (667 Calories), and the probability that any given order would be COD (0.267).

Now suppose your friend Alice remarks that her meal is cold. Knowing this, you can calculate the probability she ordered beef (0.625) or chicken (0.25) or fish (0.125). And what is the expected value of the amount she paid for her meal? \$1.50. And her expected Calorie count? 825 Calories.

Next suppose your friend Bob says he feels sorry for her and offers her some of his meal, which is hot. Straightforward application of the formulas above can determine the refined probabilities of what he ordered, along with the expected calorie count and cost.

8.1.4 Inference Strategy

Often, it is not sufficient to calculate the probabilities of the various possible input events. The correct operation of a system may require that a definite choice be made of exactly one input event. For processes

without loss, this can be done accurately. However, for processes with loss, some strategy must be used to convert probabilities to a single choice.

One simple strategy, "maximum likelihood," is to decide on whichever input event has the highest probability after the output event is known. For many applications, particularly communication with small error, this is a good strategy. It works for the symmetric binary channel when the two input probabilities are equal and the error is small. However, sometimes it does not work at all. For example, if used for the Huntington's Disease test on people without a family history, this strategy would never say that the person has a defective gene, regardless of the test results.

Inference is important in many fields of interest, such as machine learning, natural language processing and other areas of artificial intelligence. An open question, of current research interest, is which inference strategies are best suited for particular purposes.

8.2 Principle of Maximum Entropy: Simple Form

In the last section, 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 an initial input probability distribution is known (in which case it is refined 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 (the word "bias" is used here not in the technical sense of statistics, but the everyday sense of a preference that inhibits impartial judgment). 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 9.

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 best avoids unnecessary assumptions.

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 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.

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 the system is in (or, as often expressed, which state is actually "occupied"); indeed it is assumed that we do not know and cannot know this with certainty. Instead we deal with the probability of each of the states being occupied. In other words, we use probability to cope 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, the various events (possible outcomes) have to be identified along with various numerical properties associated with each of the events. 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 Section 8.1.3.

8.2.1 Berger's Burgers

The Principle of Maximum Entropy will be introduced by means of an example. This example was described in Section 8.1.3. A fast-food restaurant, Berger's Burgers, offers three meals: beef, chicken, and fish. The price, Calorie count, and probability of each meal being delivered cold are as as listed in Table 8.2.

8.2.2 Probabilities

This example has been defined so that the choice of one of the three meals constitutes an outcome. If we do not know this outcome we may still have some knowledge, and we use probabilities to express this knowledge. The question is how to assign probabilities that are consistent with whatever information we may have.

In the case of Berger's Burgers, there are three 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:

$$\begin{aligned} 1 &= \sum_i p(A_i) \\ &= p(B) + p(C) + p(F) \end{aligned} \tag{8.14}$$

8.2.3 Entropy

If any of the probabilities $p(B)$, $p(C)$, or $p(F)$ is equal to 1 then the other probabilities are all 0, and we know exactly which state the system is in; in other words, we have no uncertainty and there is no need to resort to probabilities. More generally our uncertainty is expressed quantitatively by the information which we do not have about the meal chosen, or the state occupied. This is

$$
\begin{aligned}
S &= \sum_i p(A_i)\log_2\left(\frac{1}{p(A_i)}\right) \\
&= p(B)\log_2\left(\frac{1}{p(B)}\right) + p(C)\log_2\left(\frac{1}{p(C)}\right) + p(F)\log_2\left(\frac{1}{p(F)}\right) \qquad (8.15)
\end{aligned}
$$

Here, 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, appropriately, calculate a different set of probabilities and therefore different numerical value for entropy. The Principle of Maximum Entropy can be used by each person, based on his or her specific knowledge, to find the probability distribution that avoids the inadvertent assumption of anything that person does not know. The resulting probability distributions are still observer-dependent.

8.2.4 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). This property is easily proved using the Gibbs inequality. 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 an attribute for which each of the states has a value $g(A_i)$ and for which we know the actual value G, then we should consider only those probability distributions for which the expected value is equal to G

$$G = \sum_i p(A_i)g(A_i) \qquad (8.16)$$

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

$$\$1.75 = \$1.00p(B) + \$2.00p(C) + \$3.00p(F) \qquad (8.17)$$

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

8.2.5 Maximum Entropy, Analytic Form

Here we demonstrate the Principle of Maximum Entropy for the simple case 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 revenues. 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.)

After you return, your supervisors ask about the probabilities of a customer ordering each of the three Burger Meals. In other words, they want to know $p(B)$, $p(C)$, and $p(F)$. You are horrified to realize that you did not keep the original data, 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:

$$1 = p(B) + p(C) + p(F) \tag{8.18}$$

$$\$1.75 = \$1.00p(B) + \$2.00p(C) + \$3.00p(F) \tag{8.19}$$

Since you have three unknowns and only two equations, there is not enough information to solve for the unknowns.

What should you do? 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

$$S = p(B)\log_2\left(\frac{1}{p(B)}\right) + p(C)\log_2\left(\frac{1}{p(C)}\right) + p(F)\log_2\left(\frac{1}{p(F)}\right) \tag{8.20}$$

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 beef 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** is based on the reasonable assumption 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 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 Equation 8.18 above by \$1.00 and subtract it from Equation 8.19, to eliminate $p(B)$. Then we can multiply the first by \$2.00 and subtract it from the second, thereby eliminating $p(C)$:

$$p(C) = 0.75 - 2p(F) \tag{8.21}$$

$$p(B) = 0.25 + p(F) \tag{8.22}$$

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

$$0 \le p(F) \le 0.375 \tag{8.23}$$

$$0 \le p(C) \le 0.75 \tag{8.24}$$

$$0.25 \le p(B) \le 0.625 \tag{8.25}$$

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

$$S = (0.25 + p(F))\log_2\left(\frac{1}{0.25 + p(F)}\right) + (0.75 - 2p(F))\log_2\left(\frac{1}{0.75 - 2p(F)}\right) + p(F)\log_2\left(\frac{1}{p(F)}\right) \tag{8.26}$$

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 found. For example, in this case the average Calorie count can be calculated (it is 743.2 Calories), or the probability of a meal being served cold (31.8%).

8.2.6 Summary

Let's review 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. 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 unwanted assumptions 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 is necessary. That is the topic of Chapter 9.