

Foundations of Information Theory

Balaji Dhayabaran

Chapter 1 · Draft v1.0

Foundations of Information Theory

Chapter 1: Uncertainty, Information and Communication (Draft v1.0)

© 2026 Balaji Dhayabaran. Some rights reserved.

This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0). You are free to share this document for non-commercial purposes, provided you give appropriate credit and do not modify it. To view a copy of this licence, visit <https://creativecommons.org/licenses/by-nc-nd/4.0/>.

Teachers and students are welcome to use this material in class.

This is a working draft and is updated regularly. The latest version, together with the lecture slides, is available at <https://kanithakalam.com/foundations-of-information-theory/>.

Contents

1	Uncertainty, Information and Communication	3
1.1	The Measure of Information	3
1.2	The Communication System Model	4
1.3	Likelihood, Uncertainty, and the Logarithm	6
1.4	Average Uncertainty (Entropy)	7
1.5	Two Worked Problems on Entropy	7
1.6	Information Transfer Under Noise	10
1.6.1	How Often Does Each Received Value Occur?	10
1.6.2	What the Receiver Believes After Observing y	11
1.6.3	The Remaining Uncertainty for a Particular Observation	12
1.6.4	Average Remaining Uncertainty (Conditional Entropy)	12
1.6.5	The Information Transferred	13
1.7	Mutual Information	13
1.7.1	The Reverse Question	14
1.7.2	Why the Two Answers Are Always Equal	14
1.7.3	Definition of Mutual Information	15
1.7.4	Example 1.1 Revisited	16
1.8	More Worked Problems	16
1.9	Summary and Looking Ahead	23

Chapter 1

Uncertainty, Information and Communication

What does it mean for a receiver to gain information, and how can we measure it? This chapter answers these questions by following one small communication system from start to finish: a transmitter sending bits, a channel that adds noise, and a receiver trying to work out what was sent. Starting from the single idea that information is a reduction in uncertainty, we will derive, one step at a time, the surprise of an outcome, the entropy of a source, the conditional entropy that remains after an observation, and the mutual information between what is sent and what is received. Each formula is introduced only when the previous step makes it necessary, and each one is checked against the same running example.

The only background needed is basic probability: events, probabilities, independence and averages. Familiarity with digital communication will help with the motivation, but it is not required for the mathematics.

1.1 The Measure of Information

In the context of a communication system, we measure information using the number of bits that can be successfully communicated from a transmitter to a receiver under distortions like noise and channel fading. Let us try to understand the measure of information in this context, as we are all very aware of this system model, and then extend it to a general case.

Information Transferred

Before a transmission begins, the receiver does not know which message will be sent; it only knows how likely each possible message is. When the transmission occurs, the receiver observes a corrupted, noisy version of the bits. Even though this received signal is imperfect, it still provides clues about the original message, which lowers the receiver's overall doubt. Ultimately, the amount of information successfully transferred is exactly equal to this reduction in uncertainty.

We can express this core concept logically as a simple word equation:

$$\text{Information Transferred} = \text{Initial Uncertainty} - \text{Remaining Uncertainty}$$

Whenever we try to measure information, this logical rule will always be our starting point.

1.2 The Communication System Model

Let us consider a transmitter and a receiver. The transmitter is sending information in bits serially, one after the other. Each bit is represented by a voltage for a certain period of time called the symbol time. The bit can be either zero or one. In general, let us say the a-priori probability of transmitting a '1' is q and the probability of a '0' is $(1 - q)$.

When the transmitted bits arrive at the receiver, they are corrupted by noise. The receiver knows the statistics of the source (the probability q) and of the noise, but in any particular symbol time it does not know which bit has actually been transmitted. The receiver is therefore inherently uncertain about the bit.

Let x be the transmitted bit and y be the observed value at the receiver. The observed value at the receiver can be written as:

$$y = x + w \tag{1.1}$$

where w is the additive noise. Figure 1.1 shows this model.

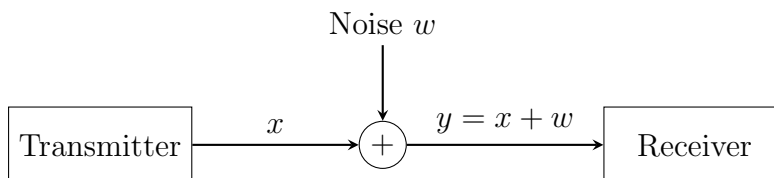


Figure 1.1: The communication system model used throughout this chapter. The transmitted bit x is corrupted by additive noise w , and the receiver observes only $y = x + w$.

From the observed value y , the receiver must decode the value of x . For now, let us assume that a '0' and a '1' are equally likely, i.e. $q = 1/2$. If the receiver is able to make a perfect decision from y , the receiver's uncertainty is completely resolved. In this case,

we say that exactly 1 bit of information has been transmitted. (We will see in Section 1.4 that when ‘0’ and ‘1’ are not equally likely, the receiver’s initial uncertainty is less than 1 bit, and so even a perfect decision transfers less than 1 bit.)

Before we look at the math, there are a few things to keep in mind about the noise w :

- In a real communication system, the noise is usually modelled as Additive White Gaussian Noise (AWGN), which follows a Gaussian PDF and can take any real value.
- In complex baseband processing, this noise is a complex number whose real and imaginary parts are independent Gaussian random variables. In our simple model, the bits are real voltages, so w is real.
- However, to keep things simple and easy to understand right now, we will assume the noise only takes a few specific values, as shown in the examples below. This discrete noise is not Gaussian; it is only a teaching simplification, and we will return to AWGN later.

Example 1.1 (Perfect Decision). Let us assume $w \in \{-1, 1\}$. To decode x , the receiver needs to know which values of y each transmitted bit can produce:

- If $x = 0$, then $y \in \{-1, 1\}$.
- If $x = 1$, then $y \in \{0, 2\}$.

In this case, the two sets are mutually exclusive, so every observed value of y points to exactly one value of x . The receiver can decode x with complete certainty. Since $q = 1/2$, the information being transmitted to the receiver can be measured as 1 bit/transmission. In other words, whenever the receiver observes y , its uncertainty about x is completely resolved.

Example 1.2 (Overlapping Sets - Ambiguity). Now consider a scenario where the noise causes the possible outcomes to overlap. Let us assume $w \in \{0, 1\}$, with both values possible. The received value y now falls into one of two sets:

- If $x = 0$, then $y \in \{0, 1\}$.
- If $x = 1$, then $y \in \{1, 2\}$.

There are four possible (x, w) combinations, but they produce only three distinct values of y :

- **When $y = 0$:** This happens only when $x = 0$ and $w = 0$. Without any uncertainty, the receiver can correctly decide $x = 0$.
- **When $y = 2$:** This happens only when $x = 1$ and $w = 1$. Again, without any uncertainty, the receiver can correctly decide $x = 1$.

- **When $y = 1$:** Here, an ambiguity arises. There are two possibilities: the transmitter could have sent $x = 0$ (with noise $w = 1$), OR it could have sent $x = 1$ (with noise $w = 0$).

When $y = 1$ occurs, the receiver is uncertain about the transmitted value of x . It could be either 0 or 1. Recall our core concept:

Information Transferred = Initial Uncertainty - Remaining Uncertainty

The main takeaway from this example is that the receiver's uncertainty is not completely resolved. Because we must subtract this remaining uncertainty from the initial uncertainty, the total information transmitted is strictly less than the 1 bit per transmission we saw in Example 1.1. We will derive exactly how to mathematically measure this remaining uncertainty in Section 1.6.

1.3 Likelihood, Uncertainty, and the Logarithm

To mathematically formalize this, we must quantify how much doubt is attached to an event. Let p denote the probability of some event (in this section, p is a general probability, not specifically the probability of a '1'). Probability is the likelihood of an event, representing its chance of being true. If an event is highly likely (p is close to 1), its occurrence tells us very little that we did not already expect. Conversely, if an event is rare (p is close to 0), its occurrence is very surprising. Therefore, our doubt about an event should *decrease* as its probability increases. A natural first guess is to measure it using the reciprocal of the probability:

$$\text{Surprise (first guess)} = \frac{1}{p} \quad (1.2)$$

If $p = 0.01$, this factor is 100. If $p = 0.99$, the factor is only 1.01. The smaller the chance of the event, the larger the magnitude of doubt the receiver holds about it occurring.

However, $\frac{1}{p}$ alone is not sufficient to measure information. Consider a transmitter sending two independent bits sequentially. This results in four possible combinations for the transmitted data: $x \in \{00, 01, 10, 11\}$. If the first bit takes a value with probability p_1 and the second bit takes a value with probability p_2 , the joint probability of this specific two-bit sequence occurring is the product of their individual probabilities:

$$p_{\text{joint}} = p_1 \times p_2 \quad (1.3)$$

Consequently, the first guess in Eq. (1.2) also becomes a product:

$$\frac{1}{p_{\text{joint}}} = \frac{1}{p_1} \times \frac{1}{p_2} \quad (1.4)$$

When the receiver decodes these two bits, the total information gained must logically be the sum of the information from each bit ($I_{\text{total}} = I_1 + I_2$). In Example 1.1, where

the transmitter sent only one equally likely bit and the receiver was able to decode it without any uncertainty, the information was 1 bit/transmission. (In Example 1.2, the remaining ambiguity made it strictly less than 1 bit.) If the transmitter now sends two independent, equally likely bits and the receiver decodes both perfectly, the information should be 2 bits/transmission. Clearly, the information and hence the measure of doubt should add up, whereas the probability of each event in x multiplies. The only well-behaved (continuous) function that converts a multiplicative input (probability) into an additive output (information) is the logarithm. Using base 2 allows us to measure this in “bits”. We call this quantity the *surprise* (or self-information) of an event with probability p :

$$S(p) = \log_2 \left(\frac{1}{p} \right) = -\log_2(p) \quad (1.5)$$

This line of reasoning, which leads from the additivity of information to the logarithm, goes back to Claude Shannon, who founded information theory in his 1948 paper [1]. Note that the surprise belongs to a single outcome. The *uncertainty* of a source, which we define next, is a property of all its possible outcomes taken together.

1.4 Average Uncertainty (Entropy)

Let X denote the random variable representing the transmitted symbol, which takes the value x_i with probability p_i ; a lower-case x denotes one particular value of X . To evaluate the overall uncertainty of the source, we must compute the average surprise over all possible outcomes. We multiply the surprise of each possible outcome by the probability of that outcome occurring, and sum it across all outcomes:

$$H(X) = \sum_i p_i \log_2 \left(\frac{1}{p_i} \right) \quad (1.6)$$

This weighted average is the entropy of the source, representing the inherent uncertainty the receiver has about X before any observation is made. It depends only on the source probabilities, not on the channel. For a source that emits independent, identically distributed symbols, it is also the long-run average surprise per symbol.

For the binary source of Section 1.2, $H(X) = q \log_2 \frac{1}{q} + (1-q) \log_2 \frac{1}{1-q}$. When $q = 1/2$, $H(X) = 1$ bit, which is why a perfect decision in Example 1.1 transferred exactly 1 bit. When $q = 0.9$, $H(X) \approx 0.469$ bits, so even a perfect decision transfers only about 0.469 bits, because the receiver already expected a ‘1’ most of the time.

1.5 Two Worked Problems on Entropy

The definitions in the previous sections are short, but they are easy to misread. The two problems in this section address the most common misunderstandings about the entropy formula itself. More worked problems, covering all the ideas of this chapter, are col-

lected in Section 1.8. Try each problem before reading its solution. Throughout, $S(p) = \log_2(1/p)$ is the surprise of an outcome with probability p , and $H(X) = \sum_i p_i \log_2(1/p_i)$ is the entropy of a source X .

Problem 1.1 (Why Not Simply Average $1/p$?). A source emits one of three symbols. For each of the two distributions

$$(a) \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) \quad (b) (0.98, 0.01, 0.01),$$

compute the average of our first guess, $\sum_i p_i \cdot \frac{1}{p_i}$, and the entropy $H(X)$. What do you notice about the average of $1/p$, and why does this make it useless as a measure of uncertainty?

Solution. First, the average of $1/p$. For (a),

$$\frac{1}{3} \cdot 3 + \frac{1}{3} \cdot 3 + \frac{1}{3} \cdot 3 = 3.$$

For (b),

$$0.98 \cdot \frac{1}{0.98} + 0.01 \cdot 100 + 0.01 \cdot 100 = 1 + 1 + 1 = 3.$$

Each term is $p_i \cdot \frac{1}{p_i} = 1$, so the average of $1/p$ is simply the number of possible outcomes:

$$\sum_{i=1}^M p_i \cdot \frac{1}{p_i} = \sum_{i=1}^M 1 = M.$$

The probabilities cancel completely. This quantity counts outcomes; it cannot tell a fair source from a nearly predictable one.

Now compute the entropy. For (a),

$$H(X) = 3 \cdot \frac{1}{3} \log_2 3 = \log_2 3 \approx 1.585 \text{ bits.}$$

For (b),

$$H(X) = 0.98 \log_2 \frac{1}{0.98} + 2 \times 0.01 \log_2 100 \approx 0.0286 + 0.1329 = 0.161 \text{ bits.}$$

Source (b) almost always emits the first symbol, so the receiver can guess correctly 98% of the time before anything is sent. Its uncertainty should be much smaller than that of source (a). Entropy captures this (0.161 bits versus 1.585 bits); the average of $1/p$ does not (3 in both cases).

Takeaway. The average of $1/p$ always equals the number of outcomes, whatever the probabilities are. It fails not only the additivity test of Section 1.3 but also the most basic test: it cannot distinguish a predictable source from an unpredictable one. The logarithm is needed for both reasons.

Problem 1.2 (Rare Events — Large Surprise, Small Contribution). A source emits ‘A’ with probability 0.999 and ‘B’ with probability 0.001.

1. Compute $S(A)$, $S(B)$, and the contribution $p_i S(p_i)$ of each symbol to $H(X)$.
2. Find $\lim_{p \rightarrow 0} p \log_2(1/p)$.
3. What value should we assign to $0 \cdot \log_2(1/0)$, and why does this make physical sense?

Solution.

$$S(B) = \log_2 \frac{1}{0.001} = \log_2 1000 \approx 9.966 \text{ bits}, \quad S(A) = \log_2 \frac{1}{0.999} \approx 0.00144 \text{ bits}.$$

Seeing ‘B’ is a genuinely surprising event: almost 10 bits. Seeing ‘A’ tells us almost nothing.

$$p_A S(A) = 0.999 \times 0.00144 \approx 0.00144 \text{ bits}, \quad p_B S(B) = 0.001 \times 9.966 \approx 0.00997 \text{ bits}.$$

$$H(X) \approx 0.00144 + 0.00997 = 0.0114 \text{ bits}.$$

The surprise of ‘B’ is about 10 bits, but because ‘B’ occurs only once in a thousand symbols, its weight $p_B = 0.001$ shrinks this to about 0.01 bits. The whole source carries only about 0.011 bits per symbol, a tiny fraction of the surprise of its rare symbol. Note that ‘B’ still contributes more than ‘A’ here; the point is not that rare events contribute nothing, but that their large surprise is heavily discounted by how seldom they happen.

To find the limit, substitute $t = 1/p$, so that $t \rightarrow \infty$ as $p \rightarrow 0$:

$$\lim_{p \rightarrow 0} p \log_2 \frac{1}{p} = \lim_{t \rightarrow \infty} \frac{\log_2 t}{t} = 0,$$

because t grows much faster than $\log_2 t$. The surprise grows without bound, but the probability shrinks faster.

The limit above shows that the contribution of an outcome goes smoothly to zero as its probability goes to zero, so defining the contribution of an impossible outcome as zero keeps $H(X)$ continuous. Physically, an outcome that never occurs can never surprise anyone, so it should add nothing to the uncertainty. Equivalently, adding an impossible symbol to a source’s alphabet should not change its entropy.

Takeaway. Surprise belongs to one outcome; entropy is an average. A rare outcome has a large surprise, but it is weighted by its small probability, so its contribution to the entropy tends to zero as its probability tends to zero.

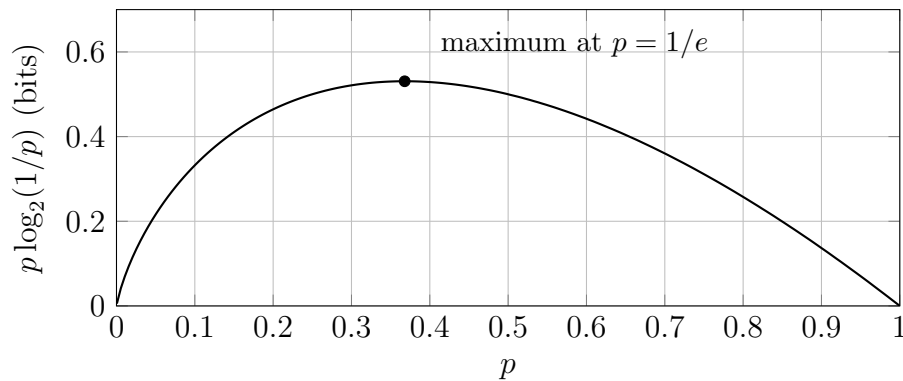


Figure 1.2: Contribution of a single outcome to the entropy as a function of its probability. It vanishes for impossible ($p = 0$) and certain ($p = 1$) outcomes and is largest, about 0.531 bits, at $p = 1/e \approx 0.368$.

1.6 Information Transfer Under Noise

We now return to Example 1.2, where the noise made the received sets overlap. There we saw that the receiver's uncertainty is not completely resolved, and so the information transferred must be less than 1 bit per transmission. We now have the tools to find out exactly how much less. Recall our core concept:

Information Transferred = Initial Uncertainty - Remaining Uncertainty

The initial uncertainty is the entropy of the source, $H(X)$, which we measured in Section 1.4. What remains is to measure the uncertainty that is left at the receiver after it observes y . As in Example 1.2, the received value is

$$y = x + w, \quad x \in \{0, 1\}, \quad w \in \{0, 1\}.$$

To calculate anything, we must also say how likely each value is. Let us assume that the source is fair, $P(x = 0) = P(x = 1) = \frac{1}{2}$, so that $H(X) = 1$ bit, and that the two noise values are equally likely, $P(w = 0) = P(w = 1) = \frac{1}{2}$. The noise is produced by the channel and has nothing to do with the bit being sent, so we assume that x and w are independent.

1.6.1 How Often Does Each Received Value Occur?

There are four possible combinations of x and w . Since x and w are independent, the probability of each combination is the product of the individual probabilities:

$$P(x, w) = P(x) P(w) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}.$$

The table below lists the combinations, the value of y each one produces, and its probability.

x	w	$y = x + w$	$P(x, w)$
0	0	0	$\frac{1}{4}$
0	1	1	$\frac{1}{4}$
1	0	1	$\frac{1}{4}$
1	1	2	$\frac{1}{4}$

The value $y = 0$ is produced by only one combination, and so is $y = 2$. The value $y = 1$ is produced by two combinations, so we add their probabilities:

$$P(y = 0) = \frac{1}{4}, \quad P(y = 1) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}, \quad P(y = 2) = \frac{1}{4}. \quad (1.7)$$

1.6.2 What the Receiver Believes After Observing y

Before anything is transmitted, the receiver's knowledge about x is the a-priori distribution $(\frac{1}{2}, \frac{1}{2})$. Once it observes a particular value of y , its knowledge changes. The question the receiver must now ask is: *given the value of y I have observed, how likely is each value of x ?*

Suppose the receiver observes $y = 1$. From the table, the total probability that leads to $y = 1$ is $\frac{1}{2}$, made up of $\frac{1}{4}$ from $x = 0$ and $\frac{1}{4}$ from $x = 1$. The chance that $x = 0$ was sent is the share of this total that came from $x = 0$:

$$P(x = 0 \mid y = 1) = \frac{1/4}{1/2} = \frac{1}{2}, \quad P(x = 1 \mid y = 1) = \frac{1/4}{1/2} = \frac{1}{2}.$$

Observing $y = 1$ leaves the receiver exactly where it started. Both explanations are equally likely, so the observation does not favour either value of x .

Now suppose the receiver observes $y = 0$. All the probability that leads to $y = 0$ comes from $x = 0$, so

$$P(x = 0 \mid y = 0) = \frac{1/4}{1/4} = 1, \quad P(x = 1 \mid y = 0) = 0.$$

In the same way, observing $y = 2$ gives

$$P(x = 0 \mid y = 2) = 0, \quad P(x = 1 \mid y = 2) = 1.$$

In every case, we divided the probability of this particular x and y occurring together by the total probability of this y . Writing this in symbols gives

$$P(x \mid y) = \frac{P(x, y)}{P(y)} = \frac{P(x) P(y \mid x)}{P(y)}, \quad (1.8)$$

where $P(y \mid x)$ is simply the probability of the noise value that turns x into y , that is, $P(w = y - x)$. Equation (1.8) is known as *Bayes' rule*. We did not need to memorise it: it follows directly from counting the ways in which each received value can occur. The

distribution $P(x | y)$ is called the *a-posteriori* distribution of x : it describes what the receiver believes about x after observing y .

1.6.3 The Remaining Uncertainty for a Particular Observation

After observing y , the receiver's knowledge about x is described by a new probability distribution, $P(x | y)$. We already know how to measure the uncertainty of a probability distribution: it is its entropy, the average surprise of Section 1.4. The only change is that the probabilities are now the updated ones. The remaining uncertainty after observing a particular value y is therefore

$$H(X | Y = y) = \sum_x P(x | y) \log_2 \left(\frac{1}{P(x | y)} \right). \quad (1.9)$$

Applying this to each received value, and using the convention $0 \cdot \log_2(1/0) = 0$ from Problem 1.2,

$$\begin{aligned} H(X | Y = 0) &= 1 \cdot \log_2 1 + 0 = 0 \text{ bits}, \\ H(X | Y = 1) &= \tfrac{1}{2} \log_2 2 + \tfrac{1}{2} \log_2 2 = 1 \text{ bit}, \\ H(X | Y = 2) &= 0 + 1 \cdot \log_2 1 = 0 \text{ bits}. \end{aligned}$$

This is exactly what we found intuitively in Example 1.2. Observing $y = 0$ or $y = 2$ settles the question completely, while observing $y = 1$ leaves the receiver with the full 1 bit of doubt it started with.

1.6.4 Average Remaining Uncertainty (Conditional Entropy)

The remaining uncertainty depends on which value of y the receiver happens to observe, and the receiver has no control over this. To describe the channel by a single number, we face the same situation as in Section 1.4, where the surprise varied from outcome to outcome, and we resolve it in the same way: we take the probability-weighted average. Each remaining uncertainty is weighted by how often the corresponding value of y occurs:

$$H(X | Y) = \sum_y P(y) H(X | Y = y). \quad (1.10)$$

For our channel,

$$H(X | Y) = \tfrac{1}{4} \cdot 0 + \tfrac{1}{2} \cdot 1 + \tfrac{1}{4} \cdot 0 = \tfrac{1}{2} \text{ bit}.$$

This average is called the *conditional entropy* of X given Y . It is the uncertainty about x that remains, on average, after the receiver observes y . We will meet a quantity of this kind again in Problem 1.8, where the second of two related bits contributes only the uncertainty that remains once the first bit is known.

Substituting Equation (1.9) into Equation (1.10) gives

$$H(X | Y) = \sum_y P(y) \sum_x P(x | y) \log_2 \left(\frac{1}{P(x | y)} \right).$$

From Equation (1.8), $P(y) P(x | y) = P(x, y)$, so the two sums combine into a single sum over all pairs (x, y) :

$$H(X | Y) = \sum_{x,y} P(x, y) \log_2 \left(\frac{1}{P(x | y)} \right). \quad (1.11)$$

This is the form in which conditional entropy usually appears in textbooks. It is worth placing it next to the formula for entropy:

$$\begin{aligned} H(X) &= \sum_x P(x) \log_2 \left(\frac{1}{P(x)} \right), \\ H(X | Y) &= \sum_{x,y} P(x, y) \log_2 \left(\frac{1}{P(x | y)} \right). \end{aligned}$$

Both formulas have the same structure: each is an average of surprise. In the first, the surprise about x is measured before the receiver observes anything. In the second, it is measured after the receiver observes y , and then averaged over everything that can happen. Just as entropy is the average uncertainty of the source, conditional entropy is the average uncertainty that the channel leaves behind.

1.6.5 The Information Transferred

We now have both quantities in our word equation. The initial uncertainty is $H(X) = 1$ bit, and the remaining uncertainty is $H(X | Y) = \frac{1}{2}$ bit. Therefore,

$$\text{Information Transferred} = H(X) - H(X | Y) = 1 - \frac{1}{2} = \frac{1}{2} \text{ bit per transmission.} \quad (1.12)$$

This answer is consistent with our intuition. It is more than zero, because half of the time ($y = 0$ or $y = 2$) the receiver learns x exactly. It is less than 1 bit, because the other half of the time ($y = 1$) the receiver learns nothing about x . On average, the channel delivers half of the bit that was sent.

1.7 Mutual Information

In Section 1.6 we measured the information transferred from the point of view of the receiver: how much does observing y reduce the uncertainty about x ? Let us give this quantity a symbol,

$$I(X; Y) = H(X) - H(X | Y) = 1 - \frac{1}{2} = \frac{1}{2} \text{ bit.} \quad (1.13)$$

We can also ask the reverse question: how much does knowing x reduce the uncertainty about y ? At first sight this seems to be a completely different question. Let us answer it for Example 1.2 using exactly the same steps as in Section 1.6, with the roles of x and y interchanged.

1.7.1 The Reverse Question

Before x is known, the uncertainty about y is the entropy of Y . We found in Section 1.6 that y takes the values 0, 1 and 2 with probabilities $\frac{1}{4}$, $\frac{1}{2}$ and $\frac{1}{4}$. Therefore,

$$H(Y) = \frac{1}{4} \log_2 4 + \frac{1}{2} \log_2 2 + \frac{1}{4} \log_2 4 = 0.5 + 0.5 + 0.5 = 1.5 \text{ bits.}$$

Now suppose x is known. If $x = 0$, then $y \in \{0, 1\}$, each value with probability $\frac{1}{2}$, so $H(Y | X = 0) = 1$ bit. If $x = 1$, then $y \in \{1, 2\}$, again each with probability $\frac{1}{2}$, so $H(Y | X = 1) = 1$ bit. Averaging over x , exactly as we averaged over y in Equation (1.10),

$$H(Y | X) = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 1 = 1 \text{ bit.}$$

The reduction in the uncertainty about y is therefore

$$I(Y; X) = H(Y) - H(Y | X) = 1.5 - 1 = \frac{1}{2} \text{ bit,} \quad (1.14)$$

which is exactly the value we obtained in Equation (1.13).

This result has a simple physical interpretation. Once x is known, the only thing that can still change y is the noise, so the remaining uncertainty about y is just the uncertainty of the noise: $H(Y | X) = H(W) = 1$ bit. The received signal varies by 1.5 bits in total; 1 bit of this variation is due to the noise, and the remaining $\frac{1}{2}$ bit is the part that carries information about x . This interpretation holds for any channel in which the noise is simply added to the transmitted signal.

1.7.2 Why the Two Answers Are Always Equal

One example can only illustrate the equality; it cannot prove it. To show that it holds for every channel, we first write $H(X)$ as a sum over pairs (x, y) . Since adding $P(x, y)$ over all values of y gives $P(x)$,

$$H(X) = \sum_x P(x) \log_2 \left(\frac{1}{P(x)} \right) = \sum_{x,y} P(x, y) \log_2 \left(\frac{1}{P(x)} \right).$$

Subtracting Equation (1.11) for $H(X | Y)$, which is a sum over the same pairs,

$$I(X; Y) = \sum_{x,y} P(x, y) \left[\log_2 \left(\frac{1}{P(x)} \right) - \log_2 \left(\frac{1}{P(x | y)} \right) \right] = \sum_{x,y} P(x, y) \log_2 \left(\frac{P(x | y)}{P(x)} \right).$$

Finally, Bayes' rule, Equation (1.8), gives $P(x | y) = P(x, y)/P(y)$, so

$$I(X; Y) = \sum_{x,y} P(x, y) \log_2 \left(\frac{P(x, y)}{P(x) P(y)} \right). \quad (1.15)$$

In Equation (1.15), x and y appear in exactly the same way: interchanging them changes nothing. Starting instead from $H(Y) - H(Y | X)$ and following the same steps leads to the same expression. Therefore, for every channel,

$$I(X; Y) = I(Y; X).$$

We can check Equation (1.15) on Example 1.2. Only the four pairs that can actually occur contribute to the sum:

(x, y)	$P(x, y)$	$P(x) P(y)$	Contribution
$(0, 0)$	$\frac{1}{4}$	$\frac{1}{2} \cdot \frac{1}{4} = \frac{1}{8}$	$\frac{1}{4} \log_2 2 = \frac{1}{4}$
$(0, 1)$	$\frac{1}{4}$	$\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$	$\frac{1}{4} \log_2 1 = 0$
$(1, 1)$	$\frac{1}{4}$	$\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$	$\frac{1}{4} \log_2 1 = 0$
$(1, 2)$	$\frac{1}{4}$	$\frac{1}{2} \cdot \frac{1}{4} = \frac{1}{8}$	$\frac{1}{4} \log_2 2 = \frac{1}{4}$

The total is $\frac{1}{2}$ bit, as before. The table also shows where the information comes from. The pairs with $y = 1$ contribute nothing, because for them $P(x, y) = P(x) P(y)$: observing $y = 1$ tells the receiver nothing about x . All of the information comes from the observations $y = 0$ and $y = 2$, which is exactly what we found in Section 1.6, now seen from a different angle.

1.7.3 Definition of Mutual Information

Since the reduction in uncertainty is the same in both directions, it does not belong to x alone or to y alone; it is shared by the two. For this reason it is called the *mutual information* between X and Y :

$$\begin{aligned} I(X; Y) &= H(X) - H(X | Y) = H(Y) - H(Y | X) \\ &= \sum_{x,y} P(x, y) \log_2 \left(\frac{P(x, y)}{P(x) P(y)} \right). \end{aligned} \quad (1.16)$$

Mutual information is the precise form of the word equation with which we began this chapter: it is the initial uncertainty minus the remaining uncertainty, and it measures the information transferred over the channel.

Equation (1.15) also tells us when no information is transferred at all. If x and y are independent, then $P(x, y) = P(x) P(y)$ for every pair, every logarithm in the sum is $\log_2 1 = 0$, and $I(X; Y) = 0$. Information can be transferred only when what is received depends on what was sent.

1.7.4 Example 1.1 Revisited

We can now confirm, using the formulas, the claim we made intuitively in Section 1.2: that the channel of Example 1.1 transfers exactly 1 bit per transmission. As before, let $P(x = 0) = P(x = 1) = \frac{1}{2}$, and let the two noise values $w \in \{-1, 1\}$ be equally likely and independent of x . The four combinations of x and w give four different received values, $y \in \{-1, 0, 1, 2\}$, each with probability $\frac{1}{4}$.

Every received value is produced by exactly one value of x , so after any observation the a-posteriori distribution of x is either $(1, 0)$ or $(0, 1)$. Hence $H(X | Y = y) = 0$ for every y , and therefore $H(X | Y) = 0$. The mutual information is

$$I(X; Y) = H(X) - H(X | Y) = 1 - 0 = 1 \text{ bit.}$$

From the other side, y takes four equally likely values, so $H(Y) = \log_2 4 = 2$ bits, and once x is known only the noise remains, so $H(Y | X) = H(W) = 1$ bit. Again,

$$I(Y; X) = H(Y) - H(Y | X) = 2 - 1 = 1 \text{ bit.}$$

When the received sets do not overlap, the channel removes all of the receiver's uncertainty, and the full 1 bit of the source is transferred.

1.8 More Worked Problems

The problems in this section test the ideas of the whole chapter. Several of them address misunderstandings about entropy that are easy to fall into; the later ones connect entropy to conditional entropy and to practical system design. As before, try each problem before reading its solution.

Problem 1.3 (Certainty Means Zero Entropy). A faulty transmitter is stuck and always sends '1'. Compute $H(X)$. Now use this source in Example 1.1 (perfect decision, $w \in \{-1, 1\}$). How much information does the receiver gain per transmission, even though it decodes every bit correctly? Relate your answer to the word equation in Section 1.1.

Solution. The source has $q = P(x = 1) = 1$ and $P(x = 0) = 0$. Using the convention from Problem 1.2,

$$H(X) = 1 \cdot \log_2 1 + 0 \cdot \log_2 \frac{1}{0} = 0 + 0 = 0 \text{ bits.}$$

The receiver knows the source statistics. Since $q = 1$, it knows that the next bit will be '1' before anything is transmitted. Its initial uncertainty is $H(X) = 0$.

In Example 1.1 the received sets for $x = 0$ and $x = 1$ do not overlap, so after observing y the receiver has no doubt left. The remaining uncertainty is also 0.

$$\text{Information Transferred} = \underbrace{0}_{\text{initial}} - \underbrace{0}_{\text{remaining}} = 0 \text{ bits per transmission.}$$

Correct decoding is not the same as receiving information. Every bit is decoded correctly, yet the receiver learns nothing it did not already know. A perfect channel cannot create information that the source does not have.

Takeaway. Information is a reduction of uncertainty. If there is no uncertainty to begin with, there is nothing to reduce, no matter how good the channel is.

Problem 1.4 (A Binary Digit Is Not a Bit of Information). A source with $q = 0.9$ produces a file of 10,000 independent binary digits. How many bits of information does the file contain? Write one sentence distinguishing a “binary digit” from a “bit” as the unit of information.

Solution. The entropy of each binary digit is

$$H(X) = 0.9 \log_2 \frac{1}{0.9} + 0.1 \log_2 \frac{1}{0.1} \approx 0.9 \times 0.152 + 0.1 \times 3.322 \approx 0.469 \text{ bits.}$$

The digits are independent, so the information adds up (Section 1.3):

$$10,000 \times 0.469 \approx 4,690 \text{ bits.}$$

A *binary digit* is a physical symbol, a 0 or a 1; a *bit* is a unit of information, the amount resolved by one yes/no question whose two answers are equally likely. A binary digit carries at most 1 bit of information, and exactly 1 bit only when 0 and 1 are equally likely and independent of the other digits.

Takeaway. The amount of information depends on the probabilities of the transmitted symbols, not on how many binary digits are used to write them down.

Problem 1.5 (Entropy as the Number of Yes/No Questions). A source emits A, B, C, D with probabilities $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8})$. Design a strategy of yes/no questions to identify the symbol, and compute the average number of questions your strategy needs. Compare it with $H(X)$. Repeat for the uniform source $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$. Why does the first source need fewer questions on average?

Solution. Ask about the most likely symbol first:

1. “Is it A?” If yes, stop (1 question).
2. “Is it B?” If yes, stop (2 questions).
3. “Is it C?” Yes means C, no means D (3 questions either way).

$$\frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 3 + \frac{1}{8} \cdot 3 = 0.5 + 0.5 + 0.375 + 0.375 = 1.75.$$

$$H(X) = \frac{1}{2} \log_2 2 + \frac{1}{4} \log_2 4 + 2 \times \frac{1}{8} \log_2 8 = 0.5 + 0.5 + 0.75 = 1.75 \text{ bits.}$$

The two agree exactly. Moreover, the number of questions needed for each symbol equals its surprise: $S(\frac{1}{2}) = 1$, $S(\frac{1}{4}) = 2$, $S(\frac{1}{8}) = 3$.

For the uniform source, the best strategy splits the symbols in half each time: “Is it A or B?”, then one more question. Every symbol needs 2 questions, so the average is 2, which equals $H(X) = \log_2 4 = 2$ bits. Using the first strategy (“Is it A?” first) on this source would be worse:

$$\frac{1}{4} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{4} \cdot 3 + \frac{1}{4} \cdot 3 = 2.25 > 2.$$

A good question is one whose two answers are equally likely; such a question resolves exactly 1 bit. In both strategies above, every question splits the remaining probability in half. The non-uniform source needs fewer questions because it is more predictable: half the time, a single question is enough.

These probabilities are powers of $\frac{1}{2}$, so the match is exact. For general probabilities, the number of questions must be a whole number, and the best strategy needs on average at least $H(X)$ questions and less than $H(X) + 1$. By asking questions about long blocks of symbols, the average per symbol can be brought as close to $H(X)$ as we like.

Takeaway. Entropy is the minimum average number of well-chosen yes/no questions needed to identify the symbol. This gives the unit “bit” a concrete meaning.

Problem 1.6 (The Formula Agrees with Step-by-Step Decisions). A source emits A with probability $\frac{1}{2}$ and B, C with probability $\frac{1}{4}$ each. Compute $H(X)$ directly. Then compute it in two stages: first decide “A or not-A”, then, only when needed, decide “B or C”. Add the uncertainties, weighting each stage by how often it happens. Do the two answers agree?

Solution. Computed directly,

$$H(X) = \frac{1}{2} \log_2 2 + \frac{1}{4} \log_2 4 + \frac{1}{4} \log_2 4 = 0.5 + 0.5 + 0.5 = 1.5 \text{ bits.}$$

In the first stage we decide “A or not-A”. The two answers have probabilities $\frac{1}{2}$ and $\frac{1}{2}$, so this stage has uncertainty

$$H\left(\frac{1}{2}, \frac{1}{2}\right) = 1 \text{ bit.}$$

This stage always happens.

In the second stage we decide “B or C”. This stage happens only when the answer to the first stage is “not-A”, which has probability $\frac{1}{2}$. Given “not-A”, B and C are equally likely, so this stage has uncertainty $H\left(\frac{1}{2}, \frac{1}{2}\right) = 1$ bit.

Combining the two stages,

$$\underbrace{1}_{\text{Stage 1}} + \underbrace{\frac{1}{2}}_{\text{how often}} \times \underbrace{1}_{\text{Stage 2}} = 1.5 \text{ bits},$$

which agrees with the direct computation.

The same holds for any probabilities p_1, p_2, p_3 :

$$H(p_1, p_2, p_3) = H(p_1, p_2 + p_3) + (p_2 + p_3) H\left(\frac{p_2}{p_2 + p_3}, \frac{p_3}{p_2 + p_3}\right).$$

Shannon [1] required this *grouping* property: the uncertainty of a choice should not depend on whether we make it all at once or in stages. He showed that, together with continuity and the requirement that uncertainty grows with the number of equally likely outcomes, it forces the formula $H(X) = \sum_i p_i \log_2(1/p_i)$, up to the choice of the base of the logarithm.

Takeaway. The entropy formula is not an arbitrary choice. It is the only formula that gives the same answer whether a decision is made in one step or broken into smaller decisions.

- Problem 1.7** (Fractional Bits and the Uniform Maximum). 1. Compute $H(X)$ for a uniform three-symbol source. What does a non-integer number of bits mean? (Hint: consider sending a block of 5 symbols, which has $3^5 = 243$ possible combinations.)
2. For a binary source, plot $H(X)$ against q from 0 to 1. Where is the maximum, and why is the curve symmetric?
3. Show that no four-symbol source can have more than 2 bits of entropy.

Solution. Part 1.

$$H(X) = \log_2 3 \approx 1.585 \text{ bits per symbol.}$$

One symbol has 3 possibilities, so we need 2 binary digits ($2^2 = 4 \geq 3$), wasting one of the four patterns. That is 2 binary digits per symbol.

A block of 5 symbols has $3^5 = 243$ possibilities. Since $2^8 = 256 \geq 243$, 8 binary digits are enough, giving $8/5 = 1.6$ binary digits per symbol. Longer blocks waste even less:

Block length n	Combinations 3^n	Binary digits needed	Per symbol
1	3	2	2.000
3	27	5	1.667
5	243	8	1.600
17	129,140,163	27	1.588
41	$\approx 3.65 \times 10^{19}$	65	1.585

As the blocks grow, the cost per symbol approaches $\log_2 3 \approx 1.585$. A fractional number of bits is an average rate: no single symbol uses 1.585 binary digits, but a long sequence of symbols needs 1.585 binary digits per symbol.

Part 2. With $P(x = 1) = q$,

$$H(q) = q \log_2 \frac{1}{q} + (1 - q) \log_2 \frac{1}{1 - q}.$$

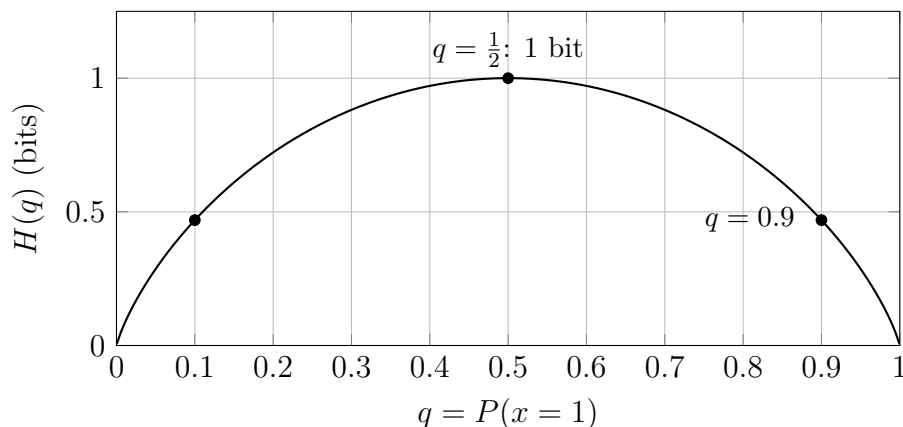


Figure 1.3: The binary entropy function. It is zero for a certain source ($q = 0$ or $q = 1$) and largest, 1 bit, for a fair source ($q = \frac{1}{2}$).

The maximum is 1 bit at $q = \frac{1}{2}$, where the receiver has no reason to prefer either symbol. The curve is symmetric because $H(q) = H(1 - q)$: replacing q by $1 - q$ is the same as swapping the labels ‘0’ and ‘1’, which only reorders the terms of the sum and so cannot change the entropy.

Part 3.

Distribution	$H(X)$ (bits)
(0.25, 0.25, 0.25, 0.25)	2.000
(0.4, 0.3, 0.2, 0.1)	1.846
(0.5, 0.25, 0.125, 0.125)	1.750
(0.7, 0.1, 0.1, 0.1)	1.357

We use the inequality $\ln z \leq z - 1$ for $z > 0$, with equality only at $z = 1$. For M symbols, summing over the symbols with $p_i > 0$,

$$H(X) - \log_2 M = \sum_i p_i \log_2 \frac{1}{p_i} - \sum_i p_i \log_2 M = \sum_i p_i \log_2 \frac{1}{Mp_i}.$$

Applying the inequality with $z = \frac{1}{Mp_i}$ and $\log_2 z = \frac{\ln z}{\ln 2}$,

$$H(X) - \log_2 M \leq \frac{1}{\ln 2} \sum_i p_i \left(\frac{1}{Mp_i} - 1 \right) = \frac{1}{\ln 2} \left(\sum_i \frac{1}{M} - 1 \right) \leq 0,$$

because there are at most M terms in the last sum. Hence $H(X) \leq \log_2 M$, with equality only when every $p_i = \frac{1}{M}$. For $M = 4$, $H(X) \leq 2$ bits.

Takeaway. Entropy can be fractional because it is an average rate over long sequences. For M symbols it is largest, $\log_2 M$ bits, when all symbols are equally likely; any imbalance makes the source more predictable and lowers its entropy.

Problem 1.8 (Additivity Needs Independence). A transmitter sends two bits. The first is equally likely to be 0 or 1, and the second is always a copy of the first. List the possible two-bit sequences and their probabilities, and compute the joint entropy. Repeat for a second bit that equals the first with probability 0.9. Compare both results with 2 bits. What exactly did Section 1.3 assume when it said that information “should add up”?

Solution. If the second bit is always a copy of the first, the possible sequences are:

Sequence	Probability
00	$\frac{1}{2}$
01	0
10	0
11	$\frac{1}{2}$

$$H = \frac{1}{2} \log_2 2 + \frac{1}{2} \log_2 2 = 1 \text{ bit.}$$

Once the first bit is known, the second tells the receiver nothing new.

If the second bit equals the first with probability 0.9:

Sequence	Probability
00	$0.5 \times 0.9 = 0.45$
01	$0.5 \times 0.1 = 0.05$
10	$0.5 \times 0.1 = 0.05$
11	$0.5 \times 0.9 = 0.45$

$$H = 2 \times 0.45 \log_2 \frac{1}{0.45} + 2 \times 0.05 \log_2 \frac{1}{0.05} \approx 1.037 + 0.432 = 1.469 \text{ bits.}$$

Notice that $1.469 = 1 + 0.469$: the first bit contributes 1 bit, and the second contributes only the uncertainty that remains about it once the first bit is known, which is $H(0.9) \approx 0.469$ bits (compare Problem 1.4). In general,

$$H(\text{bit 1, bit 2}) = H(\text{bit 1}) + H(\text{bit 2} \mid \text{bit 1}),$$

where the last term is the *conditional entropy*, the remaining uncertainty about bit 2 after bit 1 is known.

Section 1.3 wrote $p_{\text{joint}} = p_1 \times p_2$. This product rule holds only when the two bits are *independent*. Only then does knowing one bit leave the uncertainty about the other unchanged, so that the information adds up to $1 + 1 = 2$ bits.

Takeaway. Information adds up only for independent symbols. When symbols are related, each new symbol contributes only its remaining uncertainty. This remaining uncertainty is exactly the conditional entropy of Section 1.6.

Problem 1.9 (A Link to Practice — Shaped Constellations). A 4-PAM transmitter uses amplitudes $\{-3, -1, +1, +3\}$. Compute $H(X)$ and the average symbol energy $E = \sum_i p_i x_i^2$ for

- (a) uniform probabilities, (b) probabilities (0.1, 0.4, 0.4, 0.1).

What do you give up, and what do you gain?

Solution. For uniform probabilities,

$$H(X) = \log_2 4 = 2 \text{ bits}, \quad E = \frac{1}{4}(9 + 1 + 1 + 9) = 5.$$

For the shaped case, the inner points (± 1) are used more often than the outer points (± 3):

$$H(X) = 2 \times 0.1 \log_2 10 + 2 \times 0.4 \log_2 2.5 \approx 0.664 + 1.058 = 1.722 \text{ bits},$$

$$E = 2 \times 0.1 \times 9 + 2 \times 0.4 \times 1 = 1.8 + 0.8 = 2.6.$$

We give up about 0.28 bits per symbol of entropy (from 2 to 1.722 bits), and the average energy drops from 5 to 2.6, a reduction of $10 \log_{10}(5/2.6) \approx 2.84$ dB. Equivalently, for the same average energy as the uniform case, the shaped constellation can be scaled up by a factor $\sqrt{5/2.6} \approx 1.39$, spreading its points further apart and making them easier to distinguish in noise.

The entropy formula uses only the probabilities, never the amplitudes themselves. Here this is put to use: by changing only the probabilities, we change both the entropy and the energy, without changing the amplitudes at all. Whether the trade is worthwhile depends on how much information actually survives the noisy channel, which is measured by the mutual information of Section 1.7. Choosing the input probabilities to make this as large as possible is the subject of the next chapter, on channel capacity. For the AWGN channel, it turns out that carefully shaped inputs can do better than uniform ones; this idea, called *probabilistic constellation shaping*, is used in modern optical and wireless systems.

Takeaway. The number of constellation points sets only an upper limit, $\log_2 M$ bits, on the information per symbol. The actual entropy is set by how often each point is used, and choosing those probabilities is a design decision that trades information per symbol against energy.

1.9 Summary and Looking Ahead

This chapter began with a single idea: the information transferred over a channel is the reduction in the receiver's uncertainty. Every quantity we defined was needed to make that idea precise.

Quantity	Formula	Meaning
Surprise	$S(p) = \log_2 \frac{1}{p}$	Surprise of one outcome with probability p
Entropy	$H(X) = \sum_x P(x) \log_2 \frac{1}{P(x)}$	Uncertainty about x before any observation
Remaining uncertainty	$H(X Y = y) = \sum_x P(x y) \log_2 \frac{1}{P(x y)}$	Uncertainty about x after observing a particular y
Conditional entropy	$H(X Y) = \sum_y P(y) H(X Y = y)$	Average uncertainty about x left after observing y
Mutual information	$I(X; Y) = H(X) - H(X Y)$	Information transferred; also equal to $H(Y) - H(Y X)$

For our two running examples, the channel of Example 1.1 transfers $I(X; Y) = 1$ bit per transmission, and the overlapping channel of Example 1.2 transfers $I(X; Y) = \frac{1}{2}$ bit per transmission.

Throughout this chapter we assumed that the source is fair, $P(x = 1) = \frac{1}{2}$. This was a choice, not a necessity. If the transmitter sends a '1' with probability q over the channel of Example 1.2, the same steps as in Section 1.6 show that the information transferred becomes $I(X; Y) = \frac{1}{2}H(q)$, where $H(q)$ is the binary entropy function. The amount of information that gets through therefore depends not only on the channel but also on how the transmitter uses it. For this channel, it is largest when $q = \frac{1}{2}$, where it equals $\frac{1}{2}$ bit per transmission.

This raises the central question of the next chapter: for a given channel, what is the largest amount of information that can be transferred per transmission, over all the ways in which the transmitter can choose its input probabilities? This largest value is called the *capacity* of the channel.

Bibliography

- [1] C. E. Shannon, “A mathematical theory of communication,” *Bell System Technical Journal*, vol. 27, pp. 379–423 and 623–656, 1948.