Notes on Information Theory I and The Source Coding Theorem

H.L. Rappaport

October 2014

Consider a discrete memoryless channel with an input alphabet

$$\mathcal{X} = \{x_0, x_1, \ldots, x_{J-1}\}, \tag{1}$$

and an output alphabet

$$\mathcal{Y} = \{y_0, y_1, \ldots, y_{K-1}\}. \tag{2}$$

Denote the channel input by the discrete random variable X and the channel output by the discrete random variable Y. Let the probability of X obtaining the value x_j be denoted $p(x_j)$ and the probability of Y obtaining the value y_k be denoted $p(y_k)$. Furthermore, the joint probability that $X = x_j$ and $Y = y_k$ will be denoted $p(x_j, y_k)$. We immediately have the properties

$$\sum_{j=0}^{J-1} p(x_j) = \sum_{k=0}^{K-1} p(y_k) = \sum_{j=0}^{J-1} \sum_{k=0}^{K-1} p(x_j, y_k) = 1, \tag{3}$$

and

$$p(x_j) = \sum_{k=0}^{K-1} p(x_j, y_k), \qquad p(y_k) = \sum_{j=0}^{J-1} p(x_j, y_k). \tag{4}$$

The channel matrix may be written [1]

$$\mathbf{P} = \begin{bmatrix} p(y_0|x_0) & p(y_1|x_0) & \ldots & p(y_{K-1}|x_0) \\ p(y_0|x_1) & p(y_1|x_1) & \ldots & p(y_{K-1}|x_1) \\ \vdots & \vdots & & \vdots \\ p(y_0|x_{J-1}) & p(y_1|x_{J-1}) & \ldots & p(y_{K-1}|x_{J-1}) \end{bmatrix}, \tag{5}$$

where $p(y_k|x_j)$ is a transition probability or conditional probability that the channel output is y_k given that the channel input is x_j. The transition probability $p(y_k|x_j)$ is the conditional probability of correct reception when $j = k$ and a conditional probability of error when $j \neq k$.

Key Properties of Conditional Probabilities [2]

$$p(x_j, y_k) = p(y_k|x_j)\, p(x_j) = p(x_j|y_k)\, p(y_k), \tag{6}$$

where $p\left(x_j|y_k\right)$ is the conditional probability that the channel input is x_j given that the channel output is y_k. Substitution of Eq. (6) into

$$p\left(x_j\right) = \sum_{k=0}^{K-1} p\left(x_j, y_k\right), \tag{7}$$

shows

$$\sum_{k=0}^{K-1} p\left(y_k|x_j\right) = 1, \qquad p(x_j) \neq 0, \tag{8}$$

and

$$p\left(x_j\right) = \sum_{k=0}^{K-1} p\left(x_j|y_k\right) p\left(y_k\right). \tag{9}$$

Likewise, substitution of Eq. (6) into

$$p\left(y_k\right) = \sum_{j=0}^{J-1} p\left(x_j, y_k\right), \tag{10}$$

shows

$$\sum_{j=0}^{J-1} p\left(x_j|y_k\right) = 1, \qquad p\left(y_k\right) \neq 0, \tag{11}$$

and

$$p\left(y_k\right) = \sum_{j=0}^{J-1} p\left(y_k|x_j\right) p\left(x_j\right). \tag{12}$$

Entropy Definitions [1]

The entropy of the channel input which is discrete memoryless source is given by

$$H\left(\mathcal{X}\right) = -\sum_{j=0}^{J-1} p\left(x_j\right) \log_2 p\left(x_j\right). \tag{13}$$

The entropy of the channel output is given by

$$H\left(\mathcal{Y}\right) = -\sum_{k=0}^{K-1} p\left(y_k\right) \log_2 p\left(y_k\right). \tag{14}$$

The *joint entropy* of the channel input and the channel output is given by

$$H\left(\mathcal{X}, \mathcal{Y}\right) = -\sum_{j=0}^{J-1} \sum_{k=0}^{K-1} p\left(x_j, y_k\right) \log_2 p\left(x_j, y_k\right). \tag{15}$$

The *conditional entropy* of channel input after the channel output has been observed satisfies

$$H\left(\mathcal{X}|\mathcal{Y}\right) = -\sum_{k=0}^{K-1}\sum_{j=0}^{J-1} p\left(x_j, y_k\right)\log_2 p\left(x_j|y_k\right). \tag{16}$$

The *conditional entropy* of channel output given the channel input satisfies

$$H\left(\mathcal{Y}|\mathcal{X}\right) = -\sum_{k=0}^{K-1}\sum_{j=0}^{J-1} p\left(x_j, y_k\right)\log_2 p\left(y_k|x_j\right). \tag{17}$$

Entropy of n^{th} Order Extended Alphabet [1]

The second order extended alphabet is the set of all possible ordered pairs of elements of the alphabet Denoting the probability $p\left(x_i\right) = p_i$ and the probability $p\left(x_j\right) = p_j$ then since the source is memoryless the probability of the individual source alphabet symbols are independent and the probability of an order pair $x_i x_j$ satisfies

$$p\left(x_i x_j\right) = p_i p_j \tag{18}$$

. The entropy of the second order extended alphabet is then

$$H\left(\mathcal{X}^2\right) = -\sum_{i,j} p_i p_j \log_2\left(p_i p_j\right), \tag{19}$$

and the summations are over all members of the alphabet. Now

$$H\left(\mathcal{X}^2\right) = -\sum_{i,j} p_i p_j \left[\log_2 p_i + \log_2 p_j\right], \tag{20}$$

$$= -\sum_{j} p_j \sum_{i} p_i \log_2 p_i - \sum_{i} p_i \sum_{j} p_j \log_2 p_j = 2H\left(\mathcal{X}\right). \tag{21}$$

In a similar manner the entropy of the n^{th} order extended alphabet or ordered elements taken n at a time is given by

$$H\left(\mathcal{X}^n\right) = -\sum_{j_1, j_2, j_3, \ldots j_n} p_{j_1} p_{j_2} p_{j_3} \ldots p_{j_n} \log_2\left[p_{j_1} p_{j_2} p_{j_3} \ldots p_{j_n}\right], \tag{22}$$

$$= -\sum_{j_2, j_3, j_4, \ldots j_n} p_{j_2} p_{j_3} p_{j_4} \ldots p_{j_n} \sum_{j_1} p_{j_1} \log_2 p_{j_1} - \sum_{j_1, j_3, j_4, \ldots j_n} p_{j_1} p_{j_3} p_{j_4} \ldots p_{j_n} \sum_{j_2} p_{j_2} \log_2 p_{j_2}$$

$$-\sum_{j_1,j_2,j_4,\ldots j_n} p_{j_1}p_{j_2}p_{j_4}\ldots p_{j_n}\sum_{j_3} p_{j_3}\log_2 p_{j_3}\ldots - \sum_{j_1,j_2,j_3,\ldots j_{n-1}} p_{j_1}p_{j_2}p_{j_3}\ldots p_{j_{n-1}}\sum_{j_n} p_{j_n}\log_2 p_{j_n}, \quad (23)$$

$$= nH(\mathcal{X}). \quad (24)$$

Mutual Information [1]

The mutual information $I(\mathcal{X};\mathcal{Y})$ is defined by

$$I(\mathcal{X};\mathcal{Y}) = H(\mathcal{X}) - H(\mathcal{X}|\mathcal{Y}), \quad (25)$$

and represents our uncertainty about the channel input that is resolved by observing the channel output.

Properties of the mutual information include

$$I(\mathcal{X};\mathcal{Y}) = I(\mathcal{Y};\mathcal{X}), \quad (26)$$

or

$$I(\mathcal{X};\mathcal{Y}) = H(\mathcal{Y}) - H(\mathcal{Y}|\mathcal{X}), \quad (27)$$

and

$$I(\mathcal{X};\mathcal{Y}) = H(\mathcal{X}) + H(\mathcal{Y}) - H(\mathcal{X},\mathcal{Y}). \quad (28)$$

To show Eq. (26) or equivalently Eq. (27) we need to show

$$H(\mathcal{X}) - H(\mathcal{X}|\mathcal{Y}) = H(\mathcal{Y}) - H(\mathcal{Y}|\mathcal{X}). \quad (29)$$

Direct substitution of the definitions of the entropies Eqs. (13) - (17) shows

$$H(\mathcal{X}) - H(\mathcal{X}|\mathcal{Y}) - H(\mathcal{Y}) + H(\mathcal{Y}|\mathcal{X}) \quad (30)$$

$$= -\sum_j p(x_j)\log_2 p(x_j) + \sum_{j,k} p(x_j,y_k)\log_2 p(x_j|y_k)$$

$$+\sum_k p(y_k)\log_2 p(y_k) - \sum_{j,k} p(x_j,y_k)\log_2 p(y_k|x_j) \quad (31)$$

$$= -\sum_j p(x_j)\log_2 p(x_j) + \sum_k p(y_k)\log_2 p(y_k) + \sum_{j,k} p(x_j, y_k)\log_2 \frac{p(x_j|y_k)}{p(y_k|x_j)}. \tag{32}$$

From Eq. (6)

$$\frac{p(x_j|y_k)}{p(y_k|x_j)} = \frac{p(x_j)}{p(y_k)}. \tag{33}$$

So

$$\begin{aligned} &H(\mathcal{X}) - H(\mathcal{X}|\mathcal{Y}) - H(\mathcal{Y}) + H(\mathcal{Y}|\mathcal{X}) \\ &= -\sum_j p(x_j)\log_2 p(x_j) + \sum_k p(y_k)\log_2 p(y_k) \\ &+ \sum_{j,k} p(x_j, y_k)\log_2 p(x_j) - \sum_{j,k} p(x_j, y_k)\log_2 p(y_k) \qquad (34) \\ &= -\sum_j p(x_j)\log_2 p(x_j) + \sum_k p(y_k)\log_2 p(y_k) \\ &+ \sum_{j,k} p(x_j|y_k)\, p(y_k)\log_2 p(x_j) - \sum_{j,k} p(y_k|x_j)\, p(x_j)\log_2 p(y_k) = 0, \qquad (35) \end{aligned}$$

when Eqs. (9) and (12) are used, which is the desired result.

Mutual Information in terms of Joint Entropy [1]

Next consider Eq. (28)

$$I(\mathcal{X};\mathcal{Y}) = H(\mathcal{X}) + H(\mathcal{Y}) - H(\mathcal{X},\mathcal{Y}).$$

To prove this result first notice that using Eq. (25) it follows from

$$H(\mathcal{X}) - H(\mathcal{X}|\mathcal{Y}) = H(\mathcal{X}) + H(\mathcal{Y}) - H(\mathcal{X},\mathcal{Y}), \tag{36}$$

or

$$H(\mathcal{X},\mathcal{Y}) - H(\mathcal{X}|\mathcal{Y}) - H(\mathcal{Y}) = 0. \tag{37}$$

Substituting the definitions of the entropies Eqs. (13)-(17) into the left hand side of Eq. (37) gives

$$-\sum_{j,k} p(x_j, y_k)\log_2 p(x_j, y_k) + \sum_{j,k} p(x_j, y_k)\log_2 p(x_j|y_k) + \sum_k p(y_k)\log_2 p(y_k) \tag{38}$$

$$= -\sum_{j,k} p\left(x_j, y_k\right) \log_2 \frac{p\left(x_j, y_k\right)}{p\left(x_j|y_k\right)} + \sum_k p\left(y_k\right) \log_2 p\left(y_k\right) \tag{39}$$

$$= -\sum_{j,k} p\left(x_j, y_k\right) \log_2 p\left(y_k\right) + \sum_k p\left(y_k\right) \log_2 p\left(y_k\right) = 0, \tag{40}$$

where Eqs. (4) and (6) have been used thus proving the result.

Source Coding Theorem

One way to understand the source coding theorem [4, 5] is to first consider a general discrete source with an output alphabet of J symbols. It would require $\log_2 J$ bits to transmit a given symbol. Similarly, for a block of n sequential ordered elements of the alphabet there would be J^n possible messages that can be transmitted in one block. It would thus require $\log_2 J^n = n \log_2 J$ bits to send a block of n symbols.

Next specialize to the case of a discrete memoryless source also with an output alphabet of J symbols. Let the i^{th} member of the alphabet be denoted x_i and let the probability of occurrence of x_i be denoted p_i. Now the block of n symbols where n is very large is considered. Let the expected number of occurrences of x_i be denoted n_i. We have that for n very large

$$n_i = p_i n. \tag{41}$$

That is to say for n very large the number of occurrences of each of the x_i is known and the number of *likely* messages is not J^n but instead the number of distinct permutations of a block of symbols with

$$\begin{gathered} p_0 n \text{ occurrences of } x_0, \\ p_1 n \text{ occurrences of } x_1, \\ p_2 n \text{ occurrences of } x_2, \\ \vdots \\ p_{J-1} n \text{ occurrences of } x_{J-1}. \end{gathered} \tag{42}$$

So the number of possible messages that can be sent in one block of n symbols is

$$\text{Num. of messages} = \frac{n!}{(p_0 n)!\,(p_1 n)!\,(p_2 n)! \ldots (p_{J-1} n)!}. \tag{43}$$

Therefore, the number of bits required to send these messages or, equivalently, the number of bits needed to indicate to a receiver which of these messages was transmitted would be

$$N_{\text{bits}} = \log_2 \left[\frac{n!}{(p_0 n)!\,(p_1 n)!\,(p_2 n)! \ldots (p_{J-1} n)!} \right], \tag{44}$$

and

$$N_{\text{bits}} \ln 2 = \ln n! - \sum_{j=0}^{J-1} \ln \left[(p_j n)! \right]. \tag{45}$$

Using the approximation for $\ln n!$ discussed in the appendix shows

$$N_{\text{bits}} \ln 2 = n \ln n - n - \sum_{j=0}^{J-1} \left[p_j n \ln (p_j n) - p_j n \right], \tag{46}$$

$$= n \ln n - n - n \sum_{j=1}^{J} \left[p_j \ln p_j + p_j \ln n - p_j \right], \tag{47}$$

$$= -n \sum_{j=0}^{J-1} p_j \ln p_j. \tag{48}$$

Or

$$N_{\text{bits}} = -n \sum_{j=0}^{J-1} p_j \log_2 p_j = nH(\mathcal{X}). \tag{49}$$

Of course *more* bits can always be used to code all of these different messages but this presents the minimum needed when n is very large. The minimum required average number of bits per transmitted symbol is denoted $\bar{L}_{\text{min}}$ and thus satisfies

$$\bar{L}_{\text{min}} = \frac{N_{\text{bits}}}{n} = H(\mathcal{X}). \tag{50}$$

Result (50) is the source coding theorem [4]. Of course we have not shown that uniquely decodable codes exist that operate on one symbol at a time and can obtain this limit. However, we have shown that this limit exists.

Appendix

Stirling's Formula [3] gives

$$\ln \Gamma(z) \sim \left(z - \frac{1}{2}\right) \ln z - z + \frac{1}{2} \ln(2\pi) + \frac{1}{12z} + \cdots, \tag{51}$$

as $z \to \infty$ in $|\arg z| < \pi$ where

$$z! = \Gamma(z+1). \tag{52}$$

Thus

$$\ln z! \sim \left(z + \frac{1}{2}\right) \ln(z+1) - z - 1 + \frac{1}{2} \ln(2\pi) + \frac{1}{12(z+1)} + \cdots. \tag{53}$$

Therefore the leading order contributions to $\ln z!$ for $z \gg 1$ is just

$$\ln z! \sim z \ln z - z, \tag{54}$$

and we will use

$$\ln n! \approx n \ln n - n, \tag{55}$$

in our discussion of the source coding theorem.

References

[1] S. Haykin, *Digital Communications*, Wiley, N.Y. (1988), chapter 2.

[2] A. Papoulis, *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, N.Y. (1965).

[3] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions*, Dover, N.Y. (1965), pp. 255-257.

[4] Haykin 1988, p. 21.

[5] C.E. Shannon, *A Mathematical Theory of Communication*, Bell System Technical Journal (1948).