Channel Capacity

Discrete Channels

- A discrete channel is defined by the input symbol set $\mathcal{X}$, the probability $p(y|x)$, and the output symbol $\mathcal{Y}$.
- Example
  Noisyless binary channel (Fig 8.2)
- A communication system includes a discrete channel as a sub-system. (Fig 8.1) Here the source are encoded into channel symbols and the received channel symbols are decoded.

Definitions of Channel Capacity

- The information capacity of a discrete channel is

$$C_I = \max_{p(x)} I(X;Y).$$

- The operational capacity of a discrete channel is

$$C_O = \lim_{n\to\infty} \frac{\log M}{n},$$

where M is the number of distinguishable signals for n uses of the channel.

- The two definitions of capacity are equivalent, i.e.,

$$C_I = C_O.$$

Examples of Capacities

- Noisy channel with non-overlapping output $C = 1$.
- Noisy typewriter $C = \log 13$.
- Binary symmetric channel $C = 1 - H(p)$.
- Binary erasure channel $C = 1 - \alpha$.
- (Weakly) symmetric channels

$$C = \log |\mathcal{Y}| - H(\text{row p.m.f.})$$

Properties of Capacity

- Non-negative
- $C \leq \log |\mathcal{X}|$
- $C \leq \log |\mathcal{Y}|$
- A concave function of $p(x)$, so a local maximum is a global maximum.

Channel Coding Picture

- Through one use of a noisy channel, we cannot be sure which X has been sent. How can we transmit C bits *reliably* of information per use of this channel?
- For very large blocks, each channel looks like a noisy typewriter.
- The channel has a subset of inputs that produce essentially disjoint sequences.
- Figure 8.7.

A Few Definitions

- Discrete channel is a 3-tuple $(\mathcal{X}, p(y|x), \mathcal{Y})$.
- The nth extension of a channel $(\mathcal{X}, p(y|x), \mathcal{Y})$ is the channel $(\mathcal{X}^n, p(y^n|x^n), \mathcal{Y}^n)$.
- Memoryless is defined by $p(y_k|x^k, y^{k-1}) = p(y_k|x_k)$, i.e., the probability is not affected by the past transmissions.
- Non-feedback property is defined by $p(x_k|x^{k-1}, y^{k-1}) = p(x_k|x^{k-1})$, i.e., the next symbol to be transmitted is not affected by the received symbols.

Discrete Memoryless Channel Without Feedback

- We will assume that the channel is memoryless and non-feedback unless we state otherwise.
- In this case, the conditional probability of the received symbols given the transmitted symbols is

$$
\begin{aligned}
p(y^n|x^n) &= \frac{p(x^n, y^n)}{p(x^n)} \\
&= \frac{\prod p(x_k|x^{k-1}, y^{k-1})p(y_k|x^k, y^{k-1})}{\prod p(x_k|x^{k-1})} \\
&= \prod_{i=1}^{n} p(y_i|x_i)
\end{aligned}
$$

(M, n) Code

- An (M, n) code for the channel $(\mathcal{X}, p(y|x), \mathcal{Y})$ consists of the following
  - An index set
    $$\mathcal{W} = \{1, 2, \ldots, M\}$$
  - Encoder $\mathcal{W} \to \mathcal{X}^n$
    $$X^n(W) = c(W)$$
  - Decoder $\mathcal{Y}^n \to \mathcal{W}$
    $$\hat{W} = g(Y^n)$$

Probability of Error of An (M, n) Code

- Conditional probability of error given index i was sent

$$\lambda_i = Pr(g(Y^n) \neq i | X^n = X^n(i))$$

- Maximum probability of error

$$\lambda^{(n)} = \max_{i \in \{1,2,\ldots,M\}} \lambda_i$$

- Average probability of error

$$P_e^{(n)} = \frac{1}{M} \sum_{i=1}^{M} \lambda_i$$

The Rate of An (M, n) Code

- The rate of a code is defined by

$$R = \frac{\log M}{n} \text{ bits per transmission.}$$

- A rate R is *achievable* if there exists a sequence of $(\lceil 2^{nR} \rceil, n)$ codes such that $\lambda^{(n)} \to 0$ as $n \to \infty$. That is, if $\lceil 2^{nR} \rceil$ messages are distinguishable on n use of channel, then R is achievable.

- The capacity of a channel is the supremum of all achievable rates.

Jointly Typical Decoding

- Our goal is to show that the capacity of channel is the information capacity (the maximum information $I(X;Y)$ achievable by varying $p(x)$).
- In fact, we will show $R < I(X;Y)$ for a given $p(x)$.
- Given $R < C$, we analyze the $(2^{nR}, n)$ code with
  - a random codebook for encoding.
  - the joint typicality decoding.

Jointly Typical Sequences (1/2)

- The set of jointly typical sequences is

$$
\begin{aligned}
A_\epsilon^{(n)} = \{(x^n, y^n) : & |\frac{-1}{n} \log p(x^n) - H(X)| \leq \epsilon, \\
& |\frac{-1}{n} \log p(y^n) - H(Y)| \leq \epsilon, \\
& |\frac{-1}{n} \log p(x^n, y^n) - H(X, Y)| \leq \epsilon\}
\end{aligned}
$$

- We will assume the i.i.d property

$$
p(x^n, y^n) = \prod_{i=1}^{n} p(x_i, y_i)
$$

Jointly Typical Sequences (2/2)

The joint typical set $A_\epsilon^{(n)}$ has the following properties.

- $Pr((X^n, Y^n) \in A_\epsilon^{(n)}) \to 1$
- $(1-\epsilon)2^{n(H(X,Y)-\epsilon)} \leq |A_\epsilon^{(n)}| \leq 2^{n(H(X,Y)+\epsilon)}$
- If $\tilde{X}^n, \tilde{Y}^n$ are drawn independently, then

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

The probability of independently chosen X^n, Y^n of being jointly typical is one part in 2^{nI}, suggesting that there are about 2^{nI} distinguishable signals of X^n.

Main Theorems

Given a channel of capacity C,

- (The channel coding theorem) All rates below C are achievable. That is, if $R < C$, then there exists a sequence of $(2^{nR}, n)$ code such that $\lambda^{(n)} \to 0$ as $n \to \infty$.
- (Converse to the channel coding theorem) If R is achievable, then $R \leq C$. That is, if $\lambda^{(n)} \to 0$ for a $(2^{nR}, n)$ code, then $R \leq C$.

Proof of Channel Coding Theorem

- To show R is achievable, a construction of $(M = 2^{nR}, n)$ code with vanishing probability of error suffices. Our construction is
  - Generate the codebook entries according to $p(x)$. The probability of a codebook $\mathcal{C}$ being generated is

$$P(\mathcal{C}) = \prod_{w=1}^{M} \prod_{i=1}^{n} p(x_i(w)).$$

  - A message W is chosen according to the uniform distribution ($P(w) = \frac{1}{M} \; \forall w$), and the corresponding codeword $X^n(W)$ is sent over the channel

- The receiver receives Y^n, drawn from

$$P(y^n|x^n) = \prod_{i=1}^{n} p(y_i|x_i)$$

- The receiver decodes $\hat{W}$ if
  * $(X^n(\hat{W}), Y^n)$ is jointly typical
  * No other codeword is jointly typical with Y^n.

Proof of Channel Coding Theorem

- The probability of error averaged over all codebooks is

$$P(E) = \sum_{\mathcal{C}} P(\mathcal{C}) P_e^{(n)}(\mathcal{C}) = \sum_{\mathcal{C}} P(\mathcal{C}) \frac{1}{M} \sum_{w=1}^{M} \lambda_w(\mathcal{C})$$

$$= \frac{1}{M} \sum_{w=1}^{M} \sum_{\mathcal{C}} P(\mathcal{C}) \lambda_w(\mathcal{C}) = \sum_{\mathcal{C}} P(\mathcal{C}) \lambda_1(\mathcal{C}),$$

  where the last equality follows from the symmetry of the codebook generation.

- Given $W = 1$, an error occurs when
  - Y^n is not jointly typical with $X^n(1)$;
  - Y^n is jointly typical with $X^n(i), i \neq 1$

Proof of Channel Coding Theorem

- Defining E_i to be the event that $\{X^n(i), Y^n\}$ is jointly typical, then

$$
\begin{aligned}
P(E|W=1) &= P(E_1^c \cup E_2 \cup \cdots \cup E_M) \\
&\leq P(E_1^c) + \sum_{i \neq 1} P(E_i) \\
&= P(E_1^c) + (M-1)2^{-n(I(X;Y)-3\epsilon)} \\
&\leq \epsilon + (2^{nR}-1)2^{-n(I(X;Y)-3\epsilon)} \\
&\to 0, \text{ if } R < I(X;Y)
\end{aligned}
$$

Proof of Channel Coding Theorem

- Choose $p(x)$ to be $p^*(x)$, the distribution that achieves C, then the condition $R < I$ can be replaced by $R < C$.
- Since the average probability of error, say δ, of all codebooks is small, there is a codebook $\mathcal{C}^*$ with an error probability no greater than δ.
- Throw away half of the codewords in $\mathcal{C}^*$ (those with higher conditional probabilities of error). The maximal probability of error of the remaining codebook cannot be greater than 2δ, otherwise those thrown away alone makes the original codebook with error greater then δ, a contradiction. The new code has rate $R - \frac{1}{n}$ and $\lambda^{(n)} \to 0$.

Proof of the Converse

- If a rate R is achievable ($\lambda^{(n)} \to 0$), then $R \leq C$.
- Recall that the Fano's inequality gives a lower bound in the probability of error given the conditional entropy

$$H(P_e) + P_e \log(|\mathcal{X}| - 1) \geq H(X|Y).$$

- Define the probability of error (this is the average probability of error when W is uniform)

$$P_e^{(n)} = Pr(\hat{W} \neq W) = Pr(E = 1),$$

where

$$E = \begin{cases} 1, & \text{if } \hat{W} \neq W, \\ 0, & \text{if } \hat{W} = W. \end{cases}$$

Proof of the Converse

- Apply the Fano's inequality, identifying Y as Y^n and X as W

$$H(W|Y^n) \leq H(P_e^{(n)}) + P_e^{(n)} \log(|\mathcal{W}| - 1) \leq 1 + P_e^{(n)} nR$$

- Furthermore, since $X^n(W)$ is a function of W

$$\begin{aligned} H(X^n, W|Y^n) &= H(X^n|Y^n) + H(W|X^n, Y^n) \\ &= H(W|Y^n) + H(X^n|W, Y^n) \end{aligned}$$

- Putting together, we have

$$H(X^n|Y^n) \leq H(W|Y^n) \leq 1 + P_e^{(n)} nR$$

Proof of the Converse

- For the mutual information between X^n and Y^n

$$
\begin{aligned}
I(X^n;Y^n) &= H(Y^n) - H(Y^n|X^n) \\
&= H(Y^n) - \sum H(Y_i|Y_1,\ldots,Y_{i-1},X^n) \\
&= H(Y^n) - \sum H(Y_i|X_i) \\
&\le \sum H(Y_i) - \sum H(Y_i|X_i) \\
&= nI(X;Y) \le nC.
\end{aligned}
$$

Proof of the Converse

- $\lambda^{(n)} \to 0$ implies $P_e^{(n)} \to 0$.
- Let W be drawn uniformly from $\{1, \ldots, M = 2^{nR}\}$. Then

$$\begin{aligned} nR = H(W) &= H(W|Y^n) + I(W;Y^n) \\ &\leq H(W|Y^n) + I(X^n(W);Y^n) \\ &\leq 1 + P_e^{(n)} nR + I(X^n(W);Y^n) \\ &\leq 1 + P_e^{(n)} nR + nC. \end{aligned}$$

$$R \leq \frac{1}{n} + P_e^{(n)} R + C$$

Hence $R \leq C$.

Lower Bound on the Probability of Error

- Re-writing the previous equation

$$P_e^{(n)} \geq 1 - \frac{C}{R} - \frac{1}{nR}$$

- If $R > C$, the probability of error is bounded away from 0!
- It can be shown that if $R > C$, $P_e^{(n)} \to 1$ exponentially. Thus C is a clear dividing point.

Feedback Code

- What is the maximum achievable rate (capacity) with feedback? Surprisingly, feedback does not increase the capacity of a channel.
- With a feedback code, the encoder function is $X_i(W, Y^{i-1})$, since the symbol to be transmitted can depend on feedback Y^{i-1}.
- The decoder is still $g(Y^n)$, depending only on Y^n.

- Let W be drawn uniformly from $\{1, \ldots, M = 2^{nR}\}$.

$$\begin{aligned}
nR = H(W) &= H(W|Y^n) + I(W;Y^n) \\
&\leq 1 + P_e^{(n)} nR + I(W;Y^n) \\
&= 1 + P_e^{(n)} nR + H(Y^n) - H(Y^n|W) \\
&= 1 + P_e^{(n)} nR + H(Y^n) - \sum H(Y_i|W, Y^{i-1}) \\
&= 1 + P_e^{(n)} nR + H(Y^n) - \sum H(Y_i|W, Y^{i-1}, X_i) \\
&= 1 + P_e^{(n)} nR + H(Y^n) - \sum H(Y_i|X_i) \\
&\leq 1 + P_e^{(n)} nR + \sum I(X_i;Y_i) \leq 1 + P_e^{(n)} nR + nC
\end{aligned}$$

Hence $R \leq C$.

Zero-Error Codes

- A special case that $P_e^{(n)} \to 0$ is $\lambda_i = 0\ \forall i$. If a code has zero probability of error, then $H(W|Y^n) = 0$, since $W = \hat{W} = g(Y^n)$ is determined by Y^n.
- Let W be uniformly distributed over the index set, then

$$\begin{aligned} nR = H(W) &= H(W|Y^n) + I(W;Y^n) = I(W;Y^n) \\ &\leq I(X^n;Y^n) = H(Y^n) - H(Y^n|X^n) \\ &\leq \sum_i H(Y_i) - \sum_i H(Y_i|X_i) \leq \sum_i I(X_i;Y_i) \leq nC. \end{aligned}$$

So $R \leq C$.

The Joint Source Channel Coding Theorem

- We have seen two results
  - Data compression: the optimal codeword length per source symbol $R > H$
  - Channel coding: the maximum achievable rate $R < C$
- Let $\{V_1, V_2, \dots\}$ be a finite-alphabet stochastic process satisfying AEP with entropy rate $H(\mathcal{V})$. We can encode those symbol sequences, with $X^{n'}(V^n)$, that are in the typical set, requiring $H(\mathcal{V}) + \epsilon$ bits per symbol. The $X^{n'}$ are transmitted over a channel with capacity C. The decoder is $\hat{V}^n = g(Y^{n'})$.

The Joint Source Channel Coding Theorem

- Let $P_e^{(n)} = Pr(V^n \neq \hat{V}^n)$.
- (Theorem) If $H(\mathcal{V}) < C$, then there exists a source channel code such that $P_e^{(n)} \to 0$.
  - The number of elements in the typical set of $V_1, \dots, V_n$ is less then $2^{nH(\mathcal{V})}$. Therefore we can enumerate this set and use the corresponding index $\{1, 2, \dots, 2^{nH(\mathcal{V})}\}$ for the channel coding. To conform to the notation earlier, here $M = 2^{nH(\mathcal{V})}$. From the channel coding theorem, if $H(\mathcal{V}) \leq C$, then there exists an (M, n) code with $P_e^{(n)} \to 0$.

The Joint Source Channel Coding Theorem

- Conversely, if $P_e^{(n)} \to 0$ then $H(\mathcal{V}) \leq C$.
  - From Fano's inequality

$$H(V^n|\hat{V}^n) \leq 1 + P_e^{(n)} \log |\mathcal{V}|^n,$$

$$\begin{aligned} H(\mathcal{V}) &\leq \frac{H(V^n)}{n} = \frac{1}{n} H(V^n|\hat{V}^n) + \frac{1}{n} I(V^n;\hat{V}^n) \\ &\leq \frac{1}{n}(1 + P_e^{(n)} n \log |\mathcal{V}|) + \frac{1}{n} I(V^n;\hat{V}^n) \\ &\leq \frac{1}{n}(1 + P_e^{(n)} n \log |\mathcal{V}|) + \frac{1}{n} I(X^{n'};Y^{n'}) \\ &\leq \frac{1}{n} + P_e^{(n)} \log |\mathcal{V}| + C. \end{aligned}$$