Chapter 3

Channels and Channel Capacity

3.1 Discrete memoryless channels

A *channel* is a communication device with two ends, an input end, or *transmitter*, and an output end, or *receiver*. A *discrete* channel accepts for transmission the characters of some finite alphabet $A = \{a_1, \dots, a_n\}$, the *input alphabet*, and delivers characters of an *output alphabet* $B = \{b_1, \dots, b_k\}$ to the receiver. Every time an input character is accepted for transmission, an output character subsequently arrives at the receiver. That is, we do not encompass situations in which the channel responds to an input character by delivering several output characters, or no output. Such situations may be defined out of existence: once the input alphabet and the channel are fixed, the output alphabet is defined to consist of all possible outputs that may result from an input. For instance, suppose $A = \{0, 1\}$, the *binary alphabet*, and suppose that it is known that the channel is rickety, and may *fuzz* the input digit so that the receiver cannot tell which digit, 0 or 1, is being received, or may *stutter* and deliver two digits, either of which might be fuzzed, upon the transmission of a single digit. Then, with $*$ standing for "fuzzy digit," we are forced to take $B = \{0, 1, *, 00, 01, 10, 11, 0*, 1*, *0, *1, **\}$.

For a finite alphabet A, we let, as convention dictates,

$$\begin{aligned} A^\ell &= \text{ the Cartesian product of } A \text{ with itself } \ell \text{ times} \\ &= \text{ the set of words of length } \ell, \text{ over } A. \end{aligned}$$

Further, let

$$A^+ = \bigcup_{\ell=1}^{\infty} A^\ell = \text{ the set of all (non-empty) words over } A.$$

Note that if A is the input alphabet of a channel, then any finite non-empty subset of A^+ could be taken as the input alphabet of the same channel. Changing the input alphabet in this way will necessitate a change in the output alphabet. For instance, if $A = \{0, 1\}$, and the corresponding output alphabet is $B = \{0, 1, *\}$, then if we take $\widehat{A} = \{00, 11\}$, the new output alphabet will be

$$\widehat{B} = \{00, 01, 0*, 10, 11, 1*, *0, *1, **\}.$$

It is possible to vary the output alphabet by merging or amalgamating letters; for instance, if $B = \{0, 1, *, x\}$, we could take $\widetilde{B} = \{0, 1, \alpha\}$, with α meaning "either $*$ or x." This might be a shrewd simplification if, for instance, the original letters $*$ and x are different sorts of error indicators, and the distinction is of no importance.

Another common method of simplifying the output alphabet involves modifying the channel by "adding a coin flip." For instance, if $B = \{0, 1, *\}$ and you really do not want to bother with $*$, you can flip a coin whenever $*$ is received to decide if it will be read as 0 or 1. The coin need not be fair. The same idea can be used to shrink B from any finite size down to any smaller size of 2 or more. The details of the process depend on the particular situation; they are left to the ingenuity of the engineer. See Exercise 3.2.6.

It may be that there are *fundamental* input and output alphabets forced upon us by the physical nature of the channel; or, as in the case of the telegraph, for which the time-hallowed input and output alphabet is {dot, dash} (or, as Shannon [65] has it, {dot, dash, short pause (between letters), long pause (between words)}), it may be that some fundamental input alphabet is strongly recommended, although not forced, by the physical nature of the channel. In the most widespread class of examples, the *binary* channels, the input alphabet has size 2, and we usually identify the input characters with 0 and 1. Note that, for any channel that accepts at least two input characters, we can always confine ourselves to two input characters, and thus make the channel binary.

The telegraph provides a historically fundamental example of a channel; it is a somewhat uninteresting, or misleading, example for the student of information theory, because it is so reliable. Over the telegraph, if a "dot" is transmitted, then a "dot" is received (unless the lines are down), and the same goes for "dash." What makes life interesting in modern times is "channel noise"; you cannot be dead certain what the output will be for a given input. Modern channels run from outer space, to the ocean floor, to downtown Cleveland—a lot can go wrong. Specks of dust momentarily lodge in the receiver, birds fly up in front of the transmitter, a storm briefly disrupts the local electromagnetic environment—it's a wonder that successful communication ever takes place.

We take account of the uncertainty of communication by regarding the attempt to transmit a single digit as a probabilistic experiment. Before we become thoroughly engaged in working out the consequences of this view, it is time to announce a blanket assumption that will be in force from here on in this text: our channels will all be *memoryless*. This means that the likelihood of b_j being the output when a_i is the input does not vary with local conditions, nor with recent history, for each i and j. These unvarying likelihoods are called transition probabilities and will be discussed in the next section.

Please note that this assumption may well be invalid in a real situation. For instance, when you hear a "skip" from a record on a turntable,[1] your estimate

[1] If you are unacquainted with "records" and "turntables," ask the nearest elderly person about them.

of the probability of a skip in the near future changes drastically. You now estimate that there is a great likelihood of another skip soon, because experience tells you that these skips occur for an underlying *reason*, usually a piece of fluff or lint caught on the phonograph needle. Just so, in a great many situations wobbles and glitches in the communication occur for some underlying reason that will not go away for a while, and the assumption of memorylessness is rendered invalid. What can you do in such situations? There is a good deal of theory and practice available on the subject of correcting *burst* errors, as they are called in some parts. This theory and practice will not, however, be part of this course. We are calling your attention to the phenomenon of burst errors, and to the indefensibility of our blanket assumption of memorylessness in certain situations, just because one of the worst things you can do with mathematics is to misapply it to situations outside the umbrella of assumption. Probabilistic assumptions about randomness and independence are very tricky, and the assumption of memorylessness of a channel is one such.

This is not the place to go into detail, but let us assure you that you can misapply a result about randomly occurring phenomena (such as the glitches, skips, and wobbles in transmissions over our memoryless channels are assumed to be) to "show," in a dignified, sincere manner, that the probability that the sun will *not* rise tomorrow is a little greater than $1/3$. The moral is that you should stare and ponder a bit, to see if your mathematics applies to the situation at hand, and if it doesn't, don't try to force it.

Exercises 3.1

1. (a) Suppose $A = \{0, 1\}$ and $B = \{0, 1, *\}$; suppose we decide to use $\widehat{A} = \{0000, 1111\}$ as the new input alphabet, for some reason. How large will the new output alphabet be?

   (b) In general, for any input alphabet A and output alphabet B, with $|B| = k$, if we take a new input alphabet $\widehat{A} \subset A^{\ell}$, how many elements will the new output alphabet have? What will the new output alphabet be?

2. A certain binary channel has the binary alphabet as its output alphabet, as well: $A = B = \{0, 1\}$. This channel has a memory, albeit a very short one. At the start of a transmission, or right after the successful transmission of a digit, the probability of a correct transmission is p (regardless of which digit, 0 or 1, is being transmitted); right after an error (0 input, 1 output, or 1 input, 0 output), the probability of a correct transmission is q. (If this situation were real, we would plausibly have $1/2 < q < p < 1$.) In terms of p and q, find

   (a) the probability that the string 10001 is received, if 11101 was sent;

   (b) the probability that 10111 was received, if 11101 was sent;

   (c) the probability of exactly two errors in transmitting a binary word of length 5;

(d) the probability of two or fewer errors, in transmitting a binary word of length n.

3. Another binary channel has $A = B = \{0, 1\}$, and no memory; the probability of a correct transmission is p, for each digit transmitted. Find the probabilities in problem 2, above, for this channel.

3.2 Transition probabilities and binary symmetric channels

Now we shall begin to work out the consequences of the assumption of memorylessness. Let the input alphabet be $A = \{a_1, \ldots, a_n\}$, and the output alphabet be $B = \{b_1, \ldots, b_k\}$. By the assumption of memorylessness, the probability that b_j will be received, *if* a_i is the input character, depends only on i, j, and the nature of the channel, not on the weather nor the recent history of the channel. We denote this probability by q_{ij}.

The q_{ij} are called the *transition probabilities* of the channel, and the $n \times k$ matrix $Q = [q_{ij}]$ is the matrix of transition probabilities. After the input and output alphabets have been agreed upon, Q depends on the hardware, the channel itself; or, we could say that Q is a property of the channel. In principle, q_{ij} could be estimated by testing the channel: send a_i many times and record how often b_j is received. In practice, such testing may be difficult or impossible, and the q_{ij} are either estimated through theoretical considerations, or remain hypothetical. Note that $\sum_{j=1}^{k} q_{ij} = 1$, for each i; that is, the row sums of Q are all 1.

A *binary symmetric channel* (BSC, for short) is a memoryless channel with $A = B = \{0, 1\}$ like that described in Exercise 3.1.3; whichever digit, 0 or 1, is being transmitted, the probability p that it will get through correctly is called the *reliability* of the channel. Usually, $1/2 < p < 1$, and we hope p is close to 1. Letting 0 and 1 index the transition probabilities in the obvious way, the matrix of transition probabilities for a binary symmetric channel with reliability p is

$$Q = \begin{bmatrix} q_{00} & q_{01} \\ q_{10} & q_{11} \end{bmatrix} = \begin{bmatrix} p & 1-p \\ 1-p & p \end{bmatrix}.$$

The word "symmetric" in "binary symmetric channel" refers to the symmetry of Q, or to the fact that the channel treats the digits 0 and 1 symmetrically.

Observe that sending any particular binary word of length n through a binary symmetric channel with reliability p is an instance of n independent Bernoulli trials, with probability p of Success on each trial (if you count a correct transmission of a digit as a Success). Thus, the probability of exactly k errors ($n-k$ Successes) in such a transmission is $\binom{n}{k} p^{n-k}(1-p)^k$, and the average or expected number of errors is $n(1-p)$.

Exercises 3.2

1. For a particular memoryless channel we have $A = \{0, 1\}$, $B = \{0, 1, *\}$, and the channel treats the input digits symmetrically; each digit has probability p of being transmitted correctly, probability q of being switched to the other digit, and probability r of being fuzzed, so that the output is $*$. Note that $p+q+r = 1$.

   (a) Give the matrix of transition probabilities, in terms of p, q, and r.

   (b) In terms of n, p, and k, what is the probability of exactly k errors (where an *error* is either a fuzzed digit or a switched digit) in the transmission of a binary word of length n, over this channel?

   (c) Suppose that $*$ is eliminated from the output alphabet by means of coin flip, with a fair coin. Whenever $*$ is received, the coin is flipped; if heads comes up, the $*$ is read is 0, and if tails comes up, it is read as 1. What is the new matrix of transition probabilities? Is the channel now binary symmetric?

   (d) Suppose that $*$ is eliminated from the output alphabet by merging it with 1. That is, whenever $*$ is received, it is read as 1 (this amounts to a coin flip with a very unfair coin). What is the new matrix of transition probabilities? Is the channel now binary symmetric?

2. A binary symmetric channel has reliability p.

   (a) What is the minimum value of p allowable, if there is to be at least a 95% chance of no errors at all in the transmission of a binary word of length 15?

   (b) Give the inequality that p must satisfy if there is to be at least a 95% chance of no more than one error in the transmission of a binary word of length 15. For the numerically deft and/or curious: is the minimum value of p satisfying this requirement significantly less than the minimum p satisfying the more stringent requirement in part (a)?

   (c) What is the minimum value of p allowable if the average number of errors in transmitting binary words of length 15 is to be no greater than $1/2$?

3. $A = B = \{0, 1\}$, and the channel is memoryless, but is not a binary symmetric channel because it treats 0 and 1 differently. The probability is p_0 that 0 will be transmitted correctly, and p_1 that 1 will be transmitted correctly.

   In terms of p_0 and p_1, find the probabilities described in Exercise 3.1.2 (a) and (b). Also, if a binary word of length n has z zeros and $n-z$ ones, with $n \geq 2$, find, in terms of p_0, p_1, n, and z, the probability of two or fewer errors in the transmission of the word.

4. Suppose we decide to take A^2 as the new input alphabet. Then B^2 will be the new output alphabet. How will the new transition probabilities $q_{(i,i')(j,j')}$ be related to the old transition probabilities q_{ij}?

5. We have a binary symmetric channel with reliability p. We take $\widehat{A} = \{000, 111\}$ as the new input alphabet. Find the new output alphabet and the new transition probabilities.

6. Here is a quite general way of modifying the output alphabet of a discrete channel that includes the idea of "amalgamation" discussed in Section 3.1 and the idea of "amalgamation with a coin flip" broached in Exercise 3.2.1. We may as well call this method *probabilistic amalgamation*. Suppose that $A = \{a_1, \dots, a_n\}$ and $B = \{b_1, \dots, b_k\}$ are the input and output alphabets, respectively, of a discrete memoryless channel, with transition probabilities q_{ij}. Let $\widetilde{B} = \{\beta_1, \dots, \beta_m\}$, $m \geq 2$, be a new (output) alphabet, and let u_{jt}, $j = 1, \dots, k$, $t = 1, \dots, m$, be probabilities satisfying $\sum_{t=1}^{m} u_{jt} = 1$ for each $j = 1, \dots, k$. We make $\widetilde{B}$ into the new output alphabet of the channel by declaring that b_j *will be read as* β_t *with probability* u_{jt}. That is, whenever b_j is the output letter, a probabilistic experiment is performed with outcomes $\beta_1, \dots, \beta_m$ and corresponding probabilities $u_{j1}, \dots, u_{jm}$ to determine which of the new output letters will be the output.

   (a) In each of Exercises 3.2.1 (c) and (d) identify $\widetilde{B}$ and give the matrix $U = [u_{tj}]$.

   (b) In general, supposing that B has been replaced by $\widetilde{B}$ as described, express the new matrix of transition probabilities $\widetilde{Q} = [\widetilde{q}_{it}]$ for the new channel with input alphabet A and output alphabet $\widetilde{B}$ in terms of the old matrix of transition probabilities Q and the matrix of probabilities $U = [u_{jt}]$.

   (c) Suppose that $A = \{0, 1\}$, $B = \{0, 1, *\}$, and

$$Q = \begin{bmatrix} q_{00} & q_{01} & q_{0*} \\ q_{10} & q_{11} & q_{1*} \end{bmatrix} = \begin{bmatrix} .9 & .02 & .08 \\ .05 & .88 & .07 \end{bmatrix}.$$

   Find a way to probabilistically amalgamate B to $\widetilde{B} = A$, so that the resulting channel is binary symmetric, and $u_{00} = u_{11} = 1$. (That is, find a 3×2 matrix $U = [u_{jt}]$ that will do the job.) Is there any other way (i.e., possibly with $u_{00} \neq 1$ or $u_{11} \neq 1$) to probabilistically amalgamate B to A to give a BSC with a greater reliability?

3.3 Input frequencies

As before, we have a memoryless channel with input alphabet $A = \{a_1, \dots, a_n\}$, output alphabet $B = \{b_1, \dots, b_k\}$, and transition probabilities q_{ij}. For $i \in \{1, \dots, n\}$, let p_i denote the *relative frequency of transmission*, or *input frequency*, of the input character a_i. In a large number of situations, it makes sense to think of p_i as the proportion of the occurrences of a_i in the text (written in input alphabetic characters) to be transmitted through the channel.

There is a bit of ambiguity here: by “the text to be transmitted” do we mean some particular segment of input text, or a “typical” segment of input text, or the totality of all possible input text that ever will or could be transmitted? This is one of those ambiguities that will never be satisfactorily resolved, we think; we shall just admit that “the input text” may mean different things on different occasions. Whatever it means, p_i is to be thought of as the (hypothetical) probability that a character selected at random from the input text will be a_i. This probability can sometimes be estimated by examining particular segments of text. For instance, if you count the number of characters, including punctuation marks and blanks, in this text, from the beginning of this section until the end of this sentence, and then tally the number of occurrences of the letter ‘e’ in the same stretch of text, you will find that ‘e’ accounts for a little less than 1/10 of all characters; its relative frequency is estimated, by this tally, to be around 0.096. You can take this as an estimate of the input frequency of the letter ‘e’ for any channel accepting the typographical characters of this text as input. This estimate is likely to be close to the “true” input frequency of ‘e’, if such there be, provided the text segments to be transmitted are not significantly different in kind from the sample segment from which 0.096 was derived. On the other hand, you might well doubt the validity of this estimate in case the text to be transmitted were the translation of “Romeo and Juliet” into Polish.

There are situations in which there is a way to estimate the input frequencies other than by inspecting a segment of input text. For instance, suppose we are trying to transmit data by means of a *binary code*; each datum is represented by a binary word, a member of $\{0,1\}^+$, and the binary word is input to a binary channel. We take $A = \{0,1\}$. Now, the input frequencies p_0 and p_1 of 0 and 1, respectively, will depend on the frequencies with which the various data emerge from the data source, and on how these are encoded as binary words. We know, in fact we control, the latter, but the former may well be beyond our powers of conjecture. If the probabilities of the various data emerging are known a priori, and the encoding scheme is agreed upon, then p_0 and p_1 can be calculated straightforwardly (see Exercise 1 at the end of this section).

Otherwise, when the relative frequencies of the source data are not known beforehand, it is a good working rule that different data are to be regarded as equally likely. The justification for this rule is ignorance; since probability in practice is an a priori assessment of likelihood, in case there is no prior knowledge you may as well assess the known possibilities as equally likely.

We now return to the general case, with $A = \{a_1,\dots,a_n\}$ and a_i having input frequency p_i. Observe that $\sum_{i=1}^{n} p_i = 1$. Also, note that the probabilities p_i have nothing to do with the channel; they depend on how we use the input alphabet to form text. They are therefore manageable, in principle; we feel that if we know enough about what is to be transmitted, we can make arrangements (in the *encoding* of the messages to be sent) so that the input frequencies of $a_1,\dots,a_n$ are as close as desired to any prescribed values $p_1,\dots,p_n \geq 0$ satisfying $\sum_{i=1}^{n} p_i = 1$. The practical difficulties involved in approaching prescribed input frequencies are part of the next chapter's subject. For now, we will ignore those difficulties and consider the p_i to be *variables*; we pretend to be able to vary them, within the constraints that $p_i \geq 0$, $i = 1, \ldots, n$ and $\sum_i p_i = 1$. In this respect the p_i are quite different from the q_{ij}, about which we can do nothing; the transition probabilities are constant parameters, forced upon us by the choice of channel.

We now focus on the act of attempting to transmit a single input character. We regard this as a two-stage experiment. The first stage: selecting some a_i for transmission. The second stage: observing which b_j emerges at the receiving end of the channel. We take, as the set of outcomes,

$$\mathcal{S} = \{(a_i, b_j); i \in \{1, \ldots, n\}, j \in \{1, \ldots, k\}\},$$

in which (a_i, b_j) is short for "a_i was selected for transmission, and b_j was received." We commit further semantic atrocities in the interest of brevity: a_i will stand for the event

$$\{(a_i, b_1), \ldots, (a_i, b_k)\} = \text{“}a_i \text{ was selected for transmission,”}$$

as well as standing for the ith input character; similarly b_j will sometimes denote the event "b_j was received." Readers will have to be alert to the context, in order to divine what means what. For instance, in the sentence "$P(a_i) = p_i$," it is evident that a_i stands for an event, not a letter.

3.3.1 With P denoting the probability assignment to $\mathcal{S}$, and noting the abbreviations introduced above, it seems that we are given the following:

(i) $P(a_i) = p_i$, and

(ii) $P(b_j \mid a_i) = q_{ij}$, whence

(iii) $P(a_i, b_j) = P(a_i \cap b_j) = p_i q_{ij}$, and

(iv) $P(b_j) = \sum_{t=1}^{n} p_t q_{tj}$.

The probabilities $P(b_j)$ in (iv) are called the *output frequencies* of b_j, $j = 1, \ldots, k$. It is readily checked that $P(b_j) \geq 0$ and $\sum_{j=1}^{k} P(b_j) = 1$.

Now, the careful and skeptical reader will, we hope, experience a shiver of doubt in thinking all of this over. Putting aside qualms about memorylessness and the invariability of the q_{ij}, there is still an infelicity in the correspondence between the "model" and "reality" in this two-stage experiment view of transmission of a single character, and the problem is in the first stage. In order to assert (i), above, we must view the process of "selecting an input character" as similar to drawing a ball from an urn; we envision a large urn, containing balls colored $a_1, \ldots, a_n$, with proportion p_i of them colored a_i, $i = 1, \ldots, n$. Attempting to transmit a string of input symbols means successively drawing balls from this urn, with replacement and remixing after each draw; this is what our "model" says we are up to.

The problem is that this does not seem much like what we actually do when dealing with input text. If you are at some point in the text, it doesn't seem that the next character pops up at random from an urn; it seems that the probabilities of the various characters appearing next in the text ought to be affected by where we are in the text, by what has gone before. This is certainly the way it is with natural languages; for instance, in English, ‘u’ almost always follows ‘q’ and ‘b’ rarely follows ‘z’. Thus, for the English-speaking reader, the “draw from an urn” model of “selecting the next character for transmission” breaks down badly, when the input text is in English.

Notice also the situation of Exercise 3.3.1. Because of the way the source messages are encoded, it seems intuitively obvious that whenever a 0 is input, the probability of the next letter for transmission being 0 is greater than p_0, the relative frequency of 0 in the input text. (And that is, in fact, the case. You might verify that, after a 0, assuming we know nothing else of what has been transmitted already, the probability that the next letter will be 0 is $17/24$, while $p_0 < 1/2$.)

Nevertheless, we shall hold to the simplifying assumption that p_i, the proportion of a_i’s in the input text, is also the probability that the *next* letter is a_i, at any point in the input stream. This assumption is valid if we are ignorant of grammar and spelling in the input language; we are again, as with the transition probabilities, in the weird position of bringing a probability into existence by assuming ignorance. In the case of the transition probabilities, that assumption of ignorance is usually truthful; in the present case, it is more often for convenience, because it is difficult to take into account what we know of the input language. There are ways to analyze information transfer through discrete channels with account taken of grammar and/or spelling in the input language—see Shannon’s paper [65], and the discussion in Chapter 7 of this text. We shall not burden the reader here with that more difficult analysis, but content ourselves with a crude but useful simplification, in this introduction to the subject.

Exercises 3.3

1. Suppose a data, or message, source gives off, from time to time, any one of three data, or messages, M_1, M_2, and M_3. M_1 accounts for 30% of all emanations from the source, M_2 for 50%, and M_3 for 20%.

   These messages are to be transmitted using a binary channel. To this end, M_1 is encoded as 11111, M_2 as 100001, and M_3 as 1100. Find p_0 and p_1, the input frequencies of 0 and 1, respectively, into the channel to be used for this task.

   [Hint: suppose that a large number N of messages are in line to be transmitted, with $3N/10$ of them instances of M_1, $N/2$ of them M_2, and $N/5$ of them M_3. Count up the number of 0’s and the number of 1’s in the corresponding input text.]

2. Same question as in Exercise 1 above, except that nothing is known about the relative frequencies of M_1, M_2, and M_3; apply the convention of assuming that M_1, M_2, and M_3 are equally likely.

3. A binary symmetric channel with reliability p is used in a particular communication task for which the input frequencies of 0 and 1 are $p_0 = 2/3$ and $p_1 = 1/3$. Find the output frequencies of 0 and 1 in terms of p. [Hint: apply 3.3.1(iv).]
4. Let $A = \{a_1, a_2, a_3\}$, $B = \{b_1, b_2, b_3\}$,
$$Q = \begin{bmatrix} .94 & .04 & .02 \\ .01 & .93 & .06 \\ .03 & .04 & .93 \end{bmatrix},$$
$p_1 = .4$, $p_2 = .5$, and $p_3 = .1$. Find the output frequencies, $P(b_1)$, $P(b_2)$, and $P(b_3)$.
5. Suppose, in using the channel of the preceding problem, there is a *cost* associated with each attempted transmission of a single input letter. Suppose the (i, j)-entry of the following matrix gives the cost, to the user, of b_j being received when a_i was sent, in some monetary units:
$$C = \begin{bmatrix} 0 & 5 & 9 \\ 10 & 0 & 2 \\ 4 & 2 & 0 \end{bmatrix}.$$
   (a) Express, in terms of p_1, p_2, and p_3, the average cost per transmission-of-a-single-input-letter of using this channel. Evaluate when $p_1 = .4$, $p_2 = .5$, and $p_3 = .1$.
   (b) What choice of p_1, p_2, p_3 minimizes the average cost-per-use of this channel? Would the user be wise to aim to minimize that average cost?

3.4 Channel capacity

A, B, q_{ij}, and the p_i will be as in the preceding section. With the a_i standing for events, not characters, $A = \{a_1, \ldots, a_n\}$ is a system of events in the probability space associated with the two-stage experiment of sending a single character through a memoryless channel with input alphabet A and output alphabet B. Observe that we have taken on yet another risk of misunderstanding; A will sometimes be an alphabet, sometimes a system of events, and you must infer which from the context. When a system of events, A, is the *input system* of events, for the channel with input alphabet A. Similarly, $B = \{b_1, \ldots, b_k\}$ will sometimes stand for a system of events called the *output system*.

We are interested in communication, the transfer of information; it is reasonable to suppose that we ought, therefore, to be interested in the mutual information between the input and output systems,

$$\begin{aligned} I(A,B) &= \sum_{i=1}^{n}\sum_{j=1}^{k} P(a_i \cap b_j)\log\frac{P(a_i \cap b_j)}{P(a_i)P(b_j)} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{k} p_i q_{ij}\log\frac{p_i q_{ij}}{p_i \sum_{t=1}^{n} p_t q_{tj}} \\ &= \sum_{i=1}^{n} p_i \sum_{j=1}^{k} q_{ij}\log\frac{q_{ij}}{\sum_{t=1}^{n} p_t q_{tj}}. \end{aligned}$$

$I(A,B)$ is a function of the variables $p_1,\dots,p_n$, the input frequencies. It would be interesting to know the maximum value that $I(A,B)$ can have. That maximum value is called the *capacity* of the channel, and any values of $p_1,\dots,p_n$ for which that value is achieved are called *optimal input* (or *transmission*) *frequencies* for the channel. If you accept $I(A,B)$ as an index, or measure, of the potential effectiveness of communication attempts using this channel, then the capacity is the fragile acme of effectiveness. This peak is achieved by optimally adjusting the only quantities within our power to adjust, once the hardware has been established and the input alphabet has been agreed to, namely, the input frequencies. The main result of this section will show how to find the optimal input frequencies (in principle). But before launching into the technical details, let us muse a while on the meaning of what it is that we are optimizing.

3.4.1 *Shannon's interpretation of $I(A,B)$ as rate of information transfer or flow.* Suppose that input letters are arriving at the transmitter at the rate of r letters per second. The average information content of an input letter is $H(A)$; therefore, since the average of a sum is the sum of the averages, there are, on average, $rH(A)$ units of information per second arriving at the transmitter. The information flow is mussed up a bit by the channel; at what average rate is information "flowing" through the channel?

C. E. Shannon's answer [63, 65]: at the rate $rI(A,B) = r(H(A) - H(A \mid B))$. This answer becomes plausible if you bear down on the interpretation of $H(A \mid B)$ as a measure of the average *uncertainty* of the input letter, conditional upon knowing the output letter. Shannon calls $H(A \mid B)$ the "average ambiguity of the received signal," or "the equivocation," and this last terminology has taken root. Note that "the equivocation" is not dependent on the channel alone, but also on the input frequencies. In Shannon's interpretation, it is the amount of information removed, on average, by the channel from the input stream, per input letter.

The validity of this interpretation is bolstered by the role the equivocation plays in Shannon's Noisy Channel Theorem, which we will encounter later. For right now, here is an elementary example due to Shannon himself.

Let the base of log be 2, so the units of information are bits. Suppose we have a binary symmetric channel with reliability .99, and the input is streaming into the receiver at the rate of 1000 symbols (binary digits) per second, with input frequencies $p_0 = p_1 = 1/2$. These are, we shall see soon, optimal, and give

$I(A,B) = .99\log 1.98 + .01\log .02 \approx .919$. By the interpretation of $I(A,B)$ under consideration, this says that information is flowing to the receiver at the average rate of $1000(.919) = 919$ bits per second.

You might object that, on average, 990 of the 1000 digits arriving at the receiver each second are *correct* (i.e., equal to the digit transmitted), so perhaps 990 bits/second ought to be the average rate of information flow to the receiver. Shannon points out that by that reasoning, if the reliability of the channel were $1/2$, i.e., if the channel were perfectly useless, you would compute that information flow to the receiver at 500 bits/second, on average, whereas the true rate of information flow in this case ought to be zero. The problem is, whether $p = 1/2$ or $p = .99$, we do not know which of the $1000p$ correct digits (on average, each second) are correct; our uncertainty in this regard means that our estimate of the rate of information flow to the receiver ought to be revised downward from $1000p$ bits/sec. (Verify: $p\log_2 2p + (1-p)\log_2 2(1-p) < p$, $1/2 \le p < 1$.) Why this particular revision, from 990 down to 919 bits/sec? This is where $H(A \mid B) = -(.01\log .01 + .99\log .99) \approx .081$ comes in; supposing you know which letter, 0 or 1, is received, $H(A \mid B)$ is the entropy, i.e., average uncertainty, of the *input letter* (system), so it is a good measure of the amount of information to be subtracted from one (the number of bits just received) due to uncertainty about what was sent. (Convinced? Feel uncertain about something? Well, that's entropy, and it's good for you, taken in moderation.)

It is preferable to speak of $I(A,B)$ as the average information *flow through* the channel, or *flow to* the receiver, per input letter, rather than as the average amount of information *arriving at* the receiver (per input letter). The latter might reasonably be taken to be $H(B)$, which is, indeed, the average amount of information contained in the set of outcomes of the probabilistic experiment of "choosing" an input letter and then attempting to transmit it, if we were to take B as the set of outcomes; and taking B as the set of outcomes does seem to respond to the question of how much information is *arriving at* the receiver, per input letter. But $H(B)$ as a measure of information has no connection with how well the channel is communicating the input stream. For instance, for a BSC with reliability 1/2, $H(B) = \log 2$, while surely the level of communication ought to be $0 = I(A,B)$.

Use of the word "flow" in this context will aid in understanding the Noisy Channel Theorem, in Section 4.6. That theorem discloses a remarkable analogy between information flowing through a channel and fluid flowing through a pipe.

3.4.2 Supposing the transition probabilities q_{ij} are known, finding the optimal input frequencies for, and thus the capacity of, a given channel is a straightforward multi-variable optimization problem; we wish to find where $I(A,B)$, as a function of $p_1,\dots,p_n$, achieves its maximum on

$$K_n = \{(p_1,\dots,p_n) \in \mathbb{R}^n;\ p_1,\dots p_n \ge 0 \text{ and } \textstyle\sum_{i=1}^n p_i = 1\}.$$

By convention, the terms in the sum for $I(A,B)$ corresponding to pairs

(i, j) for which $q_{ij} = 0$ do not actually appear in that sum. Note that if $p_t > 0$, $t = 1, \ldots, n$ and $\sum_{t=1}^{n} p_t q_{tj} = 0$, then $q_{tj} = 0$, $t = 1, \ldots, n$. It follows that the formula for $I(A, B)$ defines a differentiable function in the positive part of $\mathbb{R}^n$, $\{(p_1, \ldots, p_n) \in \mathbb{R}^n;\ p_t > 0,\ t = 1, \ldots, n\}$. Consequently, the Lagrange Multiplier Theorem asserts that if $I(A, B)$ achieves a maximum on K_n in $K_n^+ = \{(p_1, \ldots, p_n) \in K_n;\ p_i > 0,\ i = 1, \ldots, n\}$, then the maximum is necessarily achieved at a point where $\frac{\partial}{\partial p_k}(I(A, B) - \lambda \sum_{i=1}^{n} p_i) = 0$, $k = 1, \ldots, n$, for some λ.

The main content of Theorem 3.4.3, below, is that a sort of converse of this statement holds: if the equations arising from the Lagrange Multiplier Theorem hold at a point $(p_1, \ldots, p_n) \in K_n^+$, then $I(A, B)$ necessarily achieves a maximum, on K_n, at $(p_1, \ldots, p_n)$. The proof of this statement is a bit technical, and is relegated to the next section, which is optional; although it is preferable that even students of applied mathematics understand the theoretical foundations of their subject, in this case it probably won't overly imperil your immortal soul to accept the result without proof.

Let us see where the Lagrange Multiplier method tells us to look for the optimal input frequencies. Setting $F(p_1, \ldots, p_n) = I(A, B) - \lambda \sum_{i=1}^{n} p_i$, considering only points $(p_1, \ldots, p_n)$ where all coordinates are positive, and setting $c = \log(e)$, we have

$$\begin{aligned}
\frac{\partial F}{\partial p_s} &= \frac{\partial}{\partial p_s}(I(A, B)) - \lambda \\
&= \sum_{j=1}^{k} q_{sj} \log \frac{q_{sj}}{\sum_{t=1}^{n} p_t q_{tj}} - c \sum_{i=1}^{n} p_i \sum_{j=1}^{k} \frac{q_{ij} q_{sj}}{\sum_{t=1}^{n} p_t q_{tj}} - \lambda \\
&= \sum_{j=1}^{k} q_{sj} \log \frac{q_{sj}}{\sum_{t=1}^{n} p_t q_{tj}} - c \sum_{j=1}^{k} \frac{\sum_{i=1}^{n} p_i q_{ij}}{\sum_{t=1}^{n} p_t q_{tj}} q_{sj} - \lambda \\
&= \sum_{j=1}^{k} q_{sj} \log \frac{q_{sj}}{\sum_{t=1}^{n} p_t q_{tj}} - c \sum_{j=1}^{k} q_{sj} - \lambda \\
&= \sum_{j=1}^{k} q_{sj} \log \frac{q_{sj}}{\sum_{t=1}^{n} p_t q_{tj}} - (c + \lambda).
\end{aligned}$$

Replacing $c + \lambda$ by C, and setting the partial derivative equal to 0, we obtain the *capacity equations* for the channel.

3.4.3 Theorem *Suppose a memoryless channel has input alphabet $A = \{a_1, \ldots, a_n\}$, output alphabet $B = \{b_1, \ldots, b_k\}$, and transition probabilities q_{ij}, $i \in \{1, \ldots, n\}$, $j \in \{1, \ldots, k\}$. There are optimal input frequencies for this channel. If $p_1, \ldots, p_n$ are positive real numbers, then $p_1, \ldots, p_n$ are optimal input frequencies for this channel if and only if $p_1, \ldots, p_n$ satisfy the following, for* *some value of C:*

$$\sum_{i=1}^{n} p_i = 1 \quad \textit{and} \quad \sum_{j=1}^{k} q_{sj} \log \frac{q_{sj}}{\sum_{t=1}^{n} p_t q_{tj}} = C,\ s = 1, \dots, n.$$

Furthermore, if $p_1, \dots, p_n$ are optimal input frequencies satisfying these equations, for some value of C, then C is the channel capacity.

This theorem may seem, at first glance, to be saying that all you have to do to find the capacity of a channel and the optimal input frequencies is to solve the capacity equations of the channel, the equations arising from the Lagrange Multiplier Theorem, and the condition $\sum_{i=1}^{n} p_i = 1$, for $p_1, \dots, p_n > 0$. There is a loophole, however, a possibility that slips through a crack in the wording of the theorem: it is possible that the capacity equations have no solution. See problems 9 and 14 at the end of this section. Note that in problem 9, it is not just that the equations have no solution $(p_1, \dots, p_n)$ with all the p_i positive; the equations have no solution, period.

From Theorem 3.4.3 you can infer that this unpleasant phenomenon, the capacity equations having no solution, occurs only when the capacity is achieved at points $(p_1, \dots, p_n) \in K_n$ with one or more of the p_i equal to zero. If $p_i = 0$, then a_i is never used; we have thrown away an input character; we are not using all the tricks at our disposal. Problems 9 and 14 show that it can, indeed, happen that there are input characters that we are better off without. Note, however, the result of problem 10, in which the channel quite severely mangles and bullies one of the input letters, a_n, while maintaining seamlessly perfect respect of the others; yet, in the optimal input frequencies, p_n is positive, which shows that we are better off using a_n than leaving it out, in spite of how terribly the channel treats it (provided we accept $I(A, B)$ as a measure of how well off we are). In this respect, note also the results of exercise problems 2, 6 (a special case of problem 10 when $p = 1/2$), and 7. The practical moral to be drawn from these examples seems to be that if the channel respects an input character even a little bit, if you occasionally get some information from the output (upon inputting this character) about the input, then you are better off with the character than without it. The surprising result of Exercise 14 obliterates this tentative conclusion, and shows that we may be in the presence of a mystery.

How will we know when we are in the rare necessity of banishing one or more input characters, and what do we do about determining the optimal input frequencies in such cases? According to Theorem 3.4.3, we are in such a case when and only when the capacity equations of the channel have no solution in K_n^+. In such a situation, the n-tuple $(p_1, \dots, p_n)$ of optimal input frequencies lies on one of the faces of K_n, $F_R = \{(p_1, \dots, p_n) \in K_n; p_i > 0 \text{ for } i \in R \text{ and } p_i = 0 \text{ for } i \notin R\}$, where R is a proper subset of $\{1, \dots, n\}$. For such an R, let $A_R = \{a_i \in A; i \in R\}$, the input alphabet obtained by deleting the a_i indexed by indices not in R. Finding $(p_1, \dots, p_n)$ on F_R amounts to solving the channel capacity problem with A replaced by A_R; if $(p_1, \dots, p_n) \in F_R$ is the n-tuple of optimal input frequencies, then the non-zero p_i, those indexed by $i \in R$, will satisfy the capacity equations associated with this modified problem. (These equations are obtainable from the original capacity equations by omitting those p_i and q_{ij} with $i \notin R$.)

Thus, if the capacity equations for the channel have no solution $(p_1, \dots, p_n)$ with $p_i > 0$, $i = 1, \dots, n$, we need merely solve the $2^n - n - 2$ systems of capacity equations associated with the A_R, for R satisfying $2 \le |R| \le n - 1$. It is a consequence of Theorem 3.4.3 that we may first consider all A_R with $|R| = n - 1$, and from among the various solutions select one for which the corresponding capacity is maximal. If there are no solutions, move on to A_R with $|R| = n - 2$, and so on. All of this is straightforward, but it is also a great deal of trouble; we hope that in most real situations the optimal input frequencies will be all positive.

3.4.4 As mentioned above, the proof of the main assertion of 3.4.3 is postponed until the next section, the last of this chapter. However, we can give the proof of the last assertion here. If $p_1, \dots, p_n$ satisfy the equations above, then the value of $I(A, B)$ at $(p_1, \dots, p_n)$ is

$$I(A, B) = \sum_{i=1}^{n} p_i \sum_{j=1}^{k} q_{ij} \log \frac{q_{ij}}{\sum_{t=1}^{n} p_t q_{tj}} = C \sum_{i=1}^{n} p_i = C.$$

To remember the capacity equations, other than $\sum_{i=1}^{n} p_i = 1$, it is helpful to remember that the left-hand side of

$$\sum_{j=1}^{k} q_{sj} \log \frac{q_{sj}}{\sum_{t=1}^{n} p_t q_{tj}} = C$$

is the thing multiplying p_s in the formula for

$$I(A, B) = \sum_{i=1}^{n} p_i \sum_{j=1}^{k} q_{ij} \log \frac{q_{ij}}{\sum_{t=1}^{n} p_t q_{tj}}.$$

3.4.5 *The capacity of a binary symmetric channel.* Suppose a binary symmetric channel has reliability p. Let p_0, p_1 denote the input frequencies of 0 and 1, respectively. The capacity equations are:

(1) $p_0 + p_1 = 1$,

(2) $p \log \dfrac{p}{p_0 p + p_1(1-p)} + (1-p) \log \dfrac{1-p}{p_0(1-p) + p_1 p} = C$, and

(3) $(1-p) \log \dfrac{1-p}{p_0 p + p_1(1-p)} + p \log \dfrac{p}{p_0(1-p) + p_1 p} = C$.

Setting the left-hand sides of (2) and (3) equal, and canceling $p \log p$ and $(1-p) \log(1-p)$, we obtain

$$\begin{aligned} &p \log(p_0 p + p_1(1-p)) + (1-p) \log(p_0(1-p) + p_1 p) \\ &\qquad = (1-p) \log(p_0 p + p_1(1-p)) + p \log(p_0(1-p) + p_1 p), \end{aligned}$$

whence

$$(2p-1)\log(p_0 p + p_1(1-p)) = (2p-1)\log(p_0(1-p)+p_1 p),$$

so either $p = 1/2$ or

$$p_0 p + p_1(1-p) = p_0(1-p) + p_1 p,$$

i.e.,

$$(2p-1)p_0 = (2p-1)p_1,$$

so $p_0 = p_1 = 1/2$ (in view of (1)) if $p \neq 1/2$, and the channel capacity is $C = p\log 2p + (1-p)\log 2(1-p)$, obtainable by plugging $p_0 = p_1 = 1/2$ into either (2) or (3) above.

If $p = 1/2$, then, since

$$\frac{p}{p_0 p + p_1(1-p)} = \frac{1-p}{p_0(1-p)+p_1 p} = 1$$

for all values of p_0, p_1 satisfying $p_0 + p_1 = 1$, in this case, we have $I(A,B) = 0$ for all values of p_0, p_1. This is as it should be, since when $p = 1/2$, sending a digit through this channel is like flipping a fair coin. We learn nothing about the input by examining the output, the input and output systems are statistically independent, the channel is worthless for communication.

Note that it is not obvious, a priori, that $p\log 2p + (1-p)\log 2(1-p)$ is positive for all values of $p \in [0,1]\setminus\{1/2\}$, but that this is the case follows from Theorem 2.2.13.

The foregoing shows that when $p \neq 1/2$, $p_0, p_1 = 1/2$ are the unique optimal input frequencies of a binary symmetric channel of reliability p. If we had wished only to verify that $p_0 = p_1 = 1/2$ are optimal—i.e., if the uniqueness is of no interest—then we could have saved ourselves some trouble, and found the capacity, by simply noting that $p_0 = p_1 = 1/2$ satisfy (1) and make the left-hand sides of (2) and (3) equal. The optimality of $p_0 = p_1 = 1/2$, and the expression for C, then follow from Theorem 3.4.3. For a generalization of this observation, see 3.4.7, below.

3.4.6 Here are two questions of possible practical importance that are related, and to which the answers we have are incomplete:

(i) When (under what conditions on Q) are the optimal input frequencies of a channel unique?

(ii) Do the optimal input frequencies of a channel depend continuously on the transition probabilities of the channel?

Regarding (i), the only instances we know of when the optimal input frequencies are not unique are when the capacity of the channel is zero. (Certainly, in this case, any input frequencies will be optimal; but the remarkable thing is that it is *only* in this case that we have encountered non-unique optimal input frequencies.) We hesitantly conjecture that if the channel capacity is non-zero, then the optimal input frequencies are unique. For those interested, perhaps the proof in Section 3.5 will reward study.

Regarding (ii), there is a body of knowledge related to the Implicit Function Theorem in the calculus of functions of several variables that provides an answer of sorts. Regarding the left-hand sides of the capacity equations as functions of both the p_i and q_{ij}, supposing there is a solution of the equations at positive $p_i, i = 1, \ldots, n$, and supposing that a certain large matrix of partial derivatives has maximum rank, then for every small wiggle of the q_{ij} there will be a positive solution of the new capacity equations quite close to the solution of the original system. When will that certain large matrix of partial derivatives fail to have maximum rank? We can't tell you exactly, but the short answer is: almost never. Thus, the answer to (ii) is: yes, except possibly in certain rare pathological circumstances that we haven't worked out yet.

Here is an example illustrating the possible implications and uses of the continuous dependence of the optimal input frequencies on the transition probabilities. Suppose that $A = \{0, 1\}, B = \{0, 1, *\}$, and

$$Q = \begin{bmatrix} q_{00} & q_{01} & q_{0*} \\ q_{10} & q_{11} & q_{1*} \end{bmatrix} = \begin{bmatrix} .93 & .02 & .05 \\ .01 & .95 & .04 \end{bmatrix}.$$

Then Q is "close" to $\widetilde{Q} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$ which is the matrix of transition probabilities of a BSC. (For the channel associated with $\widetilde{Q}$, $*$ has been removed as an output letter.) Therefore the optimal input frequencies of the original channel are "close" to $p_0 = p_1 = 1/2$ – and the channel capacity is "close" to log 2. Caution: there is a risk involved in rough estimation of this sort. For instance, would you say that the matrix of transition probabilities in Exercise 3.4.14 is "close" to $\begin{bmatrix} 1/2 & 1/4 & 1/4 \\ 1/4 & 1/2 & 1/4 \\ 1/4 & 1/4 & 1/2 \end{bmatrix}$? If you are in a reckless mood, you might well do so, yet the optimal input frequencies for the channel with the latter matrix of transition probabilities are $1/3, 1/3, 1/3$ (this will be shown below), while the optimal input frequencies for the channel of problem 14 are $1/2, 0, 1/2$. Disconcerting discrepancies of this sort should chasten our fudging and make us appreciate numerical error analysis of functions of several variables. But we will pursue this matter no further in this text.

3.4.7 *n*-ary symmetric channels An n-ary symmetric channel of reliability p is a discrete memoryless channel with

$$A = B \quad \text{and} \quad Q = \begin{bmatrix} p & & \frac{1-p}{n-1} \\ & \ddots & \\ \frac{1-p}{n-1} & & p \end{bmatrix};$$

that is, the main diagonal entries of Q are all the same (namely, p), and the off-diagonal entries of Q are all the same. (Their common value will have to be $\frac{1-p}{n-1}$ if the row sums are to be 1.)

It is straightforward to verify that $p_1 = \cdots = p_n = 1/n$ satisfy the capacity equations of such a channel, with

$$\begin{aligned} C =& p \log np + (1-p)\log\frac{n(1-p)}{n-1} \\ =& \log n + (p\log p + (1-p)\log\frac{1-p}{n-1}), \end{aligned}$$

so by Theorem 3.4.3, $(1/n, \ldots, 1/n)$ are optimal input frequencies for the channel and the capacity is C, above. These optimal input frequencies and this capacity are also discoverable by the method explained in the exercise section, after Exercise 3.4.12, and this method has the advantage that by it and the application of a little linear algebra theory, it can easily be seen that $p_i = 1/n$, $i = 1, \ldots, n$ are *unique* optimal input frequencies except in the case $p = 1/n$, which is precisely the case $C = 0$.

Exercises 3.4

1. Verify directly that $f(p) = p\log 2p + (1-p)\log 2(1-p)$ achieves its maximum, $\log 2$, on [0, 1] at the endpoints, 0 and 1, and its minimum, 0, at 1/2.

2. Verify that the value of $I(A,B)$ at the extreme points $\{(1,0,\ldots,0), (0,1,0,\ldots,0), \ldots, (0,\ldots,0,1)\}$ of K_n is zero.

3. Suppose $A = B = \{0,1\}$, but the channel is not symmetric; suppose a transmitted 0 has probability p of being received as 0, and a transmitted 1 has probability q of being received as 1. Let p_0 and p_1 denote the input frequencies. In terms of p, q, p_0, and p_1, write $I(A,B)$, and give the capacity equations for this channel.

4. Give $I(A,B)$ and the capacity equations for the channel described in Exercise 3.3.4.

5. $A = \{0,1\}$, $B = \{0,1,*\}$, and the channel treats the input characters symmetrically; for each input, 0 or 1, the probability that it will be received as sent is p, the probability that it will be received as the other digit is q, and the probability that it will be received as $*$ is r. Note that $p+q+r = 1$.

   Find, in terms of p, q, and r, the capacity of this channel and the optimal input frequencies.

6. $A = B = \{a,b\}$; a is always transmitted correctly; when b is transmitted, the probability is p that b will be received (and, thus, $1-p$ that a will be received). Find, in terms of p, the capacity of this channel and the optimal input frequencies. Verify that even when $p = 1/2$ (a condition of maximum disrespect for the input letter b), the capacity is positive (which is greater than the capacity would be if the letter b were discarded as an input letter—see Exercise 2, above).

7. [Part of this exercise was lifted from [37].] $A = B = \{a,b,c\}$, a is always transmitted correctly, and the channel behaves symmetrically with respect

to b and c. Each has probability p of being transmitted correctly, and probability $1-p$ of being received as the other character (c or b). (Thus, if a is received, it is certain that a was sent.)

(a) Find the capacity of this channel, and the optimal input frequencies, as functions of p.

(b) Suppose that c is omitted from the input alphabet (but not the output alphabet). Find the capacity of the channel and the optimal input frequencies in this new situation.

(c) Are there any values of p for which the capacity found in (b) is greater than that in (a)? What about the case $p = 1/2$?

8. Suppose that $A = B = \{a_1, \ldots, a_n\}$, and the channel is perfectly reliable: when a_i is sent, a_i is certain to be received. Find the capacity of this channel and the optimal input frequencies.

9. Suppose that $A = \{a_1, \ldots, a_{n+1}\}$, $B = \{a_1, \ldots, a_n\}$, and the channel respects $a_1, \ldots, a_n$ perfectly; when a_i is sent, a_i is certain to be received, $1 \leq i \leq n$.

   (a) Suppose that when a_{n+1} is sent, the output characters $a_1, \ldots, a_n$ are equally likely to be received. Show that the capacity equations for the channel have no solution in this case. Find the optimal input frequencies and the capacity of this channel.

   (b) Are there any transition probabilities $q_{n+1,j}$, $j = 1, \ldots, n$, for which there are optimal input frequencies $p_1, \ldots, p_{n+1}$ for this channel with $p_{n+1} > 0$? If so, find them, and find the corresponding optimal input frequencies and the channel capacity.

10. Suppose that $n \geq 2$, $A = \{a_1, \ldots, a_n\} = B$, and the channel respects $a_1, \ldots, a_{n-1}$ perfectly. Suppose that, when a_n is sent, the output characters $a_1, \ldots, a_n$ are equally likely to be received. Find the optimal input frequencies and the capacity of this channel.

11. We have a binary symmetric channel with reliability p, but we take $A = \{000, 111\}$. Let the input frequencies be denoted p_0 and p_1. In terms of p, p_0, and p_1, write the mutual information between inputs and outputs, and the capacity equations of this channel. Assuming that $p_0 = p_1 = 1/2$ are the optimal input frequencies, write the capacity of this channel as a function of p.

12. (a) Show that $I(A,B) \leq H(A)$. (This is a special case of a result in Section 2.4.)

    (b) Show that $I(A,B) = H(A)$ if and only if for each letter b_j received, there is exactly one input letter a_i such that $P(a_i \mid b_j) = 1$ (so $P(a_k \mid b_j) = 0$ for $k \neq i$). [Hint: recall that $H(A \mid B) = H(A) - I(A,B)$; use Theorem 2.3.5 or its proof.] In other words, $I(A,B) = H(A)$ if and only if the input is determinable with certainty from the output. In yet other words, $I(A,B) = H(A)$ if and only if the input system of events is an amalgamation of the output system.

For exercises 13–15, we are indebted to Luc Teirlinck, who observed that

$$I(A,B) = H(B) - H(B \mid A),$$

so that if

$$-H(B \mid A) = \sum_i p_i \sum_j q_{ij} \log q_{ij} \qquad \text{[verify!]}$$

does not depend on $p_1,\dots,p_n$, as it will not if the sums $S_i = \sum_j q_{ij} \log q_{ij}$ are all the same, $i = 1,\dots,n$, then $I(A,B)$ is maximized when $H(B)$ is. The obvious way to maximize $H(B)$ is to “make” $P(b_j) = \sum_{t=1}^{n} p_t q_{tj}$ equal to $1/k$, $j = 1,\dots,k$. Thus, in these cases, the optimal input frequencies $p_1,\dots,p_n$ *might* be found by solving the linear system

$$\begin{aligned} p_1 + \cdots + p_n &= 1 \\ \sum_{t=1}^{n} p_t q_{tj} &= 1/k, \quad j = 1,\dots,k. \end{aligned}$$

[The first equation is redundant: to see this, sum the r equations just above over j.] This method is not certain to succeed because the solutions of this linear system may fail to be non-negative, or may fail to exist.

Notice that the sums S_i will be all the same if each row of Q is a rearrangement of the first row.

13. Find the optimal input frequencies when

$$Q = \begin{bmatrix} 2/3 & 0 & 1/3 \\ 1/3 & 2/3 & 0 \\ 0 & 1/3 & 2/3 \end{bmatrix}.$$

Also, find the capacity of the channel.

14. Find the optimal input frequencies and the channel capacity, when

$$Q = \begin{bmatrix} 1/2 & 1/3 & 1/6 \\ 1/6 & 1/2 & 1/3 \\ 1/6 & 1/3 & 1/2 \end{bmatrix}.$$

15. Suppose that $n \ge 3$, $0 \le p \le 1$, and

$$Q = \begin{bmatrix} p & 0 & \dots & 0 & 1-p \\ 0 & p & \dots & 0 & 1-p \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & p & 1-p \\ 1-p & 0 & \dots & 0 & p \end{bmatrix}.$$

(a) For which values of p does the method of solving a linear system give the optimal input frequencies for this channel?

*(b) What are the optimal input frequencies and the channel capacity, in terms of p, in all cases?

Exercises 14 and 15 are instructive for those interested in the problem of getting conditions on Q under which the optimal input frequencies are unique and positive.

*16. Suppose a channel has input alphabet A, output alphabet B, and capacity C. Suppose we take A^k as the new input alphabet. Show that the new capacity is kC. (This result is a theorem in [81]. You may find the results of 2.4 helpful, as well as the result of exercise 2.3.6.)

3.5* Proof of Theorem 3.4.3, on the capacity equations

By the remarks of the preceding section, what remains to be shown is that (i) $I(A, B)$ does achieve a maximum on K_n and (ii) if the capacity equations are satisfied, for some C, by some $p_1, \dots, p_n > 0$, then $p_1, \dots, p_n$ are optimal input frequencies for the channel.

Since K_n is closed and bounded, to prove (i) it suffices to show that $I(A, B)$ is continuous on K_n. This may seem trivial, since $I(A, B)$ appears to be given by a formula involving only linear functions of $p_1, \dots, p_n$ and log, but please note that this formula is valid at points $(p_1, \dots, p_n) \in K_n \setminus K_n^+$ only by convention; there is trouble when one or more of the p_i is zero. Still, the verification that $I(A, B)$ is continuous at such points is straightforward, and is left to the reader to sort out. Keep in mind that $x \log x \to 0$ as $x \to 0^+$. See problem 1 at the end of this section.

A real-valued function f defined on a convex subset K of $\mathbb{R}^n$ is said to be *concave* if

$$f(tu + (1-t)v) \geq tf(u) + (1-t)f(v) \text{ for all } u, v \in K,\ t \in [0, 1].$$

If strict inequality holds whenever $u \neq v$ and $t \in (0, 1)$, we will say that f is *strictly concave*.

We shall now list some facts about concave functions to be used to finish the proof of Theorem 3.4.3. Proofs of these facts are omitted. It is recommended that the reader try to supply the proofs. Notice that 3.5.3 and 3.5.4, taken together, constitute the well-known "second derivative test" for concavity and relative maxima of functions of one variable.

3.5.1 Any sum of concave functions is concave, and if one of the summands is strictly concave, then the sum is strictly concave. A positive constant times a (strictly) concave function is (strictly) concave.

3.5.2 Any linear function is concave, and the composition of a linear function with a concave function of one variable is concave.

3.5.3 If $I \subseteq \mathbb{R}$ is an interval, $f : I \to \mathbb{R}$ is continuous, and $f'' \leq 0$ on the interior of I, then f is concave on I. If $f'' < 0$ on the interior of I, then f is strictly concave on I.

3.5.4 If $I \subseteq \mathbb{R}$ is an interval, $f : I \to \mathbb{R}$ is concave on I, and $f'(x_0) = 0$ for some $x_0 \in I$, then f achieves a maximum on I at x_0. If f is strictly concave on I and $f'(x_0) = 0$, then f achieves a maximum on I only at x_0.

Now we are ready to finish the proof of Theorem 3.4.3. Let

$$f(x) = \begin{cases} -x\log x, & x > 0 \\ 0, & x = 0. \end{cases}$$

By 3.5.3, f is strictly concave on $[0, \infty)$. Now,

$$\begin{aligned} I(A,B) &= \sum_{i=1}^{n} p_i \Big(\sum_{j=1}^{k} q_{ij}\log q_{ij}\Big) - \sum_{j=1}^{k}\Big(\sum_{i=1}^{n} p_i q_{ij}\Big)\log\Big(\sum_{t=1}^{n} p_t q_{tj}\Big) \\ &= \sum_{i=1}^{n}\Big(\sum_{j=1}^{k} q_{ij}\log q_{ij}\Big)p_i + \sum_{j=1}^{k} f\Big(\sum_{i=1}^{n} p_i q_{ij}\Big), \end{aligned}$$

so by 3.5.1 and 3.5.2, $I(A,B)$ is a concave function on K_n. It is evident that K_n is convex.

If the capacity equations are satisfied, for some C, at a point $(p_1, \dots, p_n)$ with $p_1, \dots, p_n > 0$, then $(p_1, \dots, p_n) \in K_n$ and the gradient of $I(A,B)$ at $(p_1, \dots, p_n)$ is

$$\nabla I(A,B)\big|_{(p_1,\dots,p_n)} = (C - \log e, C - \log e, \dots, C - \log e).$$

That is, the gradient of $I(A,B)$ at $(p_1, \dots, p_n)$ is a scalar multiple of $(1, \dots, 1)$, which is normal to the hyperplane with equation $x_1 + \cdots + x_n = 1$, in $\mathbb{R}^n$, of which K_n is a fragment. It follows that the directional derivative of $I(A,B)$, at $(p_1, \dots, p_n)$, in any direction parallel to this hyperplane, is zero. It follows that the function of one variable obtained by restricting $I(A,B)$ to any line segment in K_n through $(p_1, \dots, p_n)$ will have derivative zero at the value of the one variable corresponding to the point $(p_1, \dots, p_n)$. It follows that $I(A,B)$ achieves its maximum on each such line segment at $(p_1, \dots, p_n)$, by 3.5.4. Therefore, $I(A,B)$ achieves its maximum on K_n at $(p_1, \dots, p_n)$.

Exercises 3.5

1. Suppose that $(\widetilde{p}_1, \dots, \widetilde{p}_{n-1}, 0) \in K_n$, and $\widetilde{p}_1, \dots, \widetilde{p}_{n-1} > 0$. Show that

$$I(A,B)\big|_{(p_1,\dots,p_n)} \to I(A,B)\big|_{(\widetilde{p}_1,\dots,\widetilde{p}_{n-1},0)}$$

as $(p_1, \dots, p_n) \to (\widetilde{p}_1, \dots, \widetilde{p}_{n-1}, 0)$, with $(p_1, \dots, p_n) \in K_n$. [You may assume that, for each $j \in \{1, \dots, k\}$, $q_{ij} > 0$ for some $i \in \{1, \dots, n\}$. (Interpretation?) You may as well inspect the functions $f_{ij}(p_1, \dots, p_n) = p_i q_{ij} \log(\sum_{t=1}^{n} p_t q_{tj})$. No problem when $q_{ij} = 0$, and no problem when $1 \le i \le n-1$. When $i = n$, you will need to consider two cases: $q_{ij} = \cdots = q_{n-1,j} = 0$, and otherwise.]

$*$2. Under what conditions on the transition probabilities is $I(A, B)$ strictly concave on K_n?

3. Prove the statements in 3.5.1 and 3.5.2.