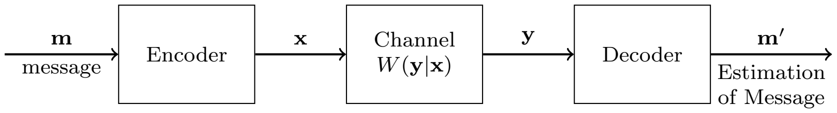
Chapter 6

Channel Coding Theorem

Parts of the discussion refer to [**yeung08**, **cover06**].

6.1 Example of Channels

Shannon's View of Communication System



Definition 6.1 — Discrete channel. A (discrete) channel consists of an input alphabet $\mathcal{X}$, an output alphabet $\mathcal{Y}$ and a probability transition matrix $W : \mathcal{X} \to \mathcal{Y}$. Here $W(y|x)$ tells the probability of outputting y when the input is x.

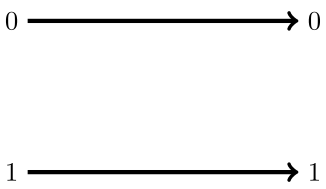
- The input is a random variable $X \sim p(x)$ on $\mathcal{X}$.
- The output Y will be the random variable on $\mathcal{Y}$ such that (X, Y) follows the joint distribution

$$(p \cdot W)(\mathbf{x}, \mathbf{y}) \triangleq p(\mathbf{x})W(\mathbf{y}|\mathbf{x}).$$

Definition 6.2 The "channel capacity" of a discrete *memoryless* channel is defined as

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

Noiseless Binary Channel

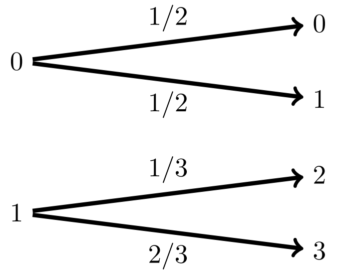


- $W(0|0) = W(1|1) = 1$.
- $C = 1$.

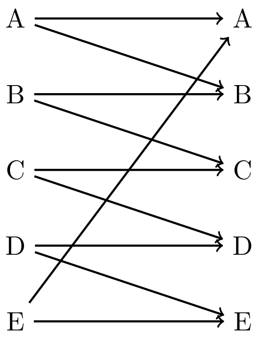
It is simple to achieve 1 bit per use for this channel using a block length 1 code.

Nonoverlapping Outputs

This example shows that the uncertainty in output is not an issue if the possible outputs of different inputs are nonoverlapping.



Noisy typerwriter



- $W(x|x) = 1/2$.

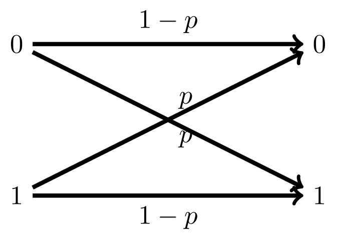
The capacity of the channel is

$$
\begin{aligned}
C &= \max_{p(x)} H(Y) - H(Y|X) \\
&= \max_{p(x)} H(Y) - \sum_{x\in\mathcal{X}} H(Y|X=x)p(x) \\
&= \max_{p(x)} H(Y) - 1 \\
&= \log_2 5 - 1 = \log_2 2.5.
\end{aligned}
$$

This is an example that block length 1 code cannot achieve C.

Using block length 2, we can use for example the codebook $\{AB, AD, CA, CC, EE\}$, which has the zero-error rate $\frac{1}{2}\log_2 5 = \log_2\sqrt{5}$. It turns out that the zero error capacity of the channel is $\log_2\sqrt{5}$.

Binary Symmetric

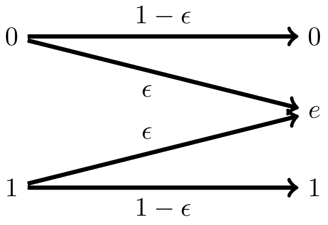


The capacity of the channel is

$$
\begin{aligned}
C &= \max_{p(x)} H(Y) - H(Y|X) \\
&= \max_{p(x)} H(Y) - \sum_{x\in\mathcal{X}} H(Y|X=x)p(x) \\
&= \max_{p(x)} H(Y) - h(p) \\
&= 1 - h(p).
\end{aligned}
$$

In the previous examples, we can have zero error codes. For the this example, however, no code with a positive rate exists.

Binary Erasure



The capacity of the channel is

$$
\begin{aligned}
C &= \max_{p(x)} H(X) - H(X|Y) \\
&= \max_{p(x)} H(X) - \sum_{y\in\dagger} H(X|Y=y)p(y) \\
&= \max_{p(x)} H(X) - p_Y(e)H(X|Y=e) \\
&= \max_{p(x)} H(X) - \epsilon H(X) \\
&= 1 - \epsilon.
\end{aligned}
$$

For binary erasure channel, feedback can help to design a simple zero-error solution. But feedback cannot increase the achievable rate, which will be proved later.

Properties of Channel Capacity

- $C \geq 0$.
- $C \leq \log|\mathcal{X}|$.
- $C \leq \log|\mathcal{Y}|$.
- $I(X;Y)$ is a continuous function of $p(x)$.
- $I(X;Y)$ is a concave function of $p(x)$.

6.2 Channel Coding Theorem

Discrete Memoryless Channel

Definition 6.3 A sequence of channels $\{W_n : \mathcal{X}^n \to \mathcal{Y}^n\}_{n=1}^{\infty}$ is called a *discrete memoryless channel* (DMC) with transition probability matrix W if $W_n(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{n} W(y_i|x_i)$. A DMC is denoted by $\{W : \mathcal{X} \to \mathcal{Y}\}$ or $\{W\}$.

The input X_i of the channel at the ith use usually further depend on other variables of the communication system, e.g., the message for transmit and the feedback from the previous outputs. The channel law of DMC requires that the distribution of the output Y_i is independent of the variables that affect X_i conditioning on X_i, since otherwise, the conditional dependence implies a different channel law.

Channel Codes

Definition 6.4 An (n, M) code for a DMC $\{W : \mathcal{X} \to \mathcal{Y}\}$ consists of an encoding function

$$f : \{1, 2, \dots, M\} \to \mathcal{X}^n$$

and a decoding function

$$\varphi : \mathcal{Y}^n \to \{1, 2, \dots, M\}.$$

The sequence $f(i) \in \mathcal{X}^n$ is called a *codeword* and the set $\{f(i) : i = 1, \dots, M\}$ is called the *codebook*.

Error Probabilities

- Let

$$\begin{aligned}\lambda_i &= \sum_{\mathbf{y} \in \mathcal{Y} : \varphi(\mathbf{y}) \neq i} W_n(\mathbf{y}|f(i)) \\ &= 1 - W_n(\varphi^{-1}(i)|f(i)).\end{aligned}$$

- The *maximal probability of error* of an (n, M) code is

$$\lambda_{\max} = \max_i \lambda_i.$$

- The *average probability of error* of an (n, M) code is

$$P_e = \frac{1}{M} \sum_i \lambda_i.$$

Achievable Rates

- The *rate* of an (n, M) code is

$$\frac{\log M}{n} \quad \text{bits per (channel) use.}$$

- A rate R is said to be *achievable* for a DMC if for any $\epsilon > 0$ and any sufficiently large n, there exists an (n, M) code such that

$$\frac{\log M}{n} > R - \epsilon \quad \text{and} \quad \lambda_{\max} < \epsilon.$$

Channel Coding Theorem

Theorem 6.1 — **Channel coding theorem.** A rate R is achievable for a DMC iff $R \leq C \triangleq \max_p I(X; Y)$.

R

- If $R \leq C$ there exist codes that achieve rate R, which is also called the achievability.
- If $R > C$ there exits no codes with rate close to R while the error is arbitrarily small, which is also called the converse.

Examples

- Noiseless channels
- Noisy typewriter
- Binary symmetric channel
- Binary erasure channel

6.3 Strong Typicality

Strongly Typical Set

Definition 6.5 For any distribution p on $\mathcal{X}$, the (strongly) typical set T_δ^n is the set of sequences $\mathbf{x} \in \mathcal{X}^n$ s.t.

$$\sum_{a \in \mathcal{X}} |P_{\mathbf{x}}(a) - p(a)| \leq \delta$$

and $P_{\mathbf{x}}(a) = 0$ whenever $p(a) = 0$.

Stronge AEP

Theorem 6.2 For random variable $X \sim p$, there exists $\eta > 0$ such that $\eta \to 0$ as $\delta \to 0$ and the following hold:

1. If $\mathbf{x} \in T_\delta^n$, then
$$2^{-n(H(X)+\eta)} \leq p(\mathbf{x}) \leq 2^{-n(H(X)-\eta)}.$$
2. For n sufficiently large,
$$\Pr\{X^n \in T_\delta^n\} > 1 - \delta.$$
3. For n sufficiently large,
$$(1-\delta)2^{n(H(X)-\eta)} \leq |T_\delta^n| \leq 2^{n(H(X)+\eta)}.$$

Proof of Strong AEP 1

- We can write $p(\mathbf{x}) = 2^{n \sum_{a \in \mathcal{X}} P_{\mathbf{x}}(a) \log p(a)}$.
- Using $\sum_a |P_{\mathbf{x}}(a) - p(a)| \leq \delta$, we have $|P_{\mathbf{x}}(a) - p(a)| \leq \delta$ for all a.
- Hence,
$$2^{n \sum_{a \in \mathcal{X}: p(a)>0} (p(a)+\delta) \log p(a)} \leq p(\mathbf{x}) \leq 2^{n \sum_{a \in \mathcal{X}: p(a)>0} (p(a)-\delta) \log p(a)}.$$
- The proof is completed by $\eta = -\delta \sum_{a \in \mathcal{X}: p(a)>0} \log p(a)$.

Proof of Strong AEP 2

For $a \in \mathcal{X}$, $N(a|X^n)$ has a binomial distribution with expectation $n \cdot p(a)$ and variance $n \cdot p(a)(1 - p(a)) \leq n/4$. Thus by Chebyshev's inequality for any a with $p(a) > 0$

$$\Pr\left\{|P_{X^n}(a) - p(a)| > \frac{\delta}{|\mathcal{X}|}\right\} \leq \frac{|\mathcal{X}|^2}{4n\delta^2}.$$

Then

$$\begin{aligned}
1 - \Pr\{X^n \in T_{[X]\delta}^n\} &= \Pr\left\{\sum_a |P_{X^n}(a) - p(a)| > \delta\right\} \\
&\leq \Pr\left\{|P_{X^n}(a) - p(a)| > \frac{\delta}{|\mathcal{X}|} \text{ for some } a\right\} \\
&\leq \sum_{a: p(a)>0} \Pr\left\{|P_{X^n}(a) - p(a)| > \frac{\delta}{|\mathcal{X}|}\right\} \\
&\leq \sum_{a: p(a)>0} \frac{|\mathcal{X}|^2}{4n\delta^2} \\
&< \delta.
\end{aligned}$$

Strong AEP 3 can be proved by 1 and 2.

Jointly Typical Set

Definition 6.6 — Jointly typical set. For $(X, Y) \sim p(x, y)$, the (strongly) jointly typical set T_δ^n is the set of $(\mathbf{x},\mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$ such that

$$\sum_{a,b} |P_{\mathbf{x},\mathbf{y}}(a,b) - p(a,b)| \le \delta$$

and $P_{\mathbf{x},\mathbf{y}}(a,b) = 0$ whenever $p(a,b) = 0$.

R

- If $(\mathbf{x},\mathbf{y}) \in T_\delta^n(X,Y)$, then $\mathbf{x} \in T_\delta^n(X)$ and $\mathbf{y} \in T_\delta^n(Y)$.
- For $\mathbf{x} \in T_\delta^n(X)$ and $\mathbf{y} \in T_\delta^n(Y)$, we may not have $(\mathbf{x},\mathbf{y}) \in T_\delta^n(X,Y)$.
- Let $Y = f(X)$. If $\mathbf{x} \in T_\delta^n(X)$ and $y_i = f(x_i)$, then $(\mathbf{x},\mathbf{y}) \in T_\delta^n(X,Y)$, and hence $\mathbf{y} \in T_\delta^n(Y)$.

Strong Joint AEP

Applying the strong AEP theorem on (X,Y)

- $p(\mathbf{x},\mathbf{y}) \approx 2^{-nH(X,Y)}$.
- For n sufficiently large, $|T_\delta^n(X,Y)| \approx 2^{nH(X,Y)}$, and
- $p(T_\delta^n(X,Y)) > 1-\delta$.

6.4 Achievability

We prove that C is achievable for a DMC $\{W\}$.

Outline of Achievability Proof

1. Generate a random code with rate close to $I(X;Y)$.
2. Define a jointly typical decoding algorithm.
3. Evaluate the expected P_e of all the codes in the ensemble.
4. Last enhance the code so that $\lambda_{\max} < \epsilon$.

Random Code Generation

- Fix $p(x)$ and $\epsilon > 0$.
- Let M be an *even* integer such that

$$I(X;Y) - \frac{\epsilon}{2} < \frac{\log M}{n} < I(X;Y) - \frac{\epsilon}{4},$$

  where $(X,Y) \sim p \cdot W$.
- Generate a codebook $\mathcal{C}$ of M codewords independently according to the distribution $p(\mathbf{x}) = \prod_i p(x_i)$. Let

$$\mathcal{C} = \{\mathbf{X}_1, \ldots, \mathbf{X}_M\}.$$

- $\mathbf{X}_i \sim p(\mathbf{x})$ are independent
- Assume that both the sender and the receiver know the instance of $\mathcal{C}$ to use.

Jointly Typical Decoding

- The first message $\mathbf{X}_1$ is transmitted, and the channel output is $\mathbf{Y}$.
- $(\mathbf{X}_1, \mathbf{Y}) \sim p(\mathbf{x})W_n(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^n p(x_i)W(y_i|x_i)$.
- Jointly typical decoding:
  - The sequence $\mathbf{Y}$ is decoded to the message k if $(\mathbf{X}_k, \mathbf{Y}) \in T_{[XY]\delta}^n$, where δ is a positive quantity to be specified later.
  - If no such message exists or if there is more than one such a message, an error is declared.

Calculate error probability. Note that λ_1 is related to the code. Instead of calculating λ_1 of a single code, we calculate the average performance, i.e., $\mathbb{E}\lambda_1(\mathcal{C})$.

Average Error Probability I

- λ_1 is a function of the code $\mathcal{C}$.
- Let $E(\mathbf{x}) = \{(\mathbf{x}, \mathbf{Y}) \in T^n_{[XY]\delta}\}$.
- We have

$$\begin{aligned}\mathbb{E}[\lambda_1] &= \Pr\{E^c(\mathbf{X}_1) \cup E(\mathbf{X}_2) \cup \cdots \cup E(\mathbf{X}_M)\} \\ &\le \Pr\{E^c(\mathbf{X}_1)\} + \sum_{k=2}^{M} \Pr\{E(\mathbf{X}_k)\}.\end{aligned}$$

- Since $(\mathbf{X}_1, \mathbf{Y}) \sim \prod_{i=1}^n p(x_i)W(y_i|x_i)$, by the property of (strongly) typical sets, for sufficiently large n,

$$\Pr\{E^c(\mathbf{X}_1)\} = \Pr\{(\mathbf{X}_1, \mathbf{Y}) \notin T^n_{[XY]\delta}\} \le \delta.$$

Average Error Probability II

- For $k > 1$, $\mathbf{X}_k$ and $\mathbf{Y}$ are independent.
- We have that for $k > 1$,

$$\begin{aligned}\Pr\{E(\mathbf{X}_k)\} &= \Pr\left\{(\mathbf{X}_k, \mathbf{Y}) \in T^n_{[XY]\delta}\right\} \\ &= \sum_{(\mathbf{x},\mathbf{y}) \in T^n_{[XY]\delta}} p(\mathbf{x}) p_{\mathbf{Y}}(\mathbf{y}) \\ &\le 2^{-n(H(X)-\eta_1)} 2^{-n(H(Y)-\eta_2)} \left|T^n_{[XY]\delta}\right| \\ &\le 2^{-n(H(X)-\eta_1)} 2^{-n(H(Y)-\eta_2)} 2^{n(H(X,Y)+\eta_3)} \\ &= 2^{-n(I(X;Y)-\tau)},\end{aligned}$$

where $\tau = \eta_1 + \eta_2 + \eta_3 \to 0$ as $\delta \to 0$.

Average Error Probability II

- For sufficiently large n

$$\begin{aligned}\mathbb{E}[\lambda_1] &\le \delta + \sum_{k=2}^{M} 2^{-n(I(X;Y)-\tau)} \\ &\le \delta + M 2^{-n(I(X;Y)-\tau)} \\ &< \delta + 2^{-n(\epsilon/4-\tau)}.\end{aligned}$$

- Note that $\tau \to 0$ as $\delta \to 0$.
- Let δ be sufficiently small such that $\delta < \epsilon/3$ and $\tau < \epsilon/4$.
- For n sufficiently large,

$$\mathbb{E}[\lambda_1] < \epsilon/2.$$

Existence: Average Error Probability

- Since the error probabilities of all codewords follow the same calculation, we have

$$\mathbb{E}P_e(\mathcal{C}) = \mathbb{E}\frac{1}{M}\sum_i \lambda_i(\mathcal{C}) = \frac{1}{M}\sum_i \mathbb{E}\lambda_i(\mathcal{C}) < \epsilon/2.$$

- Therefore, there exits at least one codebook $\mathbb{C}$ such that

$$P_e(\mathbb{C}) < \epsilon/2.$$

Existence: Maximal Error Probability

- Let $\mathbb{C}^*$ be the subset of $\mathbb{C}$ with the best half codewords (in terms of λ_i).
- The maximal error probability of the codewords in $\mathbb{C}^*$ is less than ϵ.
- The rate of $\mathbb{C}^*$ is
$$\frac{1}{n}\log\frac{M}{2} > I(X;Y) - \frac{\epsilon}{2} - \frac{1}{n} > I(X;Y) - \epsilon$$
when n is sufficiently large.

Finally, by letting $p(x)$ be the distribution that achieves the channel capacity, i.e., $I(X;Y) = C$, we prove that C is achievable.

6.5 Converse for Zero-Error Codes

Zero-Error codes

- Suppose we have a $(n, 2^{nR})$ code (f, ϕ) with $\lambda_{\max} = 0$.
- Let U be the uniform distribution on the message set $\{1, 2, \dots, 2^{nR}\}$.
- Note that $U \to \mathbf{X} \to \mathbf{Y} \to \hat{U}$ forms a markov chain, where $\mathbf{X} = f(U)$, $\hat{U} = U = \phi(\mathbf{Y})$.
- We can write

$$\begin{aligned} nR = H(U) &= H(U|\mathbf{Y}) + I(U;\mathbf{Y}) \\ &= I(U;\mathbf{Y}) \\ &\le I(\mathbf{X};\mathbf{Y}) \\ &\le \sum_{i=1}^{n} I(X_i;Y_i) \\ &\le nC. \end{aligned}$$

6.6 Fano's Inequality

Suppose that we know a random variable Y and we wish to guess the value of a correlated random variable X. Fano's inequality relates the probability of error in guessing the random variable X to its conditional entropy $H(X|Y)$. If X is a function of Y, which is equivalent to $H(X|Y) = 0$, the guessing has zero error. In general, we hope to estimate X with low probability of error only if the conditional entropy $H(X|Y)$ is small.

Fano's Inequality

Lemma 6.3 For random variables X and Y with the same alphabet $\mathcal{X}$,

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

where $P_e = \Pr\{X \ne Y\}$.

- If X is a function of Y, which is equivalent to $H(X|Y) = 0$, the guessing has zero error.
- We hope to estimate X with low probability of error only if the conditional entropy $H(X|Y)$ is small.

Proof of Fano's Inequality

- Define random variable Z with $Z = 0$ if $X = Y$ and $Z = 1$ otherwise.
- Then,

$$\begin{aligned} H(X|Y) &= H(X|Y) + H(Z|X,Y) \\ &= H(X,Z|Y) \\ &= H(Z|Y) + H(X|Z,Y) \\ &\le H(Z) + H(X|Z,Y) \\ &= H(P_e) + H(X|Z,Y). \end{aligned}$$

- $H(X|Y, Z = 0) = 0$, and $H(X|Y, Z = 1) \le \log(|\mathcal{X}| - 1)$.

6.7 Converse

A Channel Code

- Let R be an achievable rate.
- Consider an n-length code (f, φ) such that $\frac{1}{n}\log M > R - \epsilon$ and $\lambda_{\max} < \epsilon$.
- Let U be the uniform distributed random variable over the message set $\{1, 2, \dots, M\}$.
- The codeword we transmit for U is the random variable $\mathbf{X} = f(U)$.
- Let $\mathbf{Y}$ be the output of the channel for input $\mathbf{X}$, i.e., $(\mathbf{X}, \mathbf{Y}) \sim p_{\mathbf{X}}(\mathbf{x})W_n(\mathbf{y}|\mathbf{x})$.
- Let $\hat{U} = \varphi(\mathbf{Y})$.
- We have a Markov chain
$$U \to \mathbf{X} \to \mathbf{Y} \to \hat{U}.$$

For a detailed justification of the above Markov chain, see [**yeung08**].

Bound on $I(\mathbf{X}; \mathbf{Y})$

- Since the channel is memoryless,
$$\begin{aligned} H(\mathbf{Y}|\mathbf{X}) &= \sum_{\mathbf{x}} p_{\mathbf{X}}(\mathbf{x}) H(\mathbf{Y}|\mathbf{X} = \mathbf{x}) \\ &= \sum_{\mathbf{x}} p_{\mathbf{X}}(\mathbf{x}) \sum_{i=1}^{n} H(Y_i|X_i = x_i) = \sum_{i=1}^{n} H(Y_i|X_i), \end{aligned}$$
- Hence,
$$\begin{aligned} I(\mathbf{X}; \mathbf{Y}) = H(\mathbf{Y}) - H(\mathbf{Y}|\mathbf{X}) &= H(\mathbf{Y}) - \sum_{i=1}^{n} H(Y_i|X_i) \\ &\le \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) \\ &= \sum_{i=1}^{n} I(X_i; Y_i) \le nC. \end{aligned}$$

Converse

- Due to Fano's inequality and the bound on $I(\mathbf{X}; \mathbf{Y})$,
$$\begin{aligned} \log M = H(U) &= H(U|\hat{U}) + I(U; \hat{U}) \\ &\le H(U|\hat{U}) + I(\mathbf{X}; \mathbf{Y}) \\ &\le 1 + P_e \log(M - 1) + nC \\ &< 1 + \epsilon \log M + nC, \end{aligned}$$
- which implies
$$R - \epsilon < \frac{1}{n}\log M < \frac{1}{1-\epsilon}\left(\frac{1}{n} + C\right).$$
- Since for an achievable rate R, we require the above inequality holds for any $\epsilon > 0$ and all sufficiently large n, we conclude that
$$R \le C.$$

Another application of Fano's inequality (not included in this course).

Theorem 6.4 Consider a DMS $\{X_i\}_{i=1}^{\infty}$ with alphabet $\mathcal{X}$. For any variable-length code with $P_e \le \epsilon < 1/2$,
$$L \ge H_D(X) - \epsilon \log_D |\mathcal{X}| - \frac{H_D(\epsilon) + \log_D(dk)}{k}. \tag{6.1}$$
where $d = e \log_D |\mathcal{X}|$.

Proof. We bound $H(f(X^k))$ from above in terms of L. Consider any sequence $Y_1, \ldots, Y_N$ of $\mathcal{Y}$ valued random variables of random length N. Then,

$$\begin{aligned} H(N, Y_1, \ldots, Y_N) &= H(Y_1, \ldots, Y_N|N) + H(N) && (6.2)\\ &= \sum_n \Pr\{N = n\} H(Y_1, \ldots, Y_n|N = n) + H(N). && (6.3) \end{aligned}$$

Here

$$H(Y_1, \ldots, Y_n|N = n) \le \sum_i H(Y_i|N = n) \le n \log |\mathcal{Y}|, \tag{6.4}$$

and (ref. Problem 10, HW2)

$$H(N) < \log(e\mathbb{E}N). \tag{6.5}$$

Then,

$$H(N, Y_1, \ldots, Y_N) < \mathbb{E}N \log |\mathcal{Y}| + \log(e\mathbb{E}N). \tag{6.6}$$

Applying the above inequality to $Y_1, \ldots, Y_n = f(X^k)$, one may assume that $\mathbb{E}N < dk/e$ since otherwise $L = \mathbb{E}N/K \ge d/e = \log_D |\mathcal{X}|$ and the theorem would automatically hold. Thus,

$$H(f(X^k)) < \mathbb{E}N \log D + \log(dk) = kL \log D + \log(dk). \tag{6.7}$$

Note by Fano's inequality

$$H(X^k|\varphi(f(X^k))) \le P_e \log(|\mathcal{X}|^k - 1) + H(P_e) \le \epsilon k \log |\mathcal{X}| + H(\epsilon). \tag{6.8}$$

This yields

$$\begin{aligned} kH(X) &= H(X^k) && (6.9)\\ &= H(X^k|\varphi(f(X^k))) + H(\varphi(f(X^k))) && (6.10)\\ &\le \epsilon k \log |\mathcal{X}| + H(\epsilon) + H(f(X^k)). && (6.11) \end{aligned}$$

The proof is completed by (6.7) and (6.11). ■

6.8 Feedback capacity

Feedback message is common in communication systems. We consider a DMC with noise-free and immediate feedback of output symbols. Though intuitively feedback can help encoding, it is surprising to see that feedback cannot increase capacity.

Definition 6.7 An (n, M) code for a DMC $\{W : \mathcal{X} \to \mathcal{Y}\}$ with feedback consists of a sequence of encoding functions

$$\begin{aligned} f_i : \{1, 2, \ldots, M\} \times \mathcal{Y}^{i-1} &\to \mathcal{X}^n\\ (u, y_1, y_2, \cdots, y_{i-1}) &\mapsto x_i \end{aligned}$$

where y_i, $i = 1, \ldots, i-1$ are the first $i-1$ output symbols of the DMC, and a decoding function

$$\begin{aligned} \varphi : \mathcal{Y}^n &\to \{1, 2, \ldots, M\}\\ (y_1, y_2, \ldots, y_n) &\mapsto \hat{u}. \end{aligned}$$

Then we can similarly define the error probabilities and achievable rates for codes with feedback.

Definition 6.8 The feedback capacity of a DMC C_{FB} is the supremum of all the achievable rates.

Theorem 6.5 For a DMC, $C_{\text{FB}} = C$.

Proof of Achievability. $C_{\text{FB}} \ge C$. ■

Proof outline of converse

- Let U be the uniform distributed random variable over the message set $\{1, 2, \ldots, 2^{nR}\}$.
- For $i = 1, \ldots, n$, define $X_i = f_i(U, Y^{i-1})$.
- Let Y_i be the output of the channel with X_i as the input.
- We do not have the markov chain $U \to \mathbf{X} \to \mathbf{Y} \to \hat{U}$.
- Using $U \to \mathbf{Y} \to \hat{U}$, we write

$$\begin{aligned}\log M &= H(U)\\ &= H(U|\hat{U}) + I(U;\hat{U})\\ &\le H(U|\hat{U}) + I(U;\mathbf{Y}).\end{aligned}$$

- By the chain rule for entropy

$$\begin{aligned}H(\mathbf{Y}|U) &= \sum_{i=1}^{n} H(Y_i|U, Y^{i-1})\\ &= \sum_{i=1}^{n} H(Y_i|U, Y^{i-1}, X_i)\\ &= \sum_{i=1}^{n} H(Y_i|X_i).\end{aligned}$$

- Then,

$$\begin{aligned}I(U;\mathbf{Y}) &= H(\mathbf{Y}) - H(\mathbf{Y}|U)\\ &\le \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i)\\ &= \sum_{i=1}^{n} I(X_i;Y_i).\end{aligned}$$

Proof of Theorem 6.5. Since a code without feedback is a special case of a feedback code, $C_{\mathrm{FB}} \ge C$. On the other hand, we show that for any rate R that is achievable by codes with feedback, $R \le C$, which implies $C_{\mathrm{FB}} \le C$.

Let U be the uniform distributed random variable over the message set $\{1, 2, \ldots, M\}$. For $i = 1, \ldots, n$, define $X_i = f_i(U, Y^{i-1})$ and let Y_i be the output of the channel with X_i as the input. These are similar to the converse of the channel coding theorem, except that we cannot show the Markov chain in (138). Particularly, $\mathbf{X}$ and $\hat{U}$ may not be conditionally independent given $\mathbf{Y}$.

We can modify the proof of converse as follows:

$$\begin{aligned}\log M &= H(U) && (6.12)\\ &= H(U|\hat{U}) + I(U;\hat{U}) && (6.13)\\ &\le H(U|\hat{U}) + I(U;\mathbf{Y}) && (6.14)\end{aligned}$$

where the last inequality follows from the data processing inequality by considering Markov chain

$$U \to \mathbf{Y} \to \hat{U}. \qquad (6.15)$$

To bound $I(U;\mathbf{Y})$, we have by the chain rule for entropy

$$\begin{aligned}H(\mathbf{Y}|U) &= \sum_{i=1}^{n} H(Y_i|U, Y^{i-1}) && (6.16)\\ &= \sum_{i=1}^{n} H(Y_i|U, Y^{i-1}, X_i) && (6.17)\\ &= \sum_{i=1}^{n} H(Y_i|X_i) && (6.18)\end{aligned}$$

where (6.17) follows that X_i is a function of U and Y^{i-1}, and (6.18) is obtained by the fact of DMC that Y_i only depends on X_i, but not on the variables generated before X_i. Following the above derivation, we have

$$\begin{aligned} I(U;\mathbf{Y}) &= H(\mathbf{Y}) - H(\mathbf{Y}|U) && (6.19)\\ &\le \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) && (6.20)\\ &= \sum_{i=1}^{n} I(X_i;Y_i). && (6.21) \end{aligned}$$

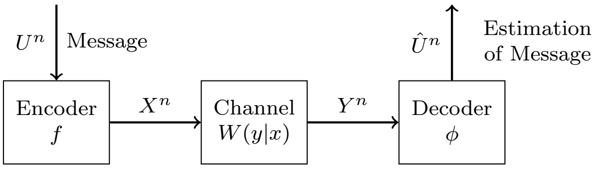
Substituting (6.21) into (6.14), the following of the proof is the same as that of the converse of the channel coding theorem. ■

6.9 Source-Channel Separation Theorem

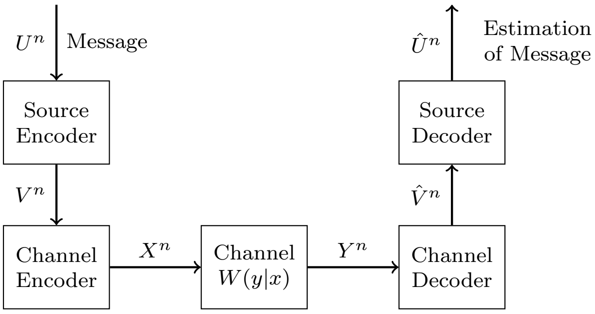
We have studied the compression of a source, where the optimal codes depend on the statistics of the source, and we have studied the communication through a channel, where the optimal codes depend on the channel transition matrix. Now we consider the combination of these two problem, i.e., the communication of a source through a channel. For example, I take a photo and want to transmit the photo to a friend.

Communicate a Source over a Channel

- Joint coding



- Separate coding



In general, the joint coding may depend on both the source and the channel, which may not be a practical solution (we hope for a universal solution). In the separate coding, the source coding depends only on the source and the channel coding depends only on the channel, which is partially universal and hence desired. For example, when compression a file, we do not need to worry what kind of channel will be used to transmit the file, and we may communicate different sources through the same channel without knowing the distribution of the source. Is the separate optimal? What is the achievable rate of the separate coding?

Source-Channel Separation Theorem

- Source: a stochastic process $U_1, U_2, \ldots$ with the entropy rate H and $-\frac{1}{n}\log p(U_1, U_2, \ldots, U_n) \to H$ in probability.
- Channel: a DMC $\{W\}$ with capacity C.

- Error probability: $P_e = P(U^n \neq \hat{U}^n)$.
- If $H < C$, there exists a separate code such that $P_e \to 0$.
- If $H > C$, the error probability is bound away from zero.

Achievability. Note that $-\frac{1}{n}\log p(U_1, U_2, \ldots, U_n) \to H$ in probability implies the similar weak AEP property, i.e., we can define a typical set $A_\epsilon^{(n)}$ such that $|A_\epsilon^{(n)}| \leq 2^{n(H+\epsilon)}$ and $\Pr\{U^n \in A_\epsilon^{(n)}\} > 1-\epsilon$ when n is sufficiently large.

Fix $\epsilon < (C-H)/2$. Use a block source code to encode and decode correctly only the source sequences in $A_\epsilon^{(n)}$, which has rate $\leq H+\epsilon < C$ and error probability $\leq \epsilon$. As $\frac{\log|A_\epsilon^{(n)}|}{n} \leq H+\epsilon < C$, by the achievability of the channel coding theorem, we have channel codes of rate at least $H+\epsilon$ and error probability $\leq \epsilon$. Therefore,

$$P(U^n \neq \hat{U}^n) = P(U^n \notin A_\epsilon^{(n)}) + P(\phi(Y^n) \neq U^n | U^n \in A_\epsilon^{(n)}) \leq 2\epsilon.$$

■

Converse. We show that $P_e \to 0$ implies $H \leq C$. For any code, we can write

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

As there exists codes with $P_e \to 0$, we have $H \leq C$. ■