Appendix F
Measures of Secrecy and Security

William Stallings

F.1 PERFECT SECRECY 2
F.2 INFORMATION AND ENTROPY 8
Information 8
Entropy 10
Properties of the Entropy Function 12
Conditional Entropy 13
F.3 ENTROPY AND SECRECY 14

Supplement to
Cryptography and Network Security, Fifth Edition
William Stallings
Prentice Hall 2010
ISBN-10: 0136097049
http://williamstallings.com/Crypto/Crypto5e.html

In this appendix, we look at measures of secrecy and security of cryptosystems from two different points of view. First, we use concepts of conditional probability, and then reformulate the results obtained in terms of entropy, which in turn depends on concepts of condition probability. The reader should first review the discussion of conditional probability and Bayes' theorem in Appendix 20A.

All of the concepts in this Appendix were first introduced in Shannon's landmark 1949 paper [SHAN49], which is included in the Document section of this book's Web site.

F.1 PERFECT SECRECY[1]

What does it mean that a crypto system is secure? Of course, if the adversary finds the entire plaintext or the entire secret key, that would be a severe failure. But even if the adversary finds a small part of the plaintext or the key, or even if the adversary determines that, say, the first letter of the plaintext is more likely to be an A than the usual frequency of an A at the beginning of a word in a typical English text, that would also be a weakness.

Idea: A cryptosystem is secure to an attack if the adversary does not learn anything after the attack compared to what he knew before the attack. In this section, we consider the case of the ciphertext-only attack. The other types of attacks can be formalized similarly. We define two types of secrecy:

- **Perfect secrecy:** the adversary does not learn anything, no matter her computational power and how much time the attack takes. This is the ideal, but cannot be realized by practical cryptosystems.
- **Computational secrecy:** the adversary does not learn anything unless she is performing more than N operations, where N is some huge number (so that the attack takes thousands of years). This is good enough and may be achieved by practical cryptosystems.

[1] This section is based on notes provided by Marius Zimand of Towson State University.

To formally define the notion of secrecy we first, we introduce some notation:

- M is a random variable that denotes a message chosen from the set of messages $\mathcal{M}$. M is characterized by its distribution (see example below).
- K is a random variable that denotes the encryption key chosen from the set of keys $\mathcal{K}$. The key K is chosen uniformly at random (i.e., all the keys are equally likely).
- C is the encryption of M, i.e., $C = \mathrm{E}(K, M)$

Simple example: Suppose the message comes from a military base. To keep things simple, let us assume that the base sends only three messages: "nothing to report," "attack with 5 planes" and "attack with 10 planes.". Then

$\mathcal{M}$ = {"nothing to report," "attack with 5 planes," "attack with 10 planes"}

This is called the *set of messages*. We can endow a set of messages with a probability distribution (in short, just *distribution*), indicating how likely each message is. For example, one possible distribution can be

$$M = \begin{pmatrix} \text{nothing to report} & \text{attack with 5 planes} & \text{attack with 10 planes} \\ 0.6 & 0.3 & 0.1 \end{pmatrix}$$

We should assume that the attacker knows the distribution M (similar to knowing the frequency of letters in English).

We are now in a position to formally define the term **perfect secrecy**, or **perfect security**. Before doing so, it is instructive to quote Shannon's description.

> “Perfect Secrecy” is defined by requiring of a system that after a cryptogram is intercepted by the enemy the a posteriori probabilities of this cryptogram representing various messages be identically the same as the a priori probabilities of the same messages before the interception. It is shown that perfect secrecy is possible but requires, if the number of messages is finite, the same number of

possible keys. If the message is thought of as being constantly generated at a given "rate" R, key must be generated at the same or a greater rate.

We develop two different versions or the definition of perfect secrecy..

Definition 1. An encryption scheme over message space $\mathcal{M}$ is **perfectly secure- version 1** if for all distributions M over $\mathcal{M}$, for any fixed message m and for any fixed ciphertext c, we have

$$\Pr[M = m \mid \mathrm{E}(K, M) = c] = \Pr[M = m]$$

Here the probabilities are taken over the distribution M and over choosing the key K uniformly at random in the space of all keys. We can make the following observations.

1. The definition is equivalent to saying that M and $\mathrm{E}(K, M)$ are independent.
2. What the definition is saying: The distribution on M is supposed to be known by the adversary. We just want that the cryptosystem does not leak any additional information. This is captured in the definition by saying that knowing the ciphertext c does not change the distribution M.
3. We have argued intuitively (Section 2.2) that the one-time pad has the above property. Now we can prove this assertion rigorously.

Theorem 1. A one-time pad is perfectly secure.

Proof of a special case (the general case is similar): Let $\mathcal{M} = \{0; 1\}$ - just two messages. Let us denote $C = \mathrm{E}(K, M) = \mathrm{K} \oplus \mathrm{M}$. We first show that

$$\Pr\left[M = 0 \mid C = 0\right] = \frac{\Pr\left[M = 0 \cap C = 0\right]}{\Pr\left[C = 0\right]} = \frac{\Pr\left[M = 0 \cap M \oplus K = 0\right]}{\Pr\left[C = 0\right]}$$

$$= \frac{\Pr\left[M = 0 \cap K = 0\right]}{\Pr\left[C = 0\right]} = \frac{\Pr\left[M = 0\right]\Pr\left[K = 0\right]}{\Pr\left[C = 0\right]}$$

Now we show that Pr[K = 0] = Pr[C = 0] = 1/2. Therefore, these two terms cancel in the above equation yielding $\Pr[M = 0 \mid C = 0] = \Pr[M = 0]$. The same argument applies for the other combinations of M and C.

$\Pr[K = 0] = 1/2$ is immediate, because there are 2 equally likely keys (namely 0 and 1).

$$\begin{aligned}
\Pr[C = 0] &= \Pr[M = 0 \cap K = 0] + \Pr[M = 1 \cap K = 1] \\
&= \Pr[M = 0] \times \Pr[K = 0] + \Pr[M = 1] \times \Pr[K = 1] \\
&= \Pr[M = 0] \times 1/2 + \Pr[M = 1] \times 1/2 \\
&= 1/2 \times (\Pr[M = 0] + \Pr[M = 1]) \\
&= 1/2
\end{aligned}$$

In the case of the one-time pad cryptosystem, the key is as long as the message, which means that the space of keys is as large as the space of messages. The next theorem shows that this is the case for any encryption scheme that is perfectly secure-version 1. In other words, any encryption scheme that is perfectly secure-version 1 suffers from the same impracticality issue as the one-time pad.

Notation: $\|A\|$ denotes the number of elements of the finite set A.

Theorem 2. If an encryption scheme is perfectly secure-version 1 over message space $\mathcal{M}$, then the set of keys $\mathcal{K}$ must satisfy $\|\mathcal{K}\| \geq \|\mathcal{M}\|$.

Proof. Let c be a ciphertext. Suppose $\|\mathcal{K}\| < \|\mathcal{M}\|$. Then when we decrypt c with all possible keys, we obtain at most $\|\mathcal{K}\|$ possible plaintexts. So there is a message m that is not obtained. Then $\Pr[M = m \mid C = c] = 0$. But clearly we can make a distribution with $P(M = m) > 0$, so this probability relation violates the definition of perfectly secure-version 1.

Thus if for example we look at messages that are say 1000 bits long, there are 2^{1000} possible messages, and we need at least 2^{1000} keys, so a key on average must be at least 1000 bits long. So, a perfectly secure version 1 is too much to ask, because it can be achieved only by very impractical encryption schemes (such as one-time pad).

The definition of an encryption that is perfectly secure - version 1 may seem to be too abstract and not be very convincing. Let us try another attempt for defining secrecy. This definition has the merit that it models the fact that the adversary does not get anything if she is doing a ciphertext-only attack.

Definition 2. An encryption scheme over message set $\mathcal{M}$ is **perfectly secure- version 2** if for any two messages m_1 and m_2 in $\mathcal{M}$ and for any algorithm A, we have

$$\Pr[A(C) = m_1 \mid C = \mathrm{E}(K, m_1)] = \Pr[A(C) = m_1 \mid C = \mathrm{E}(K, m_2)]$$

We can make the following observations.

1. Think of A as an attacker that wants to guess whether C is the encryption of m_1 or of m_2.
2. The definition assumes that the enemy does a ciphertext-only attack, because A has as input only C. Security against the other kind of attacks can be defined (more or less) similarly.
3. The probabilities are taken over the random choice of the key from $\mathcal{K}$ (and the random decisions of A if A is a probabilistic algorithm).
4. Instead of equality, suppose that the left-hand side of the above equation is greater than the right-hand side. A successful attacker would have the left-hand side big (ideally 1) and the right-hand side small (ideally 0).
5. The definition says that A is not doing any better at guessing the message when it is given an encryption of m_1 than when it is given an encryption of m_2.

Theorem 2. perfectly secure - version 2 = perfectly secure - version 1. (this means that an encryption scheme is secure according to version 1 if and only if it is secure according to version 2).

We omit the proof. It is not hard, but it is long.

Thus perfectly secure - version 2 cannot be achieved by practical encryption schemes either. So we adopt a more relaxed definition, which is computational secrecy.

Definition 3. Let ε be a small parameter (e.g., ε = 0:0001) and N be a large parameter (e.g., $N = 10^{80}$). An encryption scheme over message space $\mathcal{M}$ is computational secure (with parameters ε and N) if for any two messages m_1 and m_2 in $\mathcal{M}$ and for any algorithm A that performs N operations, we have:

$$|\Pr[A(C) = m_1 \mid C = \mathrm{E}(K, m_1)] - \Pr[A(C) = m_1 \mid C = \mathrm{E}(K, m_2)]| < \varepsilon$$

We can make the following observations.

1. There are two relaxations compared with "perfectly secure - version 2."
   - •We don't require equality between the two probabilities, just closeness within ε.
   - •And it is acceptable if the attacker can break the system by doing a huge number of operations: if an attacker must spend billions of year to break the cryptosystem, then the cryptosystem is considered secure.
2. The above definition only defines security against ciphertext-only attacks. In the same spirit, we can define computational secrecy against stronger types of attacks, such as chosen plaintext attack, or chosen ciphertext attack.
3. What should be the concrete values for N (the number of operation we allow the adversary to do) and ε (the bias we allow the adversary to achieve)? A common recommendations is that it is acceptable if no adversary running for at most $N = 2^{80}$ CPU cycles can break the system with probability greater than 2^{-64}.

Let's get a feel for these values. Computation on the order of $N = 2^{60}$ is barely within reach today. Running on a 3-GHz computer (that executes 3×10^9 cycles per second), 2^{60} cycles require $2^{60}/(3 \times 10^9)$ seconds or about 12 years. 2^{80} is $2^{20} \approx 10^6$ times longer than that. The number of seconds since the Big Bang is estimated to be in the order of 2^{58}.

An event that occurs once every hundred years can be roughly estimated to occur with probability 2^{-30} in any given second. Something that occurs with probability 2^{-60} in any given second is 2^{30} times less likely and might be expected to occur roughly once every 100 billion years.

F.2 INFORMATION AND ENTROPY

At the heart of information theory are two mathematical concepts with names that can be misleading: information and entropy. Typically, one thinks of **information** as having something to do with meaning; **entropy** is a term familiar from the second law of thermodynamics. In the discipline of information theory, information has to do with the reduction in the uncertainty about an event and entropy is an averaging of information values that happens to have a mathematical form identical to that for thermodynamic entropy.

Let us approach this new definition of information by way of an example. Imagine an investor who needs *information* (advice) about the status of certain securities, and who consults a broker with special *information* (knowledge) in that area. The broker *informs* (tells) the investor that, by coincidence, a federal investigator had come by just that morning seeking *information about* (evidence of) possible fraud by the corporation issuing that particular stock. In response to this *information* (data), the investor decides to sell, and so *informs* (notifies) the broker.

Put another way, being *uncertain* how to evaluate a portion of her portfolio, the client consults someone more *certain* than she about this side of the market. The broker relieves his client's *uncertainty* about relevant happenings by recounting the visit of the federal investigator, who had *uncertainties* to resolve of a professional nature. As an upshot of her increased *certainty* about the state of her securities, the client removes any *uncertainty* in the mind of the broker about her intention to sell.

Although the term *information* may signify notification, knowledge, or simply data, in each case the imparting of information is equivalent to the reduction in uncertainty. Information thus signifies the positive difference between two uncertainty levels.

Information

If we are to deal with information mathematically, then we need some quantity that is appropriate for measuring the amount of **information**. This problem was first raised, and solved, by Hartley in 1928 while studying telegraph communication. Hartley observed that if the probability that an event will occur is high (close to 1), there is little uncertainty that it will occur. If we subsequently learn that it has occurred, then the amount of information gained is small. Thus, one plausible measure is the reciprocal of the probability of the occurrence of an event: $1/p$. For example, an event that has an initial probability of occurrence of 0.25 conveys more information by its occurrence than one with an initial probability of 0.5. If the measure of information is $1/p$, then the occurrence of the first event conveys an information value of 4 (1/0.25) and the occurrence of the second event conveys an information value of 2 (1/0.5). But there are two difficulties in using this measure of information:

1. This measure does not seem to "work" for sequences of events. Consider a binary source that issues a stream of ones and zeros with equal probability of a one or zero for each bit. Thus, each bit has an information value of 2 (1/0.5). But if bit b_1 conveys a value of 2, what is the information conveyed by the string of two bits b_1b_2? This string can take on one of four possible outcomes, each with probability 0.25; therefore, by the $1/p$ measure, an outcome conveys an information value of 4. Similarly, the information value of 3 bits ($b_1b_2b_3$) is eight. This means that b_2 adds two units of information to the two of b_1, which is reasonable because the 2 bits have the same information value. But b_3 will add an additional four units of information. Extending the sequence, b_4 will add eight units of information, and so on. This does not seem reasonable as a measure of information.
2. Consider an event that gives rise to two or more independent variables. An example is a phase-shift-keying (PSK) signal that uses four possible phases and two amplitudes. A single signal element yields two units of information for the amplitude and four for the phase, for a total of six units by our measure. Yet each signal element is one of eight possible outcomes and hence ought to yield eight units of information by our measure.

Hartley overcame these problems by proposing that the measure of information for the occurrence of an event x be $\log(1/\mathrm{P}(x))$, where $\mathrm{P}(x)$ denotes the probability of occurrence of event x. Formally,

$$\mathrm{I}(x) = \log\,(1/\mathrm{P}(x)) = -\log \mathrm{P}(x) \qquad \textbf{(F.1)}$$

This measure "works" and leads to many useful results. The base of the logarithm is arbitrary but is invariably taken to the base 2, in which case the unit of measure is referred to as a bit. The appropriateness of this designation should be obvious as we proceed. Base 2 logarithms are assumed in the rest of this discussion. We can make the following observations:

1. A single bit that takes on the values 0 and 1 with equal probability conveys one bit of information (log(1/0.5) = 1). A string of two such bits takes on one of four equally likely outcomes with probability 0.25 and conveys two bits of information (log(1/0.25) = 2). Therefore, the second bit adds one bit of information. In a sequence of three independent bits, the third bit also adds one bit of information (log(1/0.125) = 3), and so on.
2. In the example of the PSK signal, a single signal element yields one bit of information for the amplitude and two for the phase, for a total of 3 bits, which agrees with the observation that there are eight possible outcomes.

Figure F.1 shows the information content for a single outcome as a function of the probability p of that outcome. As the outcome approaches certainty ($p = 1$), the information conveyed by its occurrence approaches zero. As the outcome approaches impossibility ($p = 0$), its information content approaches infinity.

Entropy

The other important concept in information theory is **entropy**, or **uncertainty**,[2] which was proposed in 1948 by Shannon, the founder of information theory. Shannon defined the entropy H as the average amount of information obtained from the value of a random variable. Suppose we have a random variable X, which may take on the values $x_1, x_2, \ldots, x_N$, and that the corresponding probabilities of each outcome are $P(x_1), P(x_2), \ldots, P(x_N)$. In a sequence of K occurrences of X, the outcome x_j will on average be selected $KP(x_j)$ times. Therefore, the average amount of information obtained from K outcomes is [using P_j as an abbreviation for $P(x_j)$]:

[2] Shannon used the term *entropy* because the form of the function H is the same as the form of the entropy function in statistical thermodynamics. Shannon interchangeably called H the *uncertainty function*.

$$KP_1 \log(1/P_1) + \ldots + KP_N \log(1/P_N)$$

Dividing by K yields the average amount of information per outcome for the random variable, referred to as the entropy of X, and designated by $H(X)$:

$$H(X) = \sum_{j=1}^{N} P_j \log\left(1/P_j\right) = -\sum_{j=1}^{N} P_j \log\left(P_j\right) \qquad \textbf{(F.2)}$$

The function H is often expressed as an enumeration of the probabilities of the possible outcomes: $H(P_1, P_2, \ldots, P_N)$.

As an example, consider a random variable X that takes on two possible values with respective probabilities p and $1 - p$. The entropy associated with X is

$$H(p, 1 - p) = -p\log(p) - (1 - p)\log(1 - p)$$

Figure F.2 plots $H(X)$ for this case as a function of p. Several important features of entropy are evident from this figure. First, if one of the two events is certain ($p = 1$ or $p = 0$), then the entropy is zero.[3] One of the two events has to occur and no information is conveyed by its occurrence. Second, the maximum value of H(X) = 1 is reached when the two outcomes are equally likely. This seems reasonable: the uncertainty of the outcome is maximum when the two outcomes are equally likely. This result generalizes to a random variable with N outcomes: its entropy is maximum when the outcomes are equally likely:

$$\max H(P_1, P_2, \ldots, P_N) = H(1/N, 1/N, \ldots, 1/N)$$

For example:

[3] Strictly speaking, the formula for $H(X)$ is undefined at $p = 0$. The value is assumed to be 0 for $p = 0$. This is justified because the limit of $H(X)$ as p goes to 0 is 0.

$$H(1/3, 1/3, 1/3) = 1/3 \log 3 + 1/3 \log 3 + 1/3 \log 3 = 1.585$$

whereas

$$H(1/2, 1/3, 1/6) = 1/2 \log 2 + 1/3 \log 3 + 1/6 \log 6 = 0.5 + 0.528 + 0.43 = 1.458$$

Properties of the Entropy Function

We have developed the entropy formula $H(X)$ by an intuitive line of reasoning. Another approach is to define the properties that an entropy function should have and then prove that the formula $-\sum_i P_i \log P_i$ is the only formula that has these properties. These properties, or axioms, can be stated as follows:

1. H is continuous over the range of probabilities. Thus, small changes in the probability of one of the occurrences only cause small changes in the uncertainty. This seems a reasonable requirement.
2. If there are N possible outcomes and they are equally likely, so that $P_i = 1/N$, then $H(X)$ is a monotonically increasing function of N. This is also a reasonable property because it says that the more equally likely outcomes, the larger the uncertainty.
3. If some of the outcomes of X are grouped, then H can be expressed as a weighted sum of entropies in the following fashion:

$$H(P_1, P_2, P_3, \ldots, P_N) = H(P_1 + P_2, P_3, \ldots, P_N) + (P_1 + P_2)H\left(\frac{P_1}{P_1 + P_2}, \frac{P_2}{P_1 + P_2}\right)$$

The reasoning is as follows. Before the outcome is known, the average uncertainty associated with the outcome is $H(P_1, P_2, P_3, \ldots, P_N)$. If we reveal which outcome has occurred, except that the first two outcomes are grouped together, then the average amount of uncertainty removed is $H(P_1 + P_2, P_3, \ldots, P_N)$. With probability $(P_1 + P_2)$, one of the first two outcomes occurs and the remaining uncertainty is $H[P_1/(P_1 + P_2) + P_2/(P_1 + P_2)]$.

The only definition of $H(X)$ that satisfies all three properties is the one that we have given. To see property (1), consider Figure F.2, which is clearly continuous in p. It is more difficult to depict $H(X)$ when there are more than two possible outcomes, but the fact of continuity should be clear.

For property (2), if there are N equally likely outcomes, then $H(X)$ becomes

$$H(X) = -\sum_{j=1}^{N} \frac{1}{N} \log\left(\frac{1}{N}\right) = -\log\left(\frac{1}{N}\right) = \log(N)$$

The function $\log(N)$ is a monotonically increasing function of N. Note that with four possible outcomes, the entropy is 2 bits; with eight possible outcomes, the entropy is 3 bits, and so on.

As a numerical example of property (3), we may write

$$H\left(\frac{1}{2},\frac{1}{3},\frac{1}{6}\right) = H\left(\frac{5}{6},\frac{1}{6}\right) + \frac{5}{6} H\left(\frac{3}{5},\frac{2}{5}\right)$$

$$1.458 = 0.219 + 0.43 + \frac{5}{6}(0.442 + 0.5288)$$

$$= 0.649 + 0.809$$

Conditional Entropy

Shannon defines the conditional entropy of Y given X, expressed as $\mathrm{H}(Y \mid X)$, as the uncertainty about Y given knowledge of X. This conditional entropy is defined as follows:

$$H(Y \mid X) = -\sum_{x,y} \Pr(x,y) \log \Pr(y \mid x)$$

where

x = a value contained in the set X

y = a value contained in the set Y

$\Pr(x, y)$ = probability of the joint occurrence of x for the value in X and y for the value in Y

Conditional uncertainties obey intuitively pleasing rules, such as:

$$\mathrm{H}(X, Y) = \mathrm{H}(X) + \mathrm{H}(Y \mid X)$$

F.3 ENTROPY AND SECRECY

For a symmetric encryption system, the basic equations are $C = \mathrm{E}(K, M)$ and $M = \mathrm{E}(K, C)$. These equations can be written equivalently, in terms of uncertainties as

$$H(C \mid K, M) = 0$$

and

$$H(M \mid K, C) = 0 \qquad \textbf{(F.3)}$$

respectively, because, for instance $H(C \mid K, M)$ is zero if and only if, M and K uniquely determine C, which is a basic requirement of symmetric encryption.

Shannon's definition of perfect secrecy can then be written as:

$$H(M \mid C) = H(M) \qquad \textbf{(F.4)}$$

because this equality holds if and only if M is statistically independent of C.

For any secret key cryptosystem, we can write;

$$\begin{aligned} H(M \mid C) &\le H(M, K \mid C) \\ &= H(K \mid C) + H(M \mid K, C) \\ &= H(K \mid C) \end{aligned}$$

$$\leq H(K) \tag{F.5}$$

where we have used Equation (F.3) and the fact that removal of given knowledge can only increase uncertainty. If the cryptosystem provides perfect secrecy, it follows from Equations (F.4) and (F.5) that

$$H(K) \geq H(M) \tag{F.6}$$

Inequality (F.6) is Shannon's fundamental bound for perfect secrecy. The uncertainty of the secret key must be at least as great as the uncertainty of the plaintext that it is concealing. Let us assume we are dealing with binary values; that is, the plaintext, key, and ciphertext are represented as binary strings. Then we can say that for a key of length k bits,

$$H(K) \leq -\log(2^{-k}) = k \tag{F.7}$$

with equality if and only if the key is completely random. Similarly, if the length of the plaintext is q, then

$$H(M) \leq -\log(2^{-q}) = q \tag{F.8}$$

with equality if and only if the plaintext is completely random, which means each q-bit plaintext is equally likely to occur. Combining inequalities (F.6, F.7, F.8), the requirement for perfect secrecy if the plaintext is completely random is $k \geq q$. That is, the key must be at least as long as the plaintext. For the one-time pad, we have $k = q$.

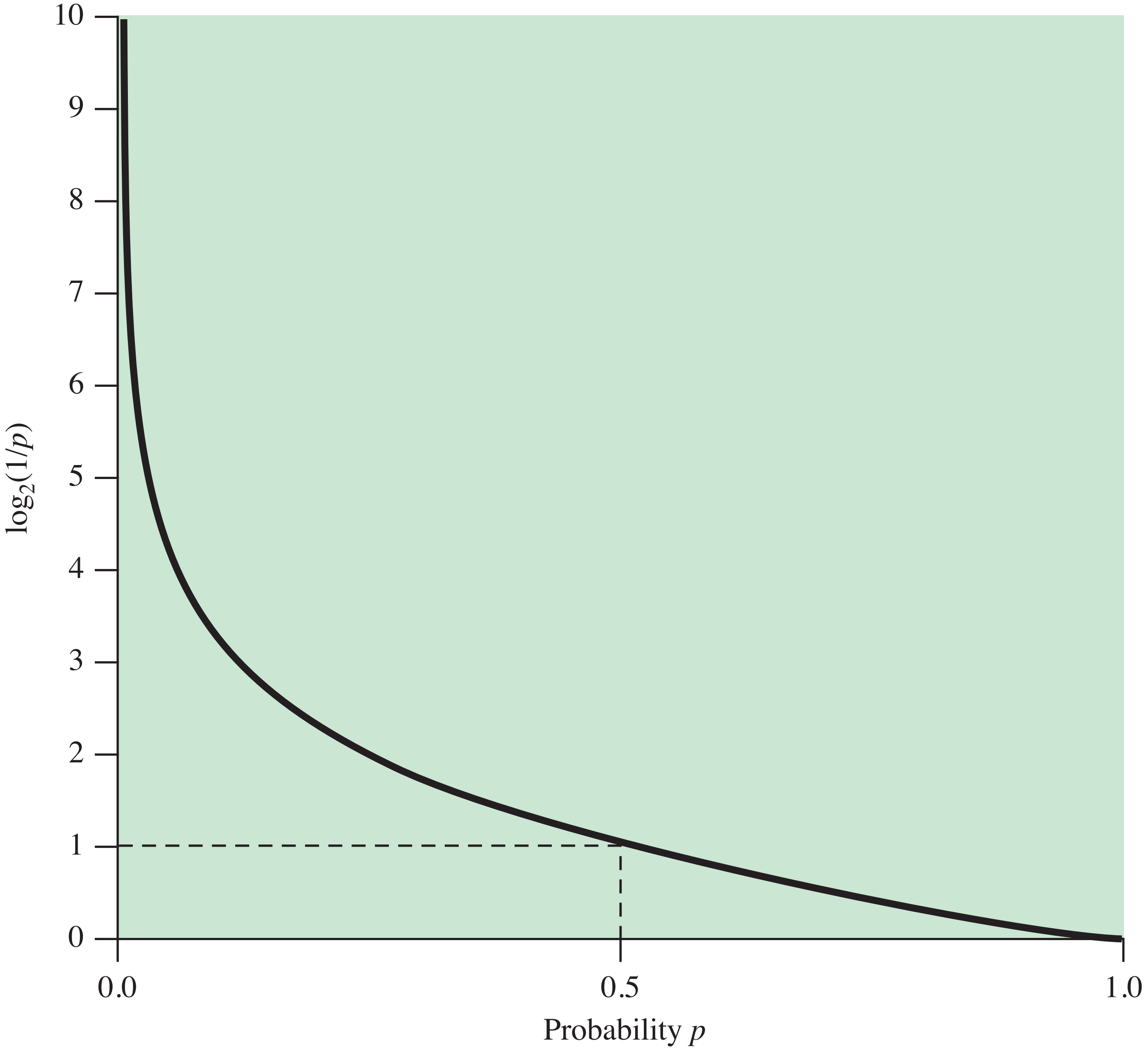

Figure F.1 Information Measure for a Single Outcome

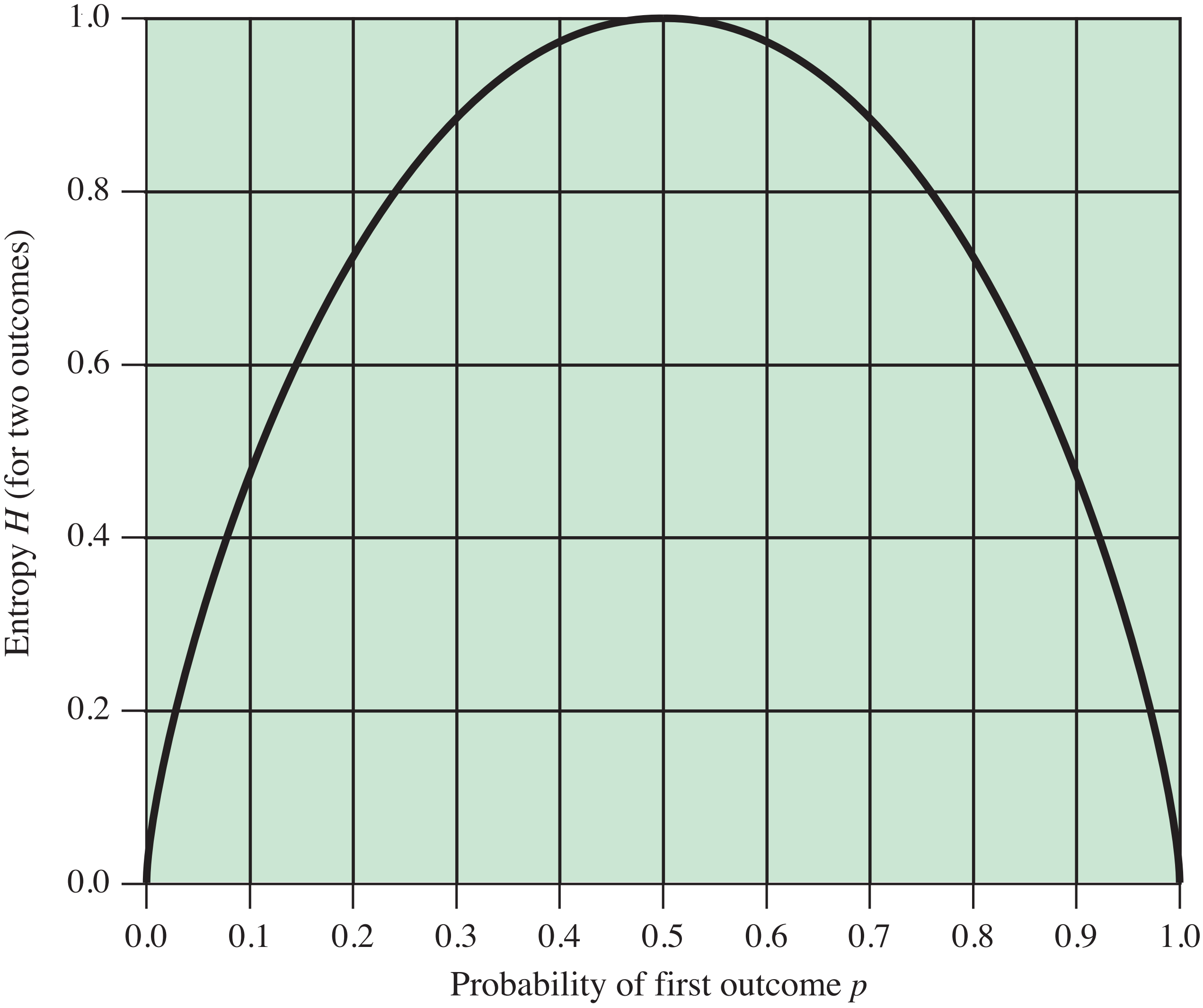

Figure F.2 Entropy Function for Random Variable with Two Outcomes