

Definitions: The *rate of a language* for messages X of length N is $r = H(X)/N$ = the average number of bits per character. (For English, r is about 1 or 1.5—use 1 in HW.)

The *absolute rate of a language* is the maximum number of bits per letter $R = \log_2 a$, where a is the alphabet size. (For English, $R = \log_2 26 \approx 4.7$ bits per letter.)

The *redundancy of a language* is $D = R - r$.

The *redundancy rate of a language* is D/R , often expressed as a percentage. (For English, this is about 79% to 68%.)

Fact. In English, there are $2^{RN} = 26^N$ N -letter messages. 2^{rN} of them are meaningful and $2^{RN} - 2^{rN}$ are meaningless.

We want to measure the effect of (some) additional information on uncertainty.

Example. Suppose X is one of n (n even) equally likely integers: $X \in \{1, 2, \dots, n\}$. Then $H(X) = \log_2 n$. But if X is known to be *even*, then the entropy decreases by 1 to $-1 + \log_2 n = \log_2(n/2)$. If Y is the message “ X is even,” then we write $H_Y(X) = -1 + \log_2 n$.

Definition. Write $p_Y(X) = P(X|Y)$ for the conditional probability of X given Y . Write $p(X, Y) = p_Y(X) \cdot p(X)$ for the joint probability of X and Y .

Definition. The *equivocation* or *conditional entropy of X given Y* is the amount of uncertainty remaining about X after Y is learned. In symbols, this is

$$H_Y(X) = - \sum_{X,Y} p(X, Y) \log_2 p_Y(X)$$

or

$$H_Y(X) = \sum_Y p(Y) \sum_X p_Y(X) \log_2 p_Y(X).$$

Example. Suppose X and Y each can be one of four equally likely messages, and each Y message limits X to one of two equally likely messages. (For example, Y_2 might be, “ X is X_1 or X_4 ”.) Then each $p_Y(X)$ is $1/2$ or 0 , so

$$H_Y(X) = 4[(1/4) \cdot 2(1/2) \log_2 2] = 1.$$

Perfect Secrecy

Let M , C , K have probability $p(M)$, $p(C)$, $p(K)$, respectively. Let $p_C(M)$ denote the probability that M was sent given that C was received. Let $p_M(C)$ denote the probability that C was received given that M was sent.

Definition. A cipher has *perfect secrecy* if $p_C(M) = p(M)$ for all C and all M . That is, intercepting C gives no information at all about M (or K).

Theorem. A cipher has perfect secrecy if and only if $p_M(C) = p(M)$ for all C and all M .

This implies that $p_M(C) = p_{M'}(C)$ for every C , M , M' . It also means that the number of keys must be \geq the number of messages if perfect secrecy is desired. One can achieve perfect secrecy using completely random keys at least as long as the messages they encipher.

Example. The Caesar cipher $E_K(m) = (m + K) \bmod 26$ does not have perfect secrecy (except for messages of length 1 letter) because the key is not long enough.

Example. A *one-time pad* Caesar cipher $E_K(m_i) = (m_i + k_i) \bmod 26$ (where $K = k_1 k_2 \dots$, a string of letters) has perfect secrecy.