Information Theory, Part I.

John MacLaren Walsh, Ph.D.

ECET 602, Spring Quarter, 2015

1 References

- *Elements of Information Theory*, 2nd ed., T. M. Cover and J. A. Thomas, John Wiley and Sons, 2006.
- *Information Theory and Reliable Communication*, R. G. Gallager, John Wiley and Sons, 1968.
- *Information Theory, Inference, and Learning Algorithms*, D. J. C. MacKay, Cambridge University Press, 2003.
- *Information Theory: Coding Theorems For Discrete Memoryless Systems.*, I. Csiszar, J. Koerner, Academic Press, 1981.
- *Information Theory and Network Coding*, R. W. Yeung, Springer Science and Business Media, 2008.

Notes follow the introductory material in the first reference very closely.

2 Overview

Using the tools of typical sequences and the asymptotic equipartition property (AEP) we establish a region of achievable rates for lossless source coding for sequences of discrete valued i.i.d. random variables. Next, using the tools of jointly typical sequences and the joint AEP, we establish a region of achievable rates for channel coding for discrete memoryless channels. The next lecture establishes that these achievable rates are the only achievable rates over all possible codes.

3 The Entropy Function (Cover and Thomas Chapter 2)

Given a random variable X drawn from a finite set $\mathcal{X}$ according to the distribution Q, the Shannon entropy of the random variable is defined as

$$\mathsf{H}_X := -\sum_{a \in \mathcal{X}} \mathsf{Q}(a) \log(\mathsf{Q}(a))$$

Note that this function is non-negative, concave, and achieves its maximum at $\mathsf{Q}(a) = 1/|\mathcal{X}|$ for all a, yielding an associated entropy of $\log_2(|\mathcal{X}|)$ (We proved this fact in class using the first order necessary conditions and a Lagrange multiplier).

4 WLLN and Convergence In Probability

Recall from lecture that a sequence of random variables X_N is said to converge in probability to a random variable X if for all $\epsilon > 0$

$$\lim_{N \to \infty} \Pr\left[\{\omega \, || X_N(\omega) - X(\omega)| > \epsilon\}\right] = 0$$

We noted that this is very different from convergence in distribution (i.e. weak convergence), using the illustrative example of a sequence of i.i.d. Gaussian random variables with zero mean and unit variance, which trivially converges in distribution (i.e. weakly), but does not converge in probability.

Next, we discussed the weak law of large numbers, which says that if $\{X_i\}$ is a sequence of independent and identically distributed random variables with mean μ, the random variable $Y_N := \frac{1}{N}\sum_{i=1}^{N} X_i$ converges in probability to μ.

5 Typical Sequences and AEP (Cover and Thomas Chapter 3)

Consider a sequence of N discrete random variables $\mathbf{X} := [X_1, X_2, \ldots, X_N]$ each taking values in the set $\mathcal{X}$, that are independently and identically distributed according to the distribution Q.

We begin by noting that

$$-\frac{1}{N}\log(\mathbb{P}(\mathbf{X})) \to \mathsf{H}(\mathsf{Q}) \tag{1}$$

in probability as $N \to \infty$. This can be seen by noting that

$$\frac{1}{N}\log(\mathbb{P}(\mathbf{X})) = \frac{1}{N}\log\left(\prod_{i=1}^{N}\mathsf{Q}(X_i)\right) = \frac{1}{N}\sum_{i=1}^{N}\log(\mathsf{Q}(X_i))$$

is the empirical average of N i.i.d. random variables, which by the WLLN converges in probability to the expectation of one of the i.i.d. random variables, which is the entropy in this case.

Bearing this in mind, we define the set $\mathcal{A}_\epsilon^N$ of typical sequences (i.e. the typical set) to be the set of all deterministic sequences $\mathbf{x} = [x_1, x_2, \ldots, x_N]$ of N elements of the set $\mathcal{X}$ whose log probability are no further than ϵ from the entropy

$$\mathcal{A}_\epsilon^N := \left\{\mathbf{x} | e^{-N(\mathsf{H}(\mathsf{Q})+\epsilon)} \leq \mathsf{Q}(\mathbf{x}) \leq e^{-N(\mathsf{H}(\mathsf{Q})-\epsilon)}\right\}$$

This set has the properties

1. $\mathbb{P}[\mathbf{X} \in \mathcal{A}_\epsilon^N] > 1 - \epsilon$ for N sufficiently large.
2. $|\mathcal{A}_\epsilon^N| \leq 2^{N(\mathsf{H}(\mathsf{Q})+\epsilon)}$
3. $|\mathcal{A}_\epsilon^N| \geq (1-\epsilon)2^{N(\mathsf{H}(\mathsf{Q})-\epsilon)}$ for sufficiently large N.

The first property follows from the directly from the definition of convergence in probability and the fact (1).

We then proved in class the first bound on the size of the typical set using the relations

$$\begin{aligned}
1 &\geq \mathbb{P}[\mathbf{X} \in \mathcal{A}_\epsilon^N] && (2)\\
&= \sum_{\mathbf{x}\in\mathcal{A}_\epsilon^N} \mathbb{P}[\mathbf{X} = \mathbf{x}] && (3)\\
&\geq \sum_{\mathbf{x}\in\mathcal{A}_\epsilon^N} 2^{-N(\mathsf{H}(\mathsf{Q})+\epsilon)} && (4)\\
&= |\mathcal{A}_\epsilon^N| 2^{-N(\mathsf{H}(\mathsf{Q})+\epsilon)} && (5)
\end{aligned}$$

Where the second inequality followed directly from the definition of the typical set. The students then proved the third property in groups by mimicking the proof of the second property, but starting with the equation from the first property.

6 A Basic Source Code

It can be roughly stated that the goal of lossless source coding is to compress a signal into a minimal number of bits, while still allowing for its successful reproduction from the compressed form. Here we consider the encoding of a sequence of N discrete random variables $\mathbf{X} := [X_1, X_2, \ldots, X_N]$ each taking values in the set $\mathcal{X}$, that are independently and identically distributed according to the distribution Q.

The properties of the typical set naturally allow us to define a family of source codes allowing for significant compression of this source. Fix an ϵ. Index all of the elements of the typical set. (In binary such an indexing would require $\lceil \log_2(|\mathcal{A}_\epsilon^N|)\rceil$ bits, which is less than $N(\mathsf{H}(\mathsf{Q}) + \epsilon) + 1$ bits.)

If the sequence to encode is in the typical set, then we transmit 0 followed by the binary representation of its index within the typical set. If the sequence to encode is not in the typical set, then we pre-pend it with 1, and send it uncompressed. This code has expected length

$$\begin{aligned} \text{expected Length} &\leq \mathbb{P}[\mathbf{X} \in \mathcal{A}_\epsilon^N](N(\mathsf{H}(\mathsf{Q}) + \epsilon) + 1) + \mathbb{P}[\mathbf{X} \notin \mathcal{A}_\epsilon^N](N \log_2(|\mathcal{X}|) + 2) \quad (6) \\ &\leq N(\mathsf{H}(\mathsf{Q}) + \epsilon) + 1 + \epsilon(N \log_2(|\mathcal{X}|) + 2) \quad (7) \end{aligned}$$

Dividing through by N shows that the number of bits per symbol approaches the entropy as $N \to \infty$.

7 Jointly Typical Sequence and Joint AEP (Chapter 7)

Next consider a sequence of pairs of random variables (X_n, Y_n) distributed for each n independently accordingly to a joint distribution $\mathsf{p}_{X,Y}$. We then will have not only convergence of $-\frac{1}{N}\log(\mathsf{p}(X_1, Y_1, X_2, Y_2, \ldots, X_N, Y_N))$ in probability to the joint entropy $\mathsf{H}(X, Y)$, but we will also have $-\frac{1}{N}\log(\mathsf{p}(X_1, X_2, \ldots, X_N))$ converging in probability to $\mathsf{H}(X)$ and $-\frac{1}{N}\log(\mathsf{p}(Y_1, Y_2, \ldots, Y_N))$ converging in probability to $\mathsf{H}(Y)$. The set of jointly typical sequences of length N is then the set of realizations $\mathbf{x} = [x_1, x_2, \ldots, x_N]$ and $\mathbf{y} = [y_1, y_2, \ldots, y_N]$ of N for which each of these convergences has completed within a common ϵ

$$\mathcal{A}_\epsilon^N := \left\{ (\mathbf{x}, \mathbf{y}) \left| \left| -\frac{1}{N}\log\left(\mathsf{p}(\mathbf{x}, \mathbf{y})\right) - \mathsf{H}(X, Y) \right| < \epsilon, \left| -\frac{1}{N}\log\left(\mathsf{p}(\mathbf{x})\right) - \mathsf{H}(X) \right| < \epsilon, \left| -\frac{1}{N}\log\left(\mathsf{p}(\mathbf{y})\right) - \mathsf{H}(Y) \right| < \epsilon \right. \right\}$$

Apply the AEP ideas, this set has the properties

1. $\Pr[((X_1, \ldots, X_N), (Y_1, \ldots, Y_N)) \in \mathcal{A}_\epsilon^N] \to 1$ as $N \to \infty$.
2. $|\mathcal{A}_\epsilon^N| \leq 2^{N(\mathsf{H}(X,Y)+\epsilon)}$

We additionally showed that if $\{\tilde{X}_n\}$ and $\{\tilde{Y}_n\}$ are independent i.i.d. sequences from $\mathsf{p}_{\tilde{X},\tilde{Y}} = \mathsf{p}_{\tilde{X}}\mathsf{p}_{\tilde{Y}} = \mathsf{p}_X\mathsf{p}_Y$, i.e. sharing the same marginal distribution as X and Y, then the probability that these independent sequences lie in the typical set of X_n, Y_n sequences is bounded by

$$\begin{aligned} \Pr\left[(\tilde{X}_1, \ldots, \tilde{X}_N, \tilde{Y}_1, \ldots, \tilde{Y}_N) \in \mathcal{A}_\epsilon^N\right] &= \sum_{(\mathbf{x},\mathbf{y}) \in \mathcal{A}_\epsilon^N} \mathsf{p}_X(\mathbf{x})\mathsf{p}_Y(\mathbf{y}) \leq \sum_{(\mathbf{x},\mathbf{y}) \in \mathcal{A}_\epsilon^N} 2^{-N(\mathsf{H}(X)-\epsilon)}2^{-N(\mathsf{H}(Y)-\epsilon)} \\ &\leq |\mathcal{A}_\epsilon^N| 2^{-N(\mathsf{H}(X)-\epsilon)}2^{-N(\mathsf{H}(Y)-\epsilon)} \leq 2^{-N(\mathsf{I}(X;Y)-3\epsilon)} \end{aligned}$$

where we have introduced the mutual information between two random variables X and Y

$$\mathsf{I}(X;Y) = \mathsf{H}(X) + \mathsf{H}(Y) - \mathsf{H}(X,Y) = \mathsf{H}(X) - \mathsf{H}(X|Y) = \mathsf{H}(Y) - \mathsf{H}(Y|X)$$

which may be interpreted as the reduction in uncertainty of X upon learning Y, or equivalently as the reduction in uncertainty of Y upon learning X.

8 Essence of Channel Coding

A discrete memoryless channel is a device which maps input sequences $\mathbf{x} \in \mathcal{X}^N$ to random output sequences $\mathbf{Y}$ taking values in $\mathcal{Y}^N$ (where $\mathcal{X}$ and $\mathcal{Y}$ are both finite sets) according to the conditional probability distribution

$$\mathsf{p}_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) := \prod_{i=1}^{N} \mathsf{p}_{Y|X}(y_i|x_i)$$

Thus, a discrete memoryless channel can be specified by the input set $\mathcal{X}$, the output set $\mathcal{Y}$, and the conditional distribution $\mathsf{p}_{Y|X}$.

With channel coding, the goal is to provide a means for conveying information over the channel. To do this, one builds a collection of M input sequences $\mathbf{x}_m \in \mathcal{X}^N$, $m \in \{1, \ldots, M\}$, which we call codewords. The collection of codewords $\{\mathbf{x}_m | m \in \{1, \ldots, M\}\}$ is called the codebook. The rate of the code, which indicates the number of bits transmitted per transmitted symbol, is given by $R = \frac{\log_2(M)}{N}$. The transmitter at the input to the channel conveys information by selecting a codeword via its index $m \in \{1, \ldots, M\}$,

then transmits $\mathbf{x}_m$ over the channel. This gives rise to the random observation $\mathbf{Y}$ at the receiver. The goal of the channel decoder is then to guess which codeword was transmitted. When the channel decoder guesses wrong, we say that it commits a sequence error. Perhaps the most important result of channel coding theory is the following:

Define the Shannon channel capacity to be

$$C := \max_{\mathsf{p}_X} \mathsf{I}(Y; X)$$

The quantity I here is called the mutual information between Y and X and is given by

$$\mathsf{I}(Y; X) := \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \mathsf{p}_{X,Y}(x, y) \log \left(\frac{\mathsf{p}_{X,Y}(x, y)}{\mathsf{p}_X(x)\mathsf{p}_Y(y)} \right)$$

Then, for any code rate $R < C$ there exists a codebook for which the probability of sequence error decreases exponentially in the block length N (and thus may be made arbitrarily small through appropriately large selection of N). Furthermore, for any code rate $R > C$, regardless of the code, the probability of sequence error goes to 1 at least exponentially fast in the block length.

We proved this in class using a random codebook and joint typicality decoding. In particular, build the channel code by generating a $2^{NR} \times N$ matrix by sampling its elements i.i.d. from some distribution p_X. The different codewords are different rows in the matrix. The decoder operates by selecting the codeword which is jointly typical (i.e. in the set of jointly typical sequences i.i.d. from $\mathsf{p}_X \mathsf{p}_{Y|X}$) with the received sequence if there is exactly one such codeword, and by declaring an error if not.

By symmetry it suffices to consider the probability of error when the codeword in the first row is transmitted. An error occurs if either the first codeword (row) is not jointly typical with the channel output denoted E_1, or if additionally there is another codeword, e.g. the one corresponding to row j, which is jointly typical with the channel output denoted by E_2^j. By the union bound

$$\Pr[\text{error}] \leq \Pr[E_1] + \sum_{j=2}^{2^{NR}} \Pr[E_2^j]$$

now, by the definition of the set of typical sequences $\Pr[E_1] \leq \epsilon$ for N large enough. Next, since the different codewords are drawn independently from one another when selecting the code, the probability that they are jointly typical with the output of the channel is governed by the property concerning independent sequences lying in the typical set which we discussed at the end of section 7. Thus we have

$$\Pr[\text{error}] \leq \epsilon + \sum_{j=2}^{2^{NR}} 2^{-N(\mathsf{I}(X,Y)-3\epsilon)} \leq \epsilon + 2^{-N(\mathsf{I}(X,Y)-R-\epsilon)}$$

(This is the error averaged over the selection of the codebook, and thus there must be at least one code with lower error.) From this we see that $R < \mathsf{I}(X, Y)$ allows the probability of error to die to zero as the block length $N \to \infty$. Selecting the distribution of X to maximize this upper bound, then, yields the largest region of achievable rates. We will establish that this is an exhaustive characterization of the region of achievable rates by proving a converse in the next lecture.