Chapter 8

Inference

The model of a communication system that we have been developing is shown in Figure 8.1, where the source is assumed to emit a stream of symbols. The channel may be a physical channel between different points in space, or it may be a memory which stores information for retrieval at a later time, or it may be a computation in which the information is processed in some way.

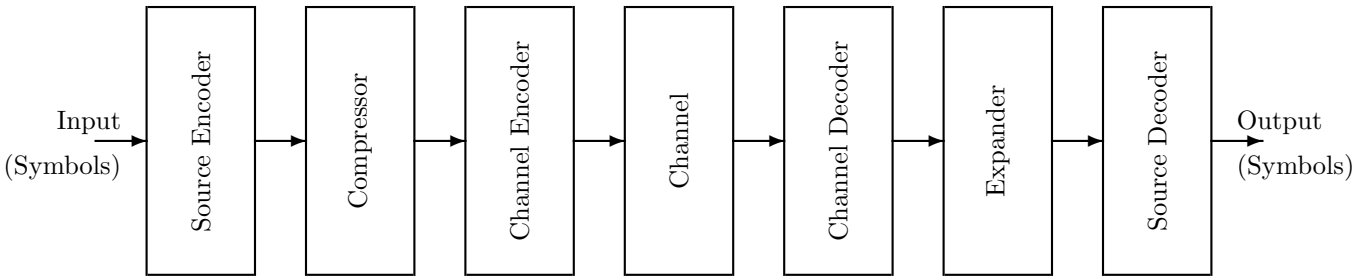


Figure 8.1: Elaborate communication system

Each of the boxes in this diagram can be represented by a "process" with a finite number of mutually exclusive, exhaustive input events, a finite number of mutually exclusive, exhaustive output events, and a means of calculating the output probability distribution from the input probability distribution. In Chapter 7, these processes were assumed to be

- **Discrete:** The inputs are selected from a discrete set of mutually exclusive possibilities; only one can occur at a time. The output is one of another discrete set of mutually exclusive values.
- **Finite:** The inputs are selected from a finite set of possibilities; only one can occur at a time. Similarly, the output is one of another finite set of values.
- **Memoryless:** The process acts on the input at some time and produces an output based on that input and not on any prior inputs.
- **Nondeterministic:** The process may produce different outputs when it is presented with the same input more than once (the model is also valid for deterministic processes).

Author: Paul Penfield, Jr.

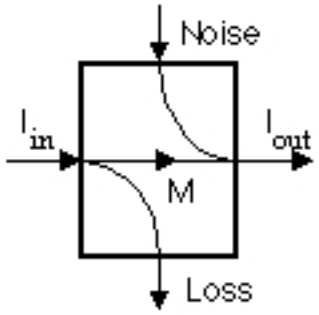


Figure 8.2: Information flow in a discrete memoryless channel

- **Lossy:** It may not be possible to determine the input state by observing the output state (the model is also valid for lossless processes).

A generalization of this model to processes with memory is sometimes required. In this case some internal state of the process would be set by the input, and the probability distribution leading to the output and the next state would depend on the current state. This sort of process-with-memory model is needed for some communications and computation systems, but will not be considered in these notes because of the greater complexity.

To refresh your own memory, included here, in Figure 8.2, is the diagram showing information flow in such processes from Chapter 6.

8.1 Estimation

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 eventually infer the symbol emitted by the source so that it can be created at the output. It is also the case for memory systems, in which the objective is to recreate the original bit pattern that was previously stored, without error.

It is not always possible to infer the input event of a process from knowledge of the output. If the system has no loss, then inference is possible, but this is generally not the case. The best that can generally be done is to refine the probabilities of the input events once the output event has been observed.

In principle, this estimation 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. Note that 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

$$p(B_j) = \sum_i c_{ji} p(A_i) \tag{8.1}$$

and the joint probability of each input with each output $p(A_i, B_j)$ and the backward conditional probabilities $p(A_i \mid B_j)$ can be found using Bayes' Theorem:

$$\begin{aligned} p(A_i, B_j) &= p(B_j)p(A_i \mid B_j) \\ &= p(A_i)p(B_j \mid A_i) \\ &= p(A_i)c_{ji} \end{aligned} \tag{8.2}$$

Now let us suppose that a particular output event B_j has been observed. The input event that "caused" this output can be estimated only to the extent of giving a probability distribution over the input events.

For each input event A_i the probability that it was the input is simply the backward conditional probability for the particular output event B_j

$$p(A_i \mid B_j) = \frac{p(A_i)}{p(B_j)} c_{ji} \tag{8.3}$$

If the process has no loss ($L = 0$) then for each j exactly one of the input events has nonzero probability, and therefore its probability is 1. In the more general case, with nonzero loss, estimation consists of refining a set of input probabilities so they are consistent with the known output. It is necessary to have a set of initial input probabilities, and naturally the refined probability distribution depends on this initial distribution.

Note that this approach only works if the input probability distribution is known. All it does is refine that distribution in the light of new knowledge, namely the observed output. If the input probability distribution is not known, then another technique is required. One such technique is the Principle of Maximum Entropy, described in Chapter 9.

It might be thought that the new input probability distribution would have less uncertainty than that of the original distribution, and this is usually, though not always, true. The uncertainty of a probability distribution is, of course, its entropy as defined earlier. The uncertainty before the output event is known is

$$U_{\text{before}} = \sum_i p(A_i) \log_2 \left(\frac{1}{p(A_i)} \right) \tag{8.4}$$

The uncertainty after some particular output event is known is

$$U_{\text{after}}(B_j) = \sum_i p(A_i \mid B_j) \log_2 \left(\frac{1}{p(A_i \mid B_j)} \right) \tag{8.5}$$

Although it is not always true that $U_{\text{after}}(B_j) \leq U_{\text{before}}$, it is not difficult to prove that the average over all output states of the residual uncertainty is less than the original uncertainty:

$$\sum_j p(B_j) U_{\text{after}}(B_j) \leq U_{\text{before}} \tag{8.6}$$

In words, this statement says that on average, our uncertainty about the input state is never increased by learning something about the output state. In other words, on average, this technique of inference helps us get a better estimate of the input state.

Once the input probability distribution has been refined, improved estimates of any properties of the input state can be obtained. Such estimates are averages, or expected values, over the probability distribution. Some examples of such estimates are given in the examples below.

8.2 Examples

Two of the following examples will be continued in subsequent chapters including the next chapter on the Principle of Maximum Entropy – the symmetric binary channel and Berger's Burgers.

8.2.1 Symmetric Binary Channel

The noiseless, lossless binary channel shown in Figure 8.3(a) is a process with two input values which may be called 0 and 1, two output values similarly named, and a transition matrix c_{ji} which guarantees that the output equals the input:

$$\begin{bmatrix} c_{00} & c_{01} \\ c_{10} & c_{11} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{8.7}$$

This channel has no loss and no noise, and the mutual information, input information, and output information are all identical.

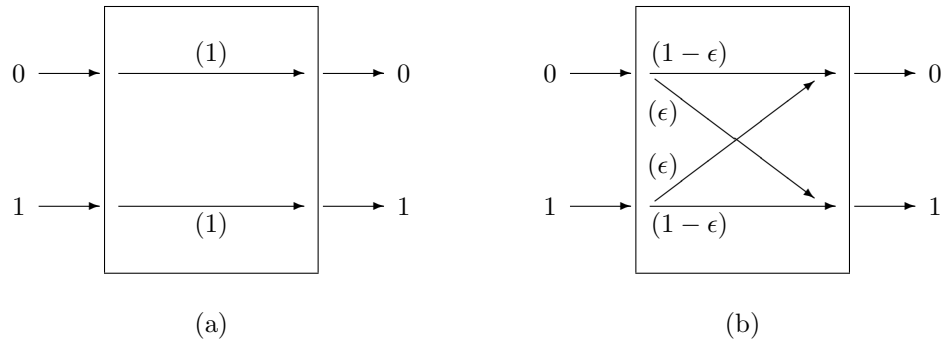


Figure 8.3: (a) Binary Channel without noise (b) Symmetric binary channel, with errors

The symmetric binary channel (Figure 8.3(b)) is similar, but occasionally makes errors. Thus if the input is 1 the output is not always 1, but with the "bit error probability" ϵ is flipped to the "wrong" value 0, and hence is "correct" only with probability $1-\epsilon$. Similarly, for the input of 0, the probability of error is ϵ. Then the transition matrix is

$$\begin{bmatrix} c_{00} & c_{01} \\ c_{10} & c_{11} \end{bmatrix} = \begin{bmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{bmatrix} \tag{8.8}$$

This channel is symmetric in the sense that the errors in both directions (from 0 to 1 and vice versa) are equally likely.

Because of the loss, the input event associated with, say, output event B_0 cannot be determined with certainty. Nevertheless, the formulas above can be used. In the important case where the two input probabilities are equal (and therefore each equal to 0.5) an output of 0 implies that the input event A_0 has probability $1-\epsilon$ and input event A_1 has probability ϵ. Thus if, as would be expected in a channel designed for low-error communication, ϵ is small, then it would be reasonable to infer that the input that produced this output was the event A_0.

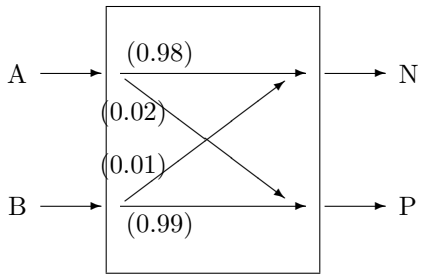
8.2.2 Huntington's Disease Testing

Huntington's Disease is a rare, progressive, hereditary brain disorder. It was named after Dr. George Huntington, a Long Island physician who published a description in 1872. People carrying the defective gene are said to be "at risk." Eventually they will develop the disease unless they die of another cause first. The symptoms appear in middle age, in people in their 40's or 50's, perhaps after the person has produced a family and has thereby possibly transmitted the defective gene to another generation. Although the disease is not fatal, those in advanced stages generally die from its complications. Until recently, people with a family history of the disease were faced with a life of uncertainty, not knowing whether they were at risk, and not knowing how to manage their personal and professional lives.

According to the Huntington's Disease Society of America, http://www.hdsa.org/, the incidence of the disease in the general public is about 1/10,000 in North America, meaning that the probability of carrying the defective gene is 0.01%, whereas the probability for those with family history (one of the two parents having the genetic disorder) is 50%.

In 1993 a test was developed which can tell if you carry the defective gene. Unfortunately the test is not perfect; there is a probability of false positive (reporting you are at risk when you actually are not) as well as a probability of a false negative (reporting you are not at risk when in fact you are). For simplicity, and because the accuracies of the tests are continually improving, we will assume that the probability of a false positive is 2%, and the probability of a false negative is 1%.

If you take the test and learn the outcome, you would of course like to infer whether you are at risk. The techniques developed above can help.



(b)

Figure 8.4: Huntinton's Disease Test

Let us model the test by a discrete memoryless process, with input A (no defective gene) and B (defective gene), and outputs P (positive) and N (negative). The process is shown in Figure 8.3(b).

First, consider the application of this test to someone with a family history, for which $p(A) = p(B) = 0.5$. Then, if the test is negative, the probability, for that person, of being at risk is $1/99 = 0.0101$ and the probability of not being at risk is $98/99 = 0.9899$. On the other hand, if the test is positive, the probability, for that person, of being at risk is $99/101 = 0.9802$ and the probability of not being at risk is $2/101 = 0.0198$. The test is very effective, in that the two outputs imply, to high probability, different inputs.

An interesting question raised by the test is whether a person with a family history would elect to take the test, or whether he or she would prefer to live not knowing what the future holds in store.

Next, consider the application of this test to someone with no family history, so that $p(A) = 0.9999$ and $p(B) = 0.0001$. Then, if the test is negative, the probability of that person being at risk is

$$\frac{(0.0001 \times 0.01)}{(0.0001 \times 0.01 + 0.9999 \times 0.98)} = 0.000001021 \tag{8.9}$$

and the probability of that person not being at risk is

$$\frac{(0.9999 \times 0.98)}{(0.0001 \times 0.01 + 0.9999 \times 0.98)} = 0.999998979 \tag{8.10}$$

On the other hand, if the test is positive, the probability of that person being at risk is

$$\frac{(0.0001 \times 0.99)}{(0.0001 \times 0.99 + 0.9999 \times 0.02)} = 0.004926 \tag{8.11}$$

and the probability of not being at risk is

$$\frac{(0.9999 \times 0.02)}{(0.0001 \times 0.99 + 0.9999 \times 0.02)} = 0.995074 \tag{8.12}$$

The test does not seem to distinguish the two possible inputs, since the overwhelming probability is that the person is not at risk, regardless of the test results. There seems to be no useful purpose served by testing people without a family history.

8.2.3 Berger's Burgers

A former 6.050/2.110 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 reduce the creation of entropy 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, 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 Value Meals 1, 2, and 3. Value Meal 1 (burger) costs \$1, contains 1000 Calories, and has a probability 0.5 of arriving cold. Value Meal 2 (chicken) costs \$2, has 600 Calories, and a probability 0.2 of arriving cold. Value 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
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 8.1: 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 that a particular customer's meal arrived cold, these probabilities can be refined to lead to a better estimate of the meal that was ordered.

Suppose you arrive at Berger's Burgers with your friends and place your orders. Assume that money is in plentiful supply 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 estimate 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, what is the probability she ordered a burger? (0.625) Chicken? (0.25) Fish? (0.125). And what is the expected value of the amount she paid for her meal? (\$1.50) And what is 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.

8.3 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. 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 always say that the person is healthy, regardless of the test results.

A discussion of other strategies is beyond the scope of these notes.