Solutions: Homework Set # 8

Problem 1 (FEEDBACK CAPACITY OF ERASURE CHANNEL WITH MEMORY)

(a) Since the bit gets through with probability $1-\alpha$, the average (effective) rate of transmission is $1-\alpha$.

Here is a simple transmission scheme: the transmitter keeps sending a bit until receives a Y from the receiver over the feedback link. The receiver sends Y if it receives the bit, otherwise it sends back an N. The probability that the channel is used k times for transmitting a bit is the probability that the channel is in the erasure state for $k-1$ time slots, and in the correct state for the k-th time slot, which is $\Pr[T=k] = \alpha^{k-1}(1-\alpha)$. Therefore the average number of times one has to use the channel to transmit one bit is

$$\begin{aligned}\mathbb{E}[T] &= \sum_{k=1}^{\infty} k\alpha^{k-1}(1-\alpha) \\ &= \frac{1}{1-\alpha}.\end{aligned}$$

Therefore the transmission rate, the average number bits transmitted per channel use, is $R = 1-\alpha$.

(b) Note that you have already seen that for discrete memoryless channels, feedback does not increase capacity. For discrete memoryless channels, we have already computed the capacity of erasure channel. Thus

$$C_{FB} = C = 1-\alpha.$$

Thus the above trivial transmission scheme achieves feedback capacity of this memoryless erasure channel. Note that although the feedback does not change the channel capacity, it can simplify the capacity achieving transmission scheme. In words, if the feedback was not present, one should use a capacity achieving code with large enough codeword length, and by encoding ad decoding the message can transmit at capacity rate, while a very simple scheme can achieves the same performance when feedback is present.

(c) If the feedback is not present, the equivalent channel is an erasure channel with probability of erasure being equal to the probability of being in the state E (π_E). This can be found by calculating the stationary distribution of the given Markov chain:

$$[\pi_E, \pi_C] = [\pi_E, \pi_C]P.$$

Thus $\pi_E = \frac{5}{7}$ and thus $\pi_C = \frac{2}{7}$.
Remark: Remember that we calculated the capacity of a discrete memoryless erasure channel (no feedback) in class. If the probability of erasure was α, the capacity of the channel was found to be $1-\alpha$. Refer to the lecture notes for the details!

(d) Here, W is the message that should be sent. To send W, we consider n times using the channel. At time i, the input is X_i and the output is Y_i based on the input X_i and the channel probabilities that in fact depends on the state of the channel (Q_i). X_i is a function of W and $Y^{i-1} = (Y_1, Y_2, \ldots, Y_{i-1})$ because we have access to a feedback. So,

I. Q^n is a function of Y^n; the state of the channel only depends on its previous state which is captured by the sequence of the channel outputs Y^n. (If Y_i is erased, Q_i has been in state E.)

II. W is independent of Q^n.

III. X_i is a function of Y^{i-1} and W. Note that Y^{i-1} is the sequence of $Y_1, Y_2, \cdots, Y_{i-1}$.

IV. given X_i and Q_i, Y_i is independent of W.

V. conditioning decreases entropy.

VI. if X_i and Q_i are given, then the i^{th} output of the channel (Y_i) is independent of Y^{i-1} and Q_1^{i-1}, Q_{i+1}^n.

Remark: The crucial thing to note is that $I(X_i; Y_i|Q_i)$ depends on i because X_i is a function of Y^{i-1}, and hence this could be a non-stationary process. However, the mutual information is a concave function in its input distribution for a fixed channel. Hence, by choosing an average distribution averaged over the Y^{i-1}, i.e.,

$$\bar{p}(X_i|W) = \sum_{y^{i-1}} p(X_i|Y^{i-1} = y^{i-1}, W) p(y^{i-1}|W),$$

the mutual information would increase. Hence, if we define the stationary process $\{\bar{X}_i\}$, where the X_i are i.i.d. with the above marginal distribution, ad using the concavity of mutual information, we obtain

$$I(X_i(y^{i-1}, W); Y_i|Q_i) \leq I(\bar{X}_i(W); Y_i|Q_i).$$

Therefore,

$$\frac{1}{n} I(W; Y^n) \leq \frac{1}{n} \sum_{i=1}^{n} I(\bar{X}_i(W); Y_i|Q_i).$$

Note that by maximizing the input distribution and taking n large enough, the last expression above would be the capacity with no feedback. On the other hand the capacity of a channel without feedback does not exceed the capacity without feedback. Thus

$$C_{NFB} \leq C_{FB} \leq \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \max_{p(\bar{X}_i)} I(\bar{X}_i; Y_i|Q_i) = C_{NFB}.$$

Therefore, $C_{FB} = C_{NFB}$.

Problem 2

(a)

$$\begin{aligned} f(p,w^*) - f(p,w) &= \sum_{x,y} p(x)q(y|x) \log \frac{w^*(x|y)}{p(x)} - \sum_{x,y} p(x)q(y|x) \log \frac{w(x|y)}{p(x)} \\ &= \sum_{x,y} p(x)q(y|x) \log \frac{w^*(x|y)}{w(x|y)} \\ &\overset{(I)}{=} \sum_{x,y} r(y)w^*(x|y) \log \frac{w^*(x|y)}{w(x|y)} \\ &= \sum_{y} r(y) \sum_{x} w^*(x|y) \log \frac{w^*(x|y)}{w(x|y)} \\ &= \sum_{y} D\left(w^*(x|y) \parallel w(x|y)\right) \\ &\overset{(II)}{\geq} 0. \end{aligned}$$

where (I) is due to the definition of $w^*(x|y)$, and (II) holds since $D(\cdot \parallel \cdot)$ is always non-negative, and $r(y) \geq 0$.

(b) We have to maximize $f(p,w)$ with respect to p subject to the constraint $\sum_x p(x) = 1$. By taking the derivative of f with respect to $p(x)$, we have

$$\begin{aligned} \frac{\partial f(p,w)}{\partial p(x)} &= \sum_{y} q(y|x) \log \frac{w(x|y)}{p(x)} - \sum_{y} p(x)q(y|x) \frac{-1}{p(x)} \\ &\overset{(III)}{=} \sum_{y} q(y|x) \log \frac{w(x|y)}{p(x)} - 1. \end{aligned}$$

where (III) holds since $\sum_y q(y|x) = 1$ for all $x \in \mathcal{X}$. From the Kuhn-Tucker conditions, assuming that the maximizing p will have positive components, these derivatives must all equal to a constant λ. Therefore,

$$\sum_{y} q(y|x) \log w(x|y) - \log p(x) \underbrace{\sum_{y} q(y|x)}_{1} - 1 = \lambda,$$

or

$$p(x) = \exp(-1-\lambda) \exp\left(\sum_{y} q(y|x) \log w(x|y) \right).$$

where the constant λ should be chosen such that $\sum_x p(x) = 1$. Hence,

$$p(x) = \frac{\exp\left(\sum_y q(y|x) \log w(x|y)\right)}{\sum_{x'} \exp\left(\sum_y q(y|x') \log w(x'|y)\right)}. \tag{1}$$

(c) Note that $C \geq f(p^{(n+1)}, w^{(n)})$, by the definition of C. Therefore,

$$\begin{aligned}
\sum_{n=0}^{N} \left| C - f(p^{(n+1)}, w^{(n)}) \right| &= \sum_{n=0}^{N} \left[C - f(p^{(n+1)}, w^{(n)}) \right] \\
&\leq \sum_{n=0}^{N} \sum_{x \in \mathcal{X}} p^*(x) \log \frac{p^{(n+1)}(x)}{p^{(n)}(x)} \\
&= \sum_{x \in \mathcal{X}} p^*(x) \log \prod_{n=0}^{N} \frac{p^{(n+1)}(x)}{p^{(n)}(x)} \\
&= \sum_{x \in \mathcal{X}} p^*(x) \log \frac{p^{(N+1)}(x)}{p^{(0)}(x)} \\
&\leq \max_x \log \frac{p^{(N+1)}(x)}{p^{(0)}(x)} \\
&\leq \max_x \log \frac{1}{p^{(0)}(x)}.
\end{aligned}$$

Note that the LHS grow with N, while the RHS does not depend on N. Therefore, we can conclude that the sequence $Cf(p^{(n+1)}, w^{(n)})$ is summable, and thus $Cf(p^{(n+1)}, w^{(n)})$ must converge to zero at least as fast as $1/n$.

Problem 3

Let $J = \{1, 2, \ldots, |\mathcal{Y}|\}$ be the set of indices for the columns of matrix W. Hence, for any $j \in J$:

$$W_j = \begin{bmatrix} w(y_j|x_1) \\ w(y_j|x_2) \\ \cdots \\ w(y_j|x_{|\mathcal{X}|}) \end{bmatrix}.$$

Let $J_1, \ldots, J_l$ be a partition of J (in other words $\cup_{k=1}^{l} J_k = J,\ J_k \cap J_{k'} = \emptyset$) with the following properties:

- For any two $j, j' \in J_k$ the column W_j is a permutation of elements of the column $W_{j'}$. In particular, notice that this implies

$$\sum_{i=1}^{|\mathcal{X}|} w(y_j|x_i) = \sum_{i=1}^{|\mathcal{X}|} w(y_{j'}|x_i) = \lambda_k \ \text{ for all } j, j' \in J_k. \tag{2}$$

- Suppose $J_k = \{j_{k_1}, j_{k_2}, \ldots, j_{k_{|J_k|}}\}$, then the row i of a partition J_k is

$$R_i = [w(y_{k_1}|x_i)\ w(y_{k_2}|x_i)\ \cdots\ w(y_{k_{|J_k|}}|x_i)].$$

For any two $i, i' \in \{1, \ldots, |\mathcal{X}|\}$, R_i and $R_{i'}$ are permutations of each other. In particular, this means that

$$\sum_{j=1}^{|J_k|} w(y_{k_j}|x_i) \log w(y_{k_j}|x_i) = \mu_k \ \text{ for all } i \in \{1, \ldots, |\mathcal{X}|\}. \tag{3}$$

We know that a partition with these properties exists since our channel is symmetric (according to the definition given in the problem).

(a) We know that KKT conditions are necessary and sufficient conditions for an input distribution to be capacity achieving. Hence, if we can prove that the uniform input distribution satisfies the KKT conditions, then we know that this distribution is capacity achieving. What we need to prove is $\frac{\partial I(X;Y)}{\partial p(x_i)} = \lambda$ for any $i \in \{1, \ldots, |\mathcal{X}|\}$, assuming $p(x_i) = \frac{1}{|\mathcal{X}|}$. From class we know that

$$\frac{\partial I(X;Y)}{\partial p(x_i)} = \sum_{j=1}^{|\mathcal{Y}|} w(y_j|x_i) \log \frac{w(y_j|x_i)}{\sum_{i=1}^{|\mathcal{X}|} p(x_i) w(y_j|x_i)} - 1,$$

hence we need to prove that the first part of this expression does not depend on i for a uniform input distribution.

$$\begin{aligned}
\frac{\partial I(X;Y)}{\partial p(x_i)} - 1 &= \sum_{k=1}^{l} \sum_{j=1}^{|J_k|} w(y_{k_j}|x_i) \log \frac{w(y_{k_j}|x_i)}{\sum_{i=1}^{|\mathcal{X}|} p(x_i) w(y_{k_j}|x_i)} \\
&= \sum_{k=1}^{l} \sum_{j=1}^{|J_k|} w(y_{k_j}|x_i) \log \frac{w(y_{k_j}|x_i)|\mathcal{X}|}{\sum_{i=1}^{|\mathcal{X}|} w(y_{k_j}|x_i)} \\
&= \sum_{k=1}^{l} \sum_{j=1}^{|J_k|} w(y_{k_j}|x_i) \log \frac{w(y_{k_j}|x_i)|\mathcal{X}|}{\lambda_k} \\
&= \sum_{k=1}^{l} \sum_{j=1}^{|J_k|} w(y_{k_j}|x_i) (\log w(y_{k_j}|x_i)|\mathcal{X}| - \log \lambda_k) \\
&= \sum_{k=1}^{l} \mu_k - \log \lambda_k + \log |\mathcal{X}|,
\end{aligned}$$

which is not a function of i. In the calculations above, the second equality follows from assuming the uniform input distribution, the third equality follows from (2) and the final equality follows from (3).

(b) From KKT conditions we also know that the capacity is:

$$\begin{aligned}
C &= \sum_{j=1}^{|\mathcal{Y}|} w(y_j|x_i) \log \frac{w(y_j|x_i)}{\sum_{i=1}^{|\mathcal{X}|} p(x_i) w(y_j|x_i)} \\
&= \sum_{j=1}^{|\mathcal{Y}|} w(y_j|x_i) \log \frac{w(y_j|x_i)|\mathcal{X}|}{\sum_{i=1}^{|\mathcal{X}|} w(y_j|x_i)} \\
&= \sum_{j=1}^{|\mathcal{Y}|} w(y_j|x_i) \left(\log w(y_j|x_i) + \log |\mathcal{X}| - \log \sum_{i=1}^{|\mathcal{X}|} w(y_j|x_i) \right) \\
&= \sum_{j=1}^{|\mathcal{Y}|} w(y_j|x_i) \log w(y_j|x_i) + \log |\mathcal{X}| - \sum_{j=1}^{|\mathcal{Y}|} w(y_j|x_i) \log \sum_{i=1}^{|\mathcal{X}|} w(y_j|x_i) \\
&= \log |\mathcal{X}| - H_{|\mathcal{Y}|}(R_i) - \sum_{j=1}^{|\mathcal{Y}|} w(y_j|x_i) \log \sum_{i=1}^{|\mathcal{X}|} w(y_j|x_i),
\end{aligned}$$

where i is an index of an arbitrary input symbol, which is what needed to be proven.0

(The definition of the symmetric channel given in this problem states that rows of partitions are permutations of each other. From this it can be deduced that the rows of the transition matrix W are also permutations of each other and hence $H_{|\mathcal{Y}|}(R_i)$ is the same for all i.)

Problem 4

(a-b) Note that the covariance matrix of Z is $(1-\lambda)\mathbf{K}_1 + \lambda\mathbf{K}_2$. However, Z is not a Gaussian random variable. In fact, we are not going to find the distribution of Z, and knowing the covariance of Z suffices for upper bounding $h(Z)$. Since a Gaussian distribution maximizes the entropy for a given covariance matrix, we have

$$h(Z) \leq \frac{1}{2}\log(2\pi e)^n |(1-\lambda)\mathbf{K}_1 + \lambda\mathbf{K}_2|.$$

(c)

$$\begin{aligned} h(Z|\theta) &= h(Z|\theta=0)\Pr(\theta=0) + h(Z|\theta=1)\Pr(\theta=1) \\ &= (1-\lambda)\frac{1}{2}\log(2\pi e)^n |\mathbf{K}_1| + \lambda\frac{1}{2}\log(2\pi e)^n |\mathbf{K}_2| \end{aligned}$$

(d) We know that conditioning always reduces entropy. Hence, $h(Z) \geq h(Z|\theta)$. Replacing the upper bound of $h(Z)$ and the value of $h(Z|\theta)$, we get

$$|(1-\lambda)\mathbf{K}_1 + \lambda\mathbf{K}_2| \geq |\mathbf{K}_1|^{1-\lambda}|\mathbf{K}_2|^{\lambda},$$

which means that determinant is a concave function.