10-704: Information Processing and Learning **Spring 2012**

Lecture 16: Shannon's Noisy Coding Theorem

Lecturer: Aarti Singh *Scribes: Rafael Stern*

16.1 Defining a Channel

The problem of sending a message through a noisy channel can be illustrated by Figure 16.1.

Figure 16.1: Sending a message through a noisy channel

W represents a message. We will often consider that it is the output of the compressor. The encoder receives W and encodes it to X^n. The encoder puts X^n into the channel and Y^n is the output the decoder receives. Finally, the decoder tries to estimate W through Y^n. The decoder's estimate is denoted by $\widehat{W}$. Our goal is to understand what kind of encodings are such that $\widehat{W}$ is the same as W with high probability and n is as small as possible.

The following notation formalizes this process:

- The possible messages are[1] elements of the set $\{1, \ldots, M\}$.
- $\mathcal{X}$ denotes the alphabet in which the message is encoded. X^n denotes the encoding function, that is, $X^n : \{1, \ldots, M\} \mapsto \mathcal{X}^n$.
- $\mathcal{Y}$ denotes the alphabet in which the output of the channel is written. $\widehat{W}$ denotes the decoding function, that is, $\widehat{W} : \mathcal{Y}^n \mapsto \{1, \ldots, M\}$.
- A channel is defined as $(\mathcal{X}^n, \mathcal{Y}^n, p(y^n|x^n))$. $p(y^n|x^n)$ represents the probability of the channel outputting y^n when x^n is received as an input.

A code is defined by $(M, n, X^n, \widehat{W})$. For a sequence of codes indexed by n, the goal of $\widehat{W}$ being the same as W with high probability is formalized through the following quantities:

- (Conditional Probability of Error) $\lambda_i^{(n)} = P(\widehat{W} \neq i | W = i)$
- (Maximal Probability of Error) $\lambda^{(n)} = \max_{i \in \{1,\ldots,M\}} \lambda_i^{(n)}$.
- (Arithmetic Probability of Error) $p_e^{(n)} = \frac{\sum_{i=1}^{M} \lambda_i^{(n)}}{M}$

[1] The encoder and decoder agree about what are the possible messages and on an index on these messages. Since there is bijective relation between messages and indices, we make no distinction between them.

The rate of a code is defined as $R = \frac{\log_2 M}{n}$. R is the ratio between how many bits of message are transmited and how many bits are used for encoding.

Definition 16.1 *A rate R is achievable if, for all n, there exists a code with $M = 2^{nR}$ s.t. $\lambda^{(n)} \overset{n\to\infty}{\longrightarrow} 0$.*

The **operational capacity** of a channel is the supremum over all achievable rates for that channel.

Let $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ the corresponding output of the channel for X. The **information capacity** of a channel is defined as $C = \max_{p(x)} I(X, Y)$.

Definition 16.2 *The channel $(\mathcal{X}^n, \mathcal{Y}^n, p(y^n|x^n))$ is a Discrete Memoryless Channel (DMC) if:*

1. *(Discrete) $\mathcal{X}, \mathcal{Y}$ are finite sets.*
2. *(Memoryless) $p(y_k|x^k, y^{k-1}) = p(y_k|x_k)$*

Also, unless otherwise stated, we will assume the channel is used with no feedback, i.e. $p(x_k|x_1^{k-1}, y^{k-1}) = p(x_k|x^{k-1})$ for all distribution over $p(x^n)$.

Corollary 16.3 *If $(\mathcal{X}^n, \mathcal{Y}^n, p(y^n|x^n))$ is a DMC without feedback, then $p(y^n|x^n) = \prod_{i=1}^n p(y_i|x_i)$.*

A DMC is completely specified by the transition probability for each individual symbol transmited, $p(y|x)$.

16.2 Jointly Typical Sets

When studying compression, we defined that a sequence is typical if its information content is close to the entropy. For i.i.d. sequences, we proved that typical sequences occur with high probability. In the problem of sending codes through a channel we are concerned with two sequences, the input and output of the channel. In this context, we develop the notion of jointly typical sets.

Definition 16.4 *$A_\epsilon^{(n)}$ is the set of typical sets of length n and tolerance ϵ and is composed of sequences (x^n, y^n) such that:*

1. $|\frac{-\log(p(x^n))}{n} - H(X)| < \epsilon$
2. $|\frac{-\log(p(y^n))}{n} - H(Y)| < \epsilon$
3. $|\frac{-\log(p(x^n, y^n))}{n} - H(X, Y)| < \epsilon$

Lemma 16.5 *If (X_i, Y_i) are i.i.d., the jointly typical sets $A_\epsilon^{(n)}$ satisfy the following properties:*

1. $P(X^n, Y^n) \in A_\epsilon^{(n)} \to 1$ *as* $n \to \infty$
2. $|A_\epsilon^{(n)}| \leq 2^{n(H(X,Y)+\epsilon)}$.
3. $|A_\epsilon^{(n)}| \geq (1-\epsilon)2^{n(H(X,Y)-\epsilon)}$, *for large* n

4. *Let $(\tilde{X}^n, \tilde{Y}^n) \sim p(x^n)p(y^n)$, i.e. $\tilde{X}^n$ and $\tilde{Y}^n$ are independent with distributions same as the marginal distributions $p(x^n)$ and $p(y^n)$ of $p(x^n, y^n)$, then*

$$P((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}) \leq 2^{-n(I(X,Y)-3\epsilon)}$$

Also, for sufficiently large n,

$$P((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}) \geq (1-\epsilon)2^{-n(I(X,Y)+3\epsilon)}$$

Proof: Items 1, 2 and 3 follow analogously to the results regarding typicality in the source coding theorem. For item 4, observe that:

$$P((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}) = \sum_{(x^n,y^n)\in A_\epsilon^{(n)}} p(x^n)p(y^n) \leq |A_\epsilon^{(n)}| \max_{(x^n,y^n)\in A_\epsilon^{(n)}} p(x^n)p(y^n)$$

From item 2, $|A_\epsilon^{(n)}| \leq 2^{n(H(X,Y)+\epsilon)}$. Also observe that for any $(x^n, y^n) \in A_\epsilon^{(n)}$, x^n is typical. Hence, $p(x^n) \leq 2^{-n(H(X)-\epsilon)}$. Similarly, $p(y^n) \leq 2^{-n(H(Y)-\epsilon)}$. Using these inequalities:

$$P((\hat{X}^n, \hat{Y}^n) \in A_\epsilon^{(n)}) \leq 2^{n(H(X,Y)+\epsilon)}2^{-n(H(X)-\epsilon)}2^{-n(H(Y)-\epsilon)} = 2^{-n(I(X;Y)-3\epsilon)}$$

Similarly, using item 3, we get for large enough n:

$$P((\hat{X}^n, \hat{Y}^n) \in A_\epsilon^{(n)}) \geq (1-\epsilon)2^{n(H(X,Y)-\epsilon)}2^{-n(H(X)+\epsilon)}2^{-n(H(Y)+\epsilon)} = (1-\epsilon)2^{-n(I(X;Y)+3\epsilon)}$$

■ A consequence of property 4 is that, for a fixed y^n, we can consider $\sim 2^{nI(X,Y)}$ x^n sequences before we are likely to find another that is jointly typical with y^n. This implies that there are $2^{nI(X,Y)}$ distinguishable input signals x^n. This observation is the key insight that leads to Shannon's noisy channel coding theorem, as discussed next.

16.3 Shannon's Noisy Coding Theorem

Theorem 16.6 *For any DMC, if $R < C$, then R is achievable. Conversely, if $R > C$, it is not achievable.*

Proof: We start proving that, if $R < C$, then R is achievable. In order to do so, it is enough to construct a particular sequence of codes with rate R such that $\lambda^{(n)} \to 0$. Our strategy to complete this goal will be to consider codes which are randomly generated and evaluate the arithmetic error probabilities of decoding these codes using jointy typicality decoding, i.e. decode the channel output as the codeword which is jointly typical with the channel output. Through the arithmetic error probabilities of these randomly generated codewords, we will obtain proof of the existence of a code with maximal probability of error going to zero as $n \to \infty$.

Our proof of achievability will proceed in two steps. We will first show the existence of a sequence of codes that can transmit $M = 2^{nR+1}$ messages (rate $R + \frac{1}{n}$) where the average error $p_e^{(n)} = \frac{\sum_{i=1}^M \lambda_i^{(n)}}{M}$ goes to zero. We will then construct a sequence of codes that transmit $\frac{M}{2} = 2^{nR}$ messages (rate R) where the maximum error λ^n goes to zero.

Consider the random code generated by the following algorithm:

1. Fix $p(x)$ such that $I(X,Y) = C$ and generate $M = 2^{nR+1}$ codewords independently[2] according to $p(x^n) = \prod_{i=1}^{n} p(x_i)$. Call the collection of codewords, $\mathcal{C}$, the codebook.
2. Assign each message W in $\{1,\ldots,M\}$ at random to a codeword $X^n(W)$ in $\mathcal{C}$.
3. Assume the codebook $\mathcal{C}$ and $p(y|x)$ are known beforehand to the decoder.
4. $\hat{W} = c$ for $c \in \{1,\ldots,M\}$ if c is the only message such that $(X^n(c), Y^n)$ are jointly typical. Otherwise, define $\hat{W} = 0$.

From step 2 in the algorithm, observe that the conditional probability of error $\lambda_i^{(n)}$ is the same for all messages $i \in \{1,\ldots,M\}$. Hence, $p_e^{(n)} = \lambda_1^{(n)}$. Here both conditional and arithmetic probability of error are expectations over the random draw of the codewords.

$$\begin{aligned}\lambda_1^{(n)} = P(\hat{W} \neq 1 | W = 1) &= P((X^n(1), Y^n(1)) \notin A_\epsilon^{(n)} \cup (\exists i \neq 1 : (X^n(i), Y^n(1)) \in A_\epsilon^{(n)})) \\ &\leq P((X^n(1), Y^n(1)) \notin A_\epsilon^{(n)}) + P((\exists i \neq 1 : (X^n(i), Y^n(1)) \in A_\epsilon^{(n)}))\end{aligned}$$

From Lemma 16.5, property 1 observe that $P((X^n(1), Y^n(1)) \notin A_\epsilon^{(n)}) \overset{n\to\infty}{\longrightarrow} 0$. Next, by the union bound and symmetry[3] of the code,

$$P(\exists i \neq 1 : (X^n(i), Y^n(1)) \in A_\epsilon^{(n)}) \leq \sum_{i=2}^{2^{nR+1}} P((X^n(i), Y^n(1)) \in A_\epsilon^{(n)}) \leq 2^{nR+1} P((X^n(2), Y^n(1)) \in A_\epsilon^{(n)})$$

Also recall that, by step 1 in the algorithm, $X^n(1)$ is independent of $X^n(2)$. Thus, since $Y^n(1)$ is a function of $X^n(1)$ and randomness introduced by the channel, it is independent of $X_n(2)$. Applying Lemma 16.5, property 4 and also step 1 of the algorithm,

$$2^{nR+1} P((X^n(2), Y^n(1)) \in A_\epsilon^{(n)}) \leq 2^{nR+1} 2^{-n(I(X,Y)-3\epsilon)} = 2^{-n(C-R-3\epsilon-\frac{1}{n})}$$

Putting the last 3 paragraphs together, conclude that:

$$\lambda_1^{(n)} \leq \epsilon + 2^{-n(C-R-3\epsilon-\frac{1}{n})}$$

and if $R < C$, for every $\delta > 0$ there exists n^* such that, if $n > n^*$, $\lambda_1^{(n)} < \delta$. Hence, for $n > n^*(\delta)$:

$$\delta > \lambda_1^{(n)} = p_e^{(n)} = \sum_{\mathcal{C}} P(\mathcal{C}) p_e^{(n)}(\mathcal{C}) \geq \min_{\mathcal{C}} \{p_e^{(n)}(\mathcal{C})\}$$

Conclude[4] that there exists a sequence of codes with rates $R + \frac{1}{n}$ such that $p_e^{(n)} \to 0$.

To control the maximal probability of error, consider a new sequence of codes constructed in the following manner: for each code in the previous sequence, remove the $\frac{M}{2}$ messages with the worst probabilities of error.

[2]Each codeword is independent of the other and also of the randomness introduced by the channel.

[3]From step 1 in the algorithm, $X^n(2)$ is identically distributed to $X^n(i)$, for any i. From step 1 and 2, $X^n(i)$ is independent of $Y^n(j)$, for any, $i \neq j$. Hence, $(X^n(i), Y^n(1))$ is identically distributed to $(X^n(2), Y^n(1))$, for any $i \neq 1$.

[4]For $n^*(\frac{1}{2}^m) = n_m$, choose the code induced by the codebooks $\arg\min_{\mathcal{C}} \{p_e^{(n_m)}(\mathcal{C})\}$. For $n_m < n < n_{m+1}$ keep the same code as the one selected for n_m. This sequence of codes is such that the rates of the codes are $R + \frac{1}{n}$ and $p_e^{(n)} \to 0$.

This trick is called expurgating a code. This new sequence is such that $\lambda^{(n)}$ is at most 2 times the old $p_e^{(n)}$. To verify this, assume the contrary. Then the worst codeword in the new codebook has error greater than $2p_e^{(n)}$ calculated in the old codebook. Observe that the words which were removed from the old codebook all have errors worse than any word in the new codebook and, thus, their errors are larger than $2p_e^{(n)}$. Observe that there are $\frac{M}{2}$ such words. Hence, the arithmetic average of the errors for the M words has to be larger than $p_e^{(n)}$. A contradiction, since the arithmetic average of these errors is precisely $p_e^{(n)}$. Hence, for this new sequence, $\lambda^{(n)} \to 0$. Moreover, the rate of the new code is $(\log(M/2))/n = (\log 2^{nR})/n = R$. Thus, the proof of achievability is complete.

Next, take $R > C$. We will prove that this rate is not achievable, i.e. the arithmetic (and hence the maximal) probability of error is bounded away from 0. For an arbitrary sequence of codes with rate R, observe that $p_e^{(n)} = P(W \neq \hat{W})$ when we take a uniform distribution on W. Recall that, from Fano's Inequality:

$$H(W|Y^n) \leq \log(2^{nR}) P(W \neq \hat{W}) + 1$$

$$P(W \neq \hat{W}) \geq \frac{H(W|Y^n) - 1}{nR}$$

Applying Fano's Inequality for the uniform distribution on W:

$$p_e^{(n)} \geq \frac{H(W) - I(W;Y^n) - 1}{nR} = \frac{nR - I(W;Y^n) - 1}{nR}$$

Next, recall that, by the Data Processing Inequality, $I(W,Y^n) \leq I(X^n;Y^n)$. Also observe that $I(X^n;Y^n) = H(Y^n) - H(Y^n|X^n) \leq \sum_{i=1}^n H(Y_i) - \sum_{i=1}^n H(Y_i|X^n, Y_{i-1}, \ldots, Y_i)$. Since the channel is a DMC, observe that $H(Y_i|X^n, Y_{i-1}, \ldots, Y_i) = H(Y_i|X_i)$. Hence, using all these inequalities, conclude that $I(W,Y^n) \leq \sum_{i=1}^n [H(Y_i) - H(Y_i|X_i)] = \sum_{i=1}^n I(X_i;Y_i) \leq nC$. Hence:

$$p_e^{(n)} \geq \frac{nR - I(W;Y^n) - 1}{nR} \geq \frac{nR - nC - 1}{nR} \overset{n\to\infty}{\longrightarrow} \frac{R - C}{R}$$

Hence, if $R > C$, $p_e^{(n)}$ does not converge to 0. Since $\lambda^{(n)} \geq p_e^{(n)}$, it also doesn't converge to 0. Since the sequence of codes was arbitrary, the rate R is not achievable, which completes the proof.

■

Channel coding theorem promises the existence of block codes that allow us to transmit information at rates below capacity with an arbitrary small probability of error if block length is large enough. However, the search for such optimal yet practical (easily implementable) codes is ongoing. Random codes used by Shannon are not practical since they require exponential look-up tables for encoding and decoding.

Examples of practical codes include block codes such as Hamming code, Reed-Muller codes, Reed-Solomon codes, Goppa codes, BCH (Bose-Chaudhary-Hocquenhem) codes, which encode messages into blocks. More recently, convolution codes have been invented that get pretty close to Shannon's capacity limit. The convolution codes don't encode in blocks, instead they read and transmit bits continuously, where transmitted bits are a linear combination of previous source bits. These codes include Turbo codes, LDPC (Low Density Parity Check) codes, Digital Fountain, codes etc.