ECE 563 FA25 HW3 Solutions

Problem 1. Entropy rates of Markov chains.

Solution. *Part A: Entropy Rate Calculation*

Consider the time-invariant Markov chain with state space $\mathcal{X} = \{1, 2, 3\}$ and probability transition matrix:

$$P = \begin{pmatrix} 1-\alpha & \alpha & 0 \\ \beta & 1-\alpha-\beta & \alpha \\ 0 & \beta & 1-\beta \end{pmatrix}$$

From the lecture notes, the stationary distribution is given by:

$$\mu = \left(\frac{1}{1+r+r^2}, \frac{r}{1+r+r^2}, \frac{r^2}{1+r+r^2} \right)$$

where $r = \alpha/\beta$.

For a stationary time-invariant Markov process, the entropy rate is:

$$H(\mathcal{X}) = H(X_2|X_1)$$

Using the stationary distribution, we compute:

$$H(X_2|X_1) = \sum_{i=1}^{3} \mu_i H(X_2|X_1 = i)$$

For each state i, the conditional entropy is:

$$\begin{aligned} H(X_2|X_1 = 1) &= H_b(1-\alpha) = -(1-\alpha)\log(1-\alpha) - \alpha\log\alpha \\ H(X_2|X_1 = 2) &= -\beta\log\beta - (1-\alpha-\beta)\log(1-\alpha-\beta) - \alpha\log\alpha \\ H(X_2|X_1 = 3) &= H_b(\beta) = -\beta\log\beta - (1-\beta)\log(1-\beta) \end{aligned}$$

Therefore, the entropy rate is:

$$\begin{aligned} H(\mathcal{X}) = &\frac{1}{1+r+r^2}\left[-(1-\alpha)\log(1-\alpha) - \alpha\log\alpha\right] \\ &+ \frac{r}{1+r+r^2}\left[-\beta\log\beta - (1-\alpha-\beta)\log(1-\alpha-\beta) - \alpha\log\alpha\right] \\ &+ \frac{r^2}{1+r+r^2}\left[-\beta\log\beta - (1-\beta)\log(1-\beta)\right] \end{aligned}$$

Substituting $r = \alpha/\beta$, this can be simplified to:

$$H(\mathcal{X}) = \frac{1}{1+r+r^2}\left[H_b(1-\alpha) + r \cdot H(\beta, 1-\alpha-\beta, \alpha) + r^2 H_b(\beta)\right]$$

where $H_b(p) = -p\log p - (1-p)\log(1-p)$ is the binary entropy function.

Part B: Card Shuffling and MCMC (Section 4.3)

What the example explains:

Section 4.3 illustrates the application of Markov chains to card shuffling and demonstrates a key concept in Markov Chain Monte Carlo (MCMC): *how long it takes for a Markov chain to reach its stationary distribution.*

The Setup: Consider a deck of 52 cards. The state space has $|\mathcal{X}| = 52!$ possibilities, so a uniform distribution has entropy $H(X) = \log 52! \approx 226$ bits.

One-at-a-time shuffling: In each shuffle, we select a card uniformly at random from one of the 52 positions and place it on top. This creates a Markov chain where:

- Each shuffle adds $H(X_{n+1}|X_n) = \log 52 \approx 5.7$ bits of entropy
- The stationary distribution is uniform by symmetry
- The chain is irreducible and aperiodic, so $H(X_n) \to \log(52!)$ as $n \to \infty$

The key question: How many shuffles are needed before the deck is uniformly distributed and independent of the initial state X_0?

Upper bound derivation: Let $N_1, N_2, \ldots$ denote the locations of cards moved to the top. Then:

$$H(X_0, X_1, \ldots, X_n) = H(X_0) + nH(N_1) = H(X_0) + n \log 52$$

Also:

$$H(X_0, X_1, \ldots, X_n) = H(X_0) + H(X_n|X_0) + H(X_1, \ldots, X_{n-1}|X_n, X_0)$$

This gives:

$$H(X_n|X_0) \leq n \log 52$$

For independence, we need $H(X_n|X_0) = \log 52!$, which requires:

$$n \geq \frac{\log 52!}{\log 52} \approx 39.6$$

Important insight: While this bound is necessary, it is *not sufficient*. The example shows that $H(X_n|X_0) < \log 52!$ for all finite n, meaning there is always some residual memory of the initial state!

Modified shuffle: The lecture then presents a modified scheme where on shuffle n, we draw M uniformly from $\{n, n+1, \ldots, 52\}$ and place the M-th card on top. After exactly 51 shuffles:

$$H(X_{51}|X_0) = \sum_{i=1}^{51} H(M_i) = \sum_{i=1}^{51} \log(52 - i + 1) = \log 52!$$

This achieves perfect randomization in exactly 51 steps, demonstrating that the shuffle design matters significantly for convergence speed.

Part C: Entropy of the English Language (Section 4.4)

What the example explains:

Section 4.4 estimates the entropy rate of English text, treating it as a stochastic process. While English is not truly stationary, we can estimate its effective entropy rate.

Model hierarchy: Using a 27-character alphabet (26 letters + space), different Markov models yield:

- **0th-order (iid):** Each letter is independent, $H(\mathcal{X}) \approx 4.76$ bits/letter
- **1st-order:** Accounts for letter frequency, $H(\mathcal{X}) \approx 4.03$ bits/letter
- **4th-order:** Captures patterns in 4-letter sequences, $H(\mathcal{X}) \approx 2.8$ bits/letter

Shannon's experiment (1952): Shannon asked people to guess the next letter in text until correct. The order of guesses reflects their internal probability model for the conditional distribution. This human experiment yielded:

$$H(\mathcal{X}) \approx 1.3 \text{ bits/letter}$$

Key insights:

1) As model order increases (capturing more context), the entropy rate decreases, reflecting the predictability structure in language.
2) Human knowledge incorporates semantic understanding beyond pure statistical patterns, leading to even lower entropy estimates.
3) This demonstrates that the entropy rate quantifies the *intrinsic randomness* or *information content* after accounting for all predictable structure.
4) The difference between theoretical models and human performance reveals the gap between statistical and semantic understanding of language.

Problem 2. More on Kraft's inequality.

Solution. The Kraft–McMillan inequality goes as follows:

Theorem 2.1. *[1, Theorem 5.5.1] Let $\mathcal{C} = \{s_1, \ldots, s_M\}$ be a uniquely decodable code over an alphabet of size D. For each $i \in [1, M]$ write ℓ_i for the length of the codeword s_i. Then we have*

$$\sum_{i=1}^{M} D^{-\ell_i} \leq 1. \tag{2.1}$$

Conversely, for any positive integers $\ell_1, \ldots, \ell_M$ that satisfy (2.1)*, there exists a uniquely decodable code over an alphabet of size D with those word lengths.*

Furthermore, the statement still holds when the source alphabet is countably infinite (i.e. the case $M = \infty$).

Proof of Theorem 2.1. This proof is based on the proof in the textbook.

Let k be an arbitray positive integer. Note that if $\mathcal{C}$ is uniquely decodable, then so is the kth extension of $\mathcal{C}$, defined as

$$\mathcal{C}^k \coloneqq \{s_{i_1}s_{i_2}\cdots s_{i_k} \ : \ i_1,\ldots,i_k \in [1,M]\}. \tag{2.2}$$

For example, the second extension of the code $\{0, 01\}$ is

$$\{00, 001, 010, 0101\}.$$

Then, note that we can write

$$\begin{aligned}(\sum_{i=1}^{M} D^{-\ell_i})^k &= \sum_{i_1=1}^{M}\cdots\sum_{i_k=1}^{M} D^{-\ell_{i_k}}\cdots D^{-\ell_{i_k}} \\ &= \sum_{s\in\mathcal{C}^k} D^{-\text{lth}(s)},\end{aligned} \tag{2.3}$$

where $\text{lth}(s)$ denotes the length of the word s. Let ℓ_{max} denote the maximum of $\ell_1,\ldots,\ell_M$. It follows that the longest codelength of $\mathcal{C}^k$ is $k\ell_{\text{max}}$. Then, we can arrange the sum in (2.3) into

$$\sum_{s\in\mathcal{C}^k} D^{-\text{lth}(s)} = \sum_{\ell=1}^{k\ell_{\text{max}}} N_k(\ell)D^{-\ell}, \tag{2.4}$$

where $N_k(\ell)$ denotes the number of codewords in $\mathcal{C}^k$ that has length ℓ. Since $\mathcal{C}^k$ is uniquely decodable, we must have $N_k(\ell) \le D^\ell$ for each ℓ. Therefore, from (2.3) and (2.4) we have

$$\begin{aligned}(\sum_{i=1}^{M} D^{-\ell_i})^k &= \sum_{\ell=1}^{k\ell_{\text{max}}} N_k(\ell)D^{-\ell} \\ &\le \sum_{\ell=1}^{k\ell_{\text{max}}} D^\ell D^{-\ell} \\ &= k\ell_{\text{max}}.\end{aligned}$$

Or equivalently,

$$\sum_{i=1}^{M} D^{-\ell_i} \le (k\ell_{\text{max}})^{\frac{1}{k}} \tag{2.5}$$

Since (2.5) holds for any positive integer k and $\lim_{k\to\infty}(k\ell_{\text{max}})^{\frac{1}{k}} = 1$, we must have

$$\sum_{i=1}^{M} D^{-\ell_i} \le 1, \tag{2.6}$$

or otherwise there will be some k large enough such that (2.5) fails. Note that (2.6) is exactly what we want to prove.

As for the case of infinitely countable source alphabet, note that for any uniquely decodable code $\mathcal{C} = \{s_1, s_2, \cdots\}$, any subset of it is still uniquely decodable. Therefore, we have

$$\begin{aligned}\sum_{i=1}^{\infty} D^{-\ell_i} &= \lim_{M\to\infty}\sum_{i=1}^{M} D^{-\ell_i} \\ &\le \lim_{M\to\infty} 1 \\ &= 1,\end{aligned} \tag{2.7}$$

where in (2.7) we used the fact that $\{s_1, s_2, \cdots, s_M\}$ is uniquely decodable.

For the converse, since prefix codes are uniquely decodable, for any $\ell_1, \ell_2, \ldots$ that satisfy (2.1), we can simply construct the prefix code with those word lengths based on the proof of the Kraft inequality [1, Theorem 5.2.1]. □

Problem 3. Adaptive Huffman codes.

Solution. Problem:

Adaptive Huffman Coding Overview:

In adaptive Huffman coding, the tree is dynamically updated as each symbol is processed. We start with an initial tree (typically with all symbols having equal or minimal weights) and update the tree after encoding each symbol based on the observed frequencies so far.

Initial Setup

Initially, before seeing any symbols, we assume all symbols have equal (or zero) frequency. We'll use the convention of starting with a simple tree or updating as we go.

Step-by-Step Construction

Symbol 1: 'a'

- Frequencies: $a = 1, b = 0, c = 0$
- Initial encoding: We can use a simple assignment or single-bit encoding
- Tree: 'a' is the only symbol seen, encoded as (e.g., root node)

Symbol 2: 'a'

- Frequencies: $a = 2, b = 0, c = 0$
- Still only 'a' has been seen
- Update weight of 'a' to 2

Symbol 3: 'b'

- Frequencies: $a = 2, b = 1, c = 0$
- First appearance of 'b'
- Huffman tree:

```
    (3)
   /   \
a(2)   b(1)
```

- Codes: $a \to 0$, $b \to 1$

Symbol 4: 'c'

- Frequencies: $a = 2, b = 1, c = 1$
- First appearance of 'c'
- Huffman tree construction:

```
      (4)
     /   \
  a(2)   (2)
         / \
      b(1) c(1)
```

- Codes: $a \to 0$, $b \to 10$, $c \to 11$

Symbol 5: 'c'

- Frequencies: $a = 2, b = 1, c = 2$
- Huffman tree construction (combining lowest weights):

```
      (5)
     /   \
   (3)    a(2)
   / \
c(2) b(1)
```

- Codes: $a \to 1$, $b \to 01$, $c \to 00$

Symbol 6: 'b'

- Frequencies: $a = 2, b = 2, c = 2$
- All symbols have equal frequency
- Huffman tree (multiple valid constructions):

```
      (6)
     /   \
   (4)    c(2)
   / \
a(2) b(2)
```

- Codes: $a \to 00$, $b \to 01$, $c \to 1$

Symbol 7: 'b'

- Frequencies: $a = 2, b = 3, c = 2$
- Huffman tree:

```
      (7)
```

```
    /   \
b(3)   (4)
      /  \
   a(2)  c(2)
```

- Codes: $a \to 10$, $b \to 0$, $c \to 11$

Symbol 8: 'c'

- Final frequencies: $a = 2, b = 3, c = 3$
- Final Huffman tree:

```
      (8)
     /   \
   (5)    c(3)
   / \
b(3)  a(2)
```

- Final codes: $a \to 01$, $b \to 00$, $c \to 1$

Summary

The adaptive Huffman coding process for `aabccbbc` involves:

1) Starting with no prior knowledge
2) After each symbol, updating the frequency counts
3) Reconstructing the Huffman tree based on current frequencies
4) Using the new tree to encode subsequent symbols

The key advantage is that the encoder and decoder can synchronize without transmitting the tree explicitly, as both update their trees identically after each symbol. The disadvantage is computational overhead from frequent tree updates.

Total bits encoded: The exact bit count depends on implementation details (how to signal new symbols, etc.), but the tree adapts to the observed distribution, ultimately converging to the optimal static Huffman code for the given frequency distribution $(a : 2, b : 3, c : 3)$.

Problem 4. Abel-Euler smoothing. Prove that if $x_n \to L$, then

$$\lim_{r \to 1^-} (1 - r) \sum_{n=0}^{\infty} r^n x_n = L. \tag{4.1}$$

Solution.

Let $\varepsilon > 0$ be given. Our goal is to show that there exists $\delta > 0$ such that for all $r \in (1 - \delta, 1)$ we have

$$|(1 - r) \sum_{n=0}^{\infty} r^n x_n - L| < \varepsilon. \tag{4.2}$$

First, since $x_n \to L$, there exists $N \geq 0$ such that for all $n \geq N$ we have $|x_n - L| < \frac{\varepsilon}{2}$.

Then, note that for $r \in (0, 1)$ we can write

$$L = (1 - r) \sum_{n=0}^{\infty} r^n L. \tag{4.3}$$

It follows that for $r \in (0, 1)$ we have

$$\begin{aligned} |(1 - r) \sum_{n=0}^{\infty} r^n x_n - L| &= |(1 - r) \sum_{n=0}^{\infty} r^n x_n - (1 - r) \sum_{n=0}^{\infty} r^n L| \\ &= |(1 - r) \sum_{n=0}^{\infty} r^n (x_n - L)| \\ &\leq (1 - r) \sum_{n=0}^{\infty} r^n |x_n - L| \\ &= (1 - r) \sum_{n=0}^{N-1} r^n |x_n - L| + (1 - r) \sum_{n=N}^{\infty} r^n |x_n - L|. \end{aligned} \tag{4.4}$$

We bound the two terms in (4.4) separately. First, we have

$$\sum_{n=0}^{N-1} r^n |x_n - L| \le \sum_{n=0}^{N-1} |x_n - L|.$$

Define $M \coloneqq \sum_{n=0}^{N-1} |x_n - L|$, which does not depend on r. Then, if we further have $1 - r < \frac{\varepsilon}{2M}$, we can get

$$\begin{aligned}(1-r)\sum_{n=0}^{N-1} r^n |x_n - L| &< \frac{\varepsilon}{2M} M \\ &= \frac{\varepsilon}{2}. \end{aligned} \tag{4.5}$$

Second, by the definition of N, we have

$$\begin{aligned}(1-r)\sum_{n=N}^{\infty} r^n |x_n - L| &< (1-r)\sum_{n=N}^{\infty} r^n \frac{\varepsilon}{2} \\ &= (1-r)\frac{\varepsilon}{2}\sum_{n=N}^{\infty} r^n \\ &\le (1-r)\frac{\varepsilon}{2}\sum_{n=0}^{\infty} r^n \\ &= \frac{\varepsilon}{2}. \end{aligned} \tag{4.6}$$

Putting (4.5) and (4.6) into (4.4), we see that if $r \in (0,1)$ and $1 - r < \frac{\varepsilon}{2M}$, then we have

$$\begin{aligned}|(1-r)\sum_{n=0}^{\infty} r^n x_n - L| &< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \\ &= \varepsilon. \end{aligned}$$

In other words, we can select $\delta = \min(\frac{\varepsilon}{2M}, 1)$, and then for all $r \in (1-\delta, 1)$, we have (4.2).

These arguments prove (4.1).

Problem 5. The Z channel.

Solution.

Let p denote $p(0|1)$, which we will replace with $\frac{1}{4}$ at the end.

Since X is a binary variable, its distribution can be characterized by a single parameter $\alpha \in [0,1]$. To be more precise, let $p_X(0) = \alpha$. Then, our goal is to calculate

$$\begin{aligned} C &= \max_{\alpha \in [0,1]} I(X;Y) \\ &= \max_{\alpha \in [0,1]} H(Y) - H(Y|X). \end{aligned}$$

We calculate the two terms separately. First, note that

$$\begin{aligned} H(Y|X) &= p_X(0)H(Y|X=0) + p_X(1)H(Y|X=1) \\ &= (1-\alpha)0 + \alpha H(p), \end{aligned} \tag{5.1}$$

where we used the fact that Y is deterministic given $X = 0$ and $Y \sim \text{Ber}(p)$ given $X = 1$. Here $H(\cdot)$ denotes the binary entropy. Second, note that the marginal distribution of Y is

$$\begin{aligned} p_Y(0) &= p_X(0)p_{Y|X}(0|0) + p_X(1)p_{Y|X}(0|1) \\ &= 1 - \alpha + p\alpha, \\ p_Y(1) &= p_X(0)p_{Y|X}(1|0) + p_X(1)p_{Y|X}(1|1) \\ &= (1-p)\alpha. \end{aligned}$$

In other words, Y follows a Bernoulli-$(1-\alpha+p\alpha)$ distribution. It follows that

$$\begin{aligned} H(Y) &= H(1-\alpha+p\alpha) \\ &= H((1-p)\alpha). \end{aligned} \tag{5.2}$$

Combining (5.1) and (5.2) gives

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H((1-p)\alpha) - \alpha H(p). \end{aligned} \tag{5.3}$$

To maximize the capacity in (5.3), it suffices to maximize $f(\alpha) \coloneqq H((1-p)\alpha) - \alpha H(p)$ for $\alpha \in [0,1]$. A simple calculation gives (recall that p is a constant here)

$$\begin{aligned} f'(\alpha) &= (1-p)\log(\frac{1}{(1-p)\alpha} - 1) - H(p), \\ f''(\alpha) &= -\log e \frac{1-p}{p}\frac{1}{\alpha^2} < 0. \end{aligned}$$

Solving $f'(\alpha) = 0$ gives

$$\alpha = \frac{1}{(1-p)(1+2^{\frac{H(p)}{1-p}})}. \tag{5.4}$$

It follows that the maximal value of f is

$$\begin{aligned} f(\frac{1}{(1-p)(1+2^{\frac{H(p)}{1-p}})}) &= H(\frac{1}{1+2^{\frac{H(p)}{1-p}}}) - \frac{H(p)}{(1-p)(1+2^{\frac{H(p)}{1-p}})} \\ &= \frac{1}{1+2^{\frac{H(p)}{1-p}}}\log(1+2^{\frac{H(p)}{1-p}}) + \frac{2^{\frac{H(p)}{1-p}}}{1+2^{\frac{H(p)}{1-p}}}\log\frac{1+2^{\frac{H(p)}{1-p}}}{2^{\frac{H(p)}{1-p}}} - \frac{H(p)}{(1-p)(1+2^{\frac{H(p)}{1-p}})} \\ &= \log(1+2^{\frac{H(p)}{1-p}}) - \frac{2^{\frac{H(p)}{1-p}}}{1+2^{\frac{H(p)}{1-p}}}\frac{H(p)}{1-p} - \frac{H(p)}{(1-p)(1+2^{\frac{H(p)}{1-p}})} \\ &= \log(1+2^{\frac{H(p)}{1-p}}) - \frac{H(p)}{1-p} \\ &= \log(1+2^{-\frac{H(p)}{1-p}}). \end{aligned}$$

In other words, the capacity of the Z-channel with $p(0|1) = p$ is

$$\begin{aligned} C &= \max_{\alpha\in[0,1]} I(X;Y) \\ &= \log(1+2^{-\frac{H(p)}{1-p}}). \end{aligned} \tag{5.5}$$

Now we put $p = \frac{1}{4}$ in (5.5), and then the capacity of the Z-channel with $p(0|1) = \frac{1}{4}$ is

$$\begin{aligned} \log(1+2^{-\frac{H(\frac{1}{4})}{1-\frac{1}{4}}}) &= \log(1+2^{-\frac{4}{3}(\frac{1}{4}\log 4 + \frac{3}{4}\log\frac{4}{3})}) \\ &= \log(1+2^{-\frac{8}{3}+\log 3}) \\ &= \log(1+\frac{3}{4\sqrt[3]{4}}) \\ &\approx 0.5582. \end{aligned}$$

Problem 6. The probabilistic method. Show that there exists a subset $S \subseteq [1,n]$ of size $\Theta(n^{\frac{1}{3}})$ that contains no 3-term arithmetic progressions (3-APs). In other words, the size of such S scales as $\Theta(n^{\frac{1}{3}})/n = \Theta(n^{-\frac{2}{3}})$.

Solution.

Let $p \in [0,1]$ to be specified later (possibly depending on n). Consider a random subset $S \subseteq [1,n]$ constructed as follows: For each $i \in [1,n]$, we put i into S independently with probability p. It follows that the expected size of S is

$$\begin{aligned} \mathbb{E}[|S|] &= \sum_{i=1}^{n} \mathbb{P}(i \in S) \\ &= np. \end{aligned} \tag{6.1}$$

Then, let N_S be the number of 3-APs in S. A similar argument shows that

$$\begin{aligned} \mathbb{E}[N_S] &= \sum_{T \text{ is a 3-AP in } [1,n]} \mathbb{P}(T \subseteq S) \\ &= N_{\text{3AP},n} p^3, \end{aligned} \tag{6.2}$$

where $N_{3\text{AP},n}$ denotes the number of 3-APs in $[1,n]$ (which is a deterministic quantity). A simple upper bound of $N_{3\text{AP},n}$ is

$$\begin{aligned} N_{3\text{AP},n} &\leq \binom{n}{2} \\ &\leq \frac{n^2}{2}, \end{aligned} \tag{6.3}$$

since the two endpoints of any 3-AP in $[1,n]$ uniquely determine a size-2 subset of $[1,n]$. Then, putting (6.3) into (6.2) gives

$$\mathbb{E}[N_S] \leq \frac{n^2}{2}p^3. \tag{6.4}$$

Combining (6.1) and (6.4) yields

$$\mathbb{E}[|S| - N_S] \geq np - \frac{n^2}{2}p^3. \tag{6.5}$$

From (6.5) we can deduce that

$$\mathbb{P}\left(|S| - N_S \geq np - \frac{n^2}{2}p^3\right) > 0. \tag{6.6}$$

In other words, there exists a subset $S \subseteq [1,n]$ such that its size subtracted by the number of 3-APs in it is at least $np - \frac{n^2}{2}p^3$.

Now we can assign the value of p. If we choose $p = n^{-\frac{2}{3}}$, then (6.6) implies that there exists $S \subseteq [1,n]$ such that $|S| - N_S \geq n^{\frac{1}{3}} - \frac{1}{2} = \Theta(n^{\frac{1}{3}})$. Then, we can remove one element from each 3-AP in S, and the resulting set S' will be 3-AP-free and has size at least $n^{\frac{1}{3}} - \frac{1}{2} = \Theta(n^{\frac{1}{3}})$. Q.E.D.

Remark 6.1. In fact, by maximizing $np - \frac{n^2}{2}p^3$ over p, we can see that if we set $p = \frac{1}{\sqrt{n}}$, then the quantity $np - \frac{n^2}{2}p^3$ can be as large as $\Theta(\sqrt{n})$. This leads to a 3-AP-free subset of $[1,n]$ of size $\Theta(n^{\frac{1}{2}})$, which is larger than the requirement $\Theta(n^{\frac{1}{3}})$ of this problem.

Remark 6.2. A more careful discussion can show that the quantity $N_{3AP,n}$ in (6.3) is actually $N_{3AP,n} = \frac{n^2}{4} + O(n)$. However, since we only care about the asymptotic behavior of these quantities, different constant factors ($\frac{1}{2}$ v.s. $\frac{1}{4}$) lead to the same result.

Problem 7. Feedback channels and joint source-channel coding theorem.

Solution.

Part A: Feedback Does Not Increase DMC Capacity

Theorem 7.1. *For a discrete memoryless channel (DMC), the capacity with feedback equals the capacity without feedback.*

Proof. Let C denote the capacity without feedback and C_{FB} the capacity with feedback.

Step 1: $C_{FB} \leq C$ is trivial.

Any code that doesn't use feedback is also valid when feedback is available, so $C \leq C_{FB}$.

Step 2: Prove $C_{FB} \leq C$.

Consider a $(2^{nR}, n)$ code with feedback. The encoder at time i can depend on the entire feedback $Y^{i-1} = (Y_1, \ldots, Y_{i-1})$ and the message W. Thus:

$$X_i = f_i(W, Y^{i-1})$$

The message W is uniformly distributed over $\{1, 2, \ldots, 2^{nR}\}$.

By Fano's inequality, if the probability of error $P_e^{(n)} \to 0$, then:

$$H(W|Y^n) \leq n\epsilon_n$$

where $\epsilon_n \to 0$ as $n \to \infty$.

Now, we compute the mutual information:

$$\begin{aligned} nR &= H(W) \\ &= I(W; Y^n) + H(W|Y^n) \\ &\leq I(W; Y^n) + n\epsilon_n \\ &= H(Y^n) - H(Y^n|W) + n\epsilon_n \end{aligned}$$

Using the chain rule:

$$H(Y^n|W) = \sum_{i=1}^{n} H(Y_i|Y^{i-1}, W)$$

For a DMC with transition probabilities $p(y|x)$:

$$\begin{aligned} H(Y_i|Y^{i-1}, W) &= H(Y_i|Y^{i-1}, W, X_i) \quad (\text{since } X_i = f_i(W, Y^{i-1})) \\ &= H(Y_i|X_i) \quad (\text{DMC property: } Y_i \perp Y^{i-1}, W|X_i) \end{aligned}$$

Therefore:

$$H(Y^n|W) = \sum_{i=1}^{n} H(Y_i|X_i)$$

Continuing:

$$\begin{aligned} nR &\leq H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i) + n\epsilon_n \\ &= \sum_{i=1}^{n} [H(Y_i) - H(Y_i|X_i)] + n\epsilon_n \\ &= \sum_{i=1}^{n} I(X_i; Y_i) + n\epsilon_n \\ &\leq \sum_{i=1}^{n} C + n\epsilon_n \\ &= nC + n\epsilon_n \end{aligned}$$

where we used the fact that $I(X_i; Y_i) \leq \max_{p(x)} I(X; Y) = C$ for each i.
Dividing by n and taking $n \to \infty$:

$$R \leq C + \epsilon_n \to C$$

This shows $C_{FB} \leq C$.
Conclusion: Combining both inequalities, $C_{FB} = C$. □

Intuition: The key insight is that for a memoryless channel, the output Y_i depends only on the current input X_i, not on past inputs or outputs. While feedback allows the encoder to adapt future inputs based on past outputs, this doesn't increase the instantaneous information transmission capacity. The memoryless property ensures that each use of the channel is independent, and the constraint $I(X_i; Y_i) \leq C$ holds regardless of how X_i is chosen.

Part B: Joint Source-Channel Coding Theorem

Theorem 7.2 (Joint Source-Channel Coding)**.** *Consider a discrete memoryless source (DMS) with entropy rate $H(\mathcal{S})$ and a DMC with capacity C.*

If $H(\mathcal{S}) < C$, then there exists a sequence of source-channel codes such that the source can be transmitted over the channel with arbitrarily small probability of error.

Conversely, if $H(\mathcal{S}) > C$, no sequence of codes can achieve reliable communication.

Proof. **Achievability ($H(\mathcal{S}) < C$):**
Let $H(\mathcal{S}) < C$. Choose R such that $H(\mathcal{S}) < R < C$.

Step 1 (Source Coding): By the source coding theorem, for any $\epsilon > 0$ and sufficiently large k, there exists a source code that maps k source symbols into ℓ bits where:

$$\frac{\ell}{k} < H(\mathcal{S}) + \epsilon < R$$

with arbitrarily small decoding error probability.

Step 2 (Channel Coding): By the channel coding theorem, for any $\delta > 0$ and sufficiently large n, there exists a channel code that can reliably transmit m bits over n channel uses where:

$$\frac{m}{n} > C - \delta > R$$

with arbitrarily small error probability.

Step 3 (Concatenation): Choose ϵ and δ small enough so that:

$$\frac{\ell}{k} < R < \frac{m}{n}$$

Then $\ell < m$, so we can transmit the ℓ-bit compressed source sequence through the channel code (padding with zeros if necessary).

The overall probability of error is bounded by:

$$P_e^{(total)} \leq P_e^{(source)} + P_e^{(channel)} \to 0$$

as $k, n \to \infty$.

The rate of source transmission is k/n source symbols per channel use, which can be made arbitrarily close to $C/H(\mathcal{S})$.

Converse ($H(\mathcal{S}) > C$):

Suppose $H(\mathcal{S}) > C$ and assume there exists a reliable communication system that maps k source symbols to n channel uses with $P_e^{(n)} \to 0$.

Let $S^k = (S_1, \ldots, S_k)$ be the source sequence and $\hat{S}^k$ be the estimate at the receiver.

By Fano's inequality:

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

Then:

$$\begin{aligned} kH(\mathcal{S}) &= H(S^k) \\ &= I(S^k; \hat{S}^k) + H(S^k|\hat{S}^k) \\ &\leq I(S^k; Y^n) + 1 + P_e^{(n)} k \log |\mathcal{S}| \\ &\leq H(Y^n) + 1 + P_e^{(n)} k \log |\mathcal{S}| \\ &\leq nC + 1 + P_e^{(n)} k \log |\mathcal{S}| \end{aligned}$$

Dividing by k and letting $n, k \to \infty$ with n/k fixed:

$$H(\mathcal{S}) \leq \frac{n}{k} C$$

Since this must hold for any code, we need $H(\mathcal{S}) \leq \frac{n}{k} C$, or equivalently $\frac{k}{n} \leq \frac{C}{H(\mathcal{S})}$.

But if $H(\mathcal{S}) > C$, then $\frac{C}{H(\mathcal{S})} < 1$, meaning we can only transmit less than one source symbol per channel use in the limit. This implies the source cannot be reliably transmitted at its natural rate. □

Key Insights:

- The theorem shows that *separation* is optimal: we can separately design source and channel codes without loss of optimality.
- The critical condition $H(\mathcal{S}) < C$ means the source entropy rate must be less than channel capacity.
- Unlike feedback, proper source-channel code design can exploit the interplay between source and channel, but Shannon showed that simple concatenation (source code + channel code) achieves the optimal performance.

REFERENCES

[1] T. M. Cover, *Elements of information theory*. John Wiley & Sons, 1999.