1) a) Using the fact that conditioning cannot increase entropy, we get that

\[ H(X_n|X_1, \ldots, X_{n-1}) \leq H(X_n|X_{n-i+1}, \ldots, X_{n-1}) = H(X_i|X_1, \ldots, X_{i-1}), \]

where the last equality follows from the stationarity property of the sequence \( \{X_i\}_{i=1}^{\infty} \).

b) From part (a), it follows that

\[ H(X_n|X^{n-1}) \leq \frac{1}{n-1} \sum_{i=1}^{n-1} H(X_i|X^{i-1}) = \frac{1}{n-1} H(X^{n-1}), \]

where the last equality follows from the chain rule. Then,

\[ H(X_n) = H(X^{n-1}) + H(X_n|X^{n-1}) \leq \frac{n}{n-1} H(X^{n-1}). \]

c) Using the chain rule and part (a), we get

\[ H(X_n) = \sum_{i=1}^{n} H(X_i|X^{i-1}) \geq \sum_{i=1}^{n} H(X_n|X^{n-1}) = n H(X_n|X^{n-1}). \]

2) We have

\[ H_{256}(P) = -\sum_{x \in \mathcal{X}} p(x) \log_{256} p(x) = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x) \log_{256}(2) = \frac{1}{8} H_2(X). \]

3) a) We have that \( H(X_1) = H(X_2) \) as \( X_1 \) and \( X_2 \) are identically distributed. Then,

\[ \rho = 1 - \frac{H(X_2|X_1)}{H(X_1)} = \frac{H(X_1) - H(X_2|X_1)}{H(X_1)} = \frac{H(X_2) - H(X_2|X_1)}{H(X_1)} = \frac{I(X_1; X_2)}{H(X_1)}. \]

b) From the non-negativity of Shannon entropy, it follows that \( \rho \leq 1 \). Moreover, using the fact that conditioning cannot increase entropy, we get \( H(X_2|X_1) \leq H(X_2) = H(X_1) \). Then,

\[ \rho = 1 - \frac{H(X_2|X_1)}{H(X_1)} \geq 1 - \frac{H(X_1)}{H(X_1)} = 0. \]

c) If \( H(X_2|X_1) = 0 \), then \( \rho = 0 \). That is, a sufficient condition for \( \rho = 1 \) is that \( X_2 = f(X_1) \) for some deterministic function \( f \).

4) a) Using the independence of \( X \) and \( Q \), we get

\[ I(X; Q, A) = I(X; Q) + I(X; A|Q) = I(X; A|Q). \]

Using the fact that \( A = a(X, Q) \), we get \( H(A|X, Q) = 0 \). Thus,

\[ I(X; A|Q) = H(A|Q) - H(A|X, Q) = H(A|Q). \]
b) We have

\[ I(X; Q_1, A_1, Q_2, A_2) = I(X; Q_1, A_1) + I(X; Q_2, A_2 | Q_1, A_1). \]

In order to prove that \( I(X; Q_1, A_1, Q_2, A_2) \leq 2I(X; Q_1, A_1) \), it suffices to show that \( I(X; Q_2, A_2 | Q_1, A_1) \leq I(X; Q_2, A_2) \). We have

\[ I(X; Q_2, A_2 | Q_1, A_1) = H(Q_2, A_2 | Q_1, A_1) - H(Q_2, A_2 | Q_1, A_1, X). \]

Using the fact that conditioning cannot increase entropy, we have \( H(Q_2, A_2 | Q_1, A_1) \leq H(Q_2, A_2) \). Moreover, using the fact that \((Q_1, A_1)\) and \((Q_2, A_2)\) are conditionally independent given \(X\), we have \( H(Q_2, A_2 | Q_1, A_1, X) = H(Q_2, A_2 | X) \). Then,

\[ I(X; Q_2, A_2 | Q_1, A_1) \leq H(Q_2, A_2) - H(Q_2, A_2 | X) = I(X; Q_2, A_2) = I(X; Q_1, A_1). \]

5) See Theorem 1 from notes of Lecture 13.

6) a) We only prove the upper bound as the lower bound follows from a similar argument. For all \( x^n \in \mathcal{T}_e^{(n)}(Q) \), we have

\[ P^n(\{x^n\}) = \prod_{i=1}^n p(x_i) = \prod_{a \in \mathcal{X}} p(a)^{\nu_a n(a)} \leq \prod_{a \in \mathcal{X}} p(a)^{n(1-\epsilon)q(a)}, \]

where the last equality follows from the definition of letter typical sets. Then,

\[ P^n(\{x^n\}) = \prod_{a \in \mathcal{X}} p(a)^{n(1-\epsilon)q(a)} = 2^{n(1-\epsilon)\sum_{a \in \mathcal{X}} q(a) \log(p(a))}, \]

Note that \( \sum_{a \in \mathcal{X}} q(a) \log(p(a)) = \sum_{a \in \mathcal{X}} q(a) \log(p(a)/q(a)) - \sum_{a \in \mathcal{X}} q(a) \log(q(a)) = -D_{\text{KL}}(P\|Q) - H(Q) \). Then,

\[ P^n(\{x^n\}) = 2^{-n(1-\epsilon)(D_{\text{KL}}(P\|Q)+H(Q))}. \]

Thus, using the union bound, we get

\[
P^n(\mathcal{T}_e^{(n)}(Q)) \leq |\mathcal{T}_e^{(n)}(Q)| 2^{-n(1-\epsilon)(D_{\text{KL}}(P\|Q)+H(Q))} \\
\leq 2^{n(1+\epsilon)H(Q) - n(1-\epsilon)(D_{\text{KL}}(P\|Q)+H(Q))} \\
= 2^{-n(D_{\text{KL}}(P\|Q)+\epsilon \sum_{a \in \mathcal{X}} q(a) \log(p(a)/q(a)))}.
\]

Using the fact that \( Q \ll P \), we have \( \mu_P, \mu_Q > 0 \) where \( \mu_Q = \min_{a \in \supp(Q)} q(a) \) and \( \mu_P = \min_{a \in \supp(Q)} p(a) \). Thus, by defining \( \delta(\epsilon) \), we get the desired upper bound.

\[ \delta(\epsilon) = -\epsilon \log(\mu_P \mu_Q). \]

b) Let \( Q = P_{XY} \) and \( P = P_x \otimes P_Y \). Clearly, \( Q \ll P \). Thus, the desired bounds follow from part (a) for

\[ \tilde{\delta}(\epsilon) = -\epsilon \log(\mu_{XY} \mu_X \mu_Y), \]

where \( \mu_X, \mu_Y, \mu_{XY} > 0 \) are defined as \( \mu_X = \min_{a \in \supp(P_X)} p_X(a) \), \( \mu_Y = \min_{a \in \supp(P_Y)} p_Y(a) \), and \( \mu_{XY} = \min_{a \in \supp(P_{XY})} p_{XY}(a) \).
7) Using the total probability theorem, we have

\[ P^{(c_n)}(y^n|x^n) = \sum_{m \in \mathcal{M}} P_M(m)P^{(c_n)}(y^n|m, x^n) = \sum_{m \in \mathcal{M}} P_M(m) \prod_{i=1}^{n} P^{(c_n)}(y_i|m, x^n, y^{i-1}). \]

We have

\[ P^{(c_n)}(y_i|m, x^n, y^{i-1}) = \frac{P^{(c_n)}(x_{i+1}, \ldots, x_n|m, x^i, y^i)P^{(c_n)}(y_i|m, x^i, y^{i-1})}{P^{(c_n)}(x_{i+1}, \ldots, x_n|m, x^i, y^{i-1})} = P^{(c_n)}(y_i|m, x^i, y^{i-1}). \]

where the equality follows from the fact that the channel is without feedback. Then, using the fact that the channel is memoryless, we get

\[ P^{(c_n)}(y_i|m, x^n, y^{i-1}) = P^{(c_n)}(y_i|m, x^i, y^{i-1}) = P_{Y|X}(y_i|x_i). \]

Thus,

\[ P^{(c_n)}(y^n|x^n) = \sum_{m \in \mathcal{M}} P_M(m) \prod_{i=1}^{n} P_{Y|X}(y_i|x_i) = \prod_{i=1}^{n} P_{Y|X}(y_i|x_i), \]

where the last equality follows from the fact that \( \prod_{i=1}^{n} P_{Y|X}(y_i|x_i) \) does not depend on \( m \) and that \( \sum_{m \in \mathcal{M}} P_M(m) = 1 \).

8) We have \( X \rightarrow Y \rightarrow E \). Then, \( I(X; E|Y) = 0 \) and \( I(X; Y) = I(X; Y|E) \). Moreover,

\[ I(X; Y, E) = I(X; E) + I(X; Y|E) = I(X; Y|E), \]

where the inequality follows from the fact that \( p_{E|X}(1|x) = p_E(1) = \alpha \) for \( x = 0, 1 \). Then,

\[ I(X; Y|E) = \alpha I(X; Y|E = 1) + (1 - \alpha)I(X; Y|E = 0) = (1 - \alpha)I(X; Y|E = 0) = (1 - \alpha)H(X). \]

Thus,

\[ \max_{P_X} I(X; Y) = \max_{P_X} (1 - \alpha)H(P_X) = 1 - \alpha, \]

where the maximum is achieved by \( p_X = \text{Unif}(\mathcal{X}). \)

9) We have that

\[ p_{Y|X}(y|x) = \begin{cases} 1/2, & y = x, y = \text{con}(x), x \in \mathcal{X}_m, \\ 0, & \text{otherwise}. \end{cases} \]

Then, \( H(P_{Y|X}(|X = x)) = 1 \) and \( H(Y|X) = \mathbb{E}[H(P_{Y|X}(|X))] = 1 \) for all \( P_X \). Thus,

\[ \max P_X I(X; Y) = \max P_X (H(Y) - H(Y|X)) = \max P_X H(Y) - 1. \]

To find the capacity, we need to find \( P_X \) that maximizes \( H(Y) \). From the fact that uniform distribution maximizes Shannon entropy, we choose a \( P_X \) that induces uniform distribution on \( Y \). If \( X \sim \text{Unif}(\mathcal{X}_m) \), then for all \( y \in \mathcal{Y}_out \), we
have

$$p_Y(y) = p_X(y)p_{Y|X}(y|y) + p_X(\text{con}^{-1}(y))p_{Y|X}(y|\text{con}^{-1}(y)) = \frac{1}{|\mathcal{X}_m|} \frac{1}{2} + \frac{1}{|\mathcal{X}_m|} \frac{1}{2} = \frac{1}{26}.$$ 

That is, if $X \sim \text{Unif}(\mathcal{X}_m)$, then $Y \sim \text{Unif}(\mathcal{Y}_{\text{out}})$. Thus, the channel capacity is $\max_{P_X} I(X;Y) = \log(26) - 1$ achieved when the input distribution is uniform.