Problem 1. (a) 

\[ Y_i = X_i \oplus Z_i, \]

where

\[ Z_i = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1 - p \end{cases} \]

and \( Z_i \) are not necessarily independent.

\[
I(X_1, \ldots, X_n; Y_1, \ldots, Y_n) = H(X_1, \ldots, X_n) - H(X_1, \ldots, X_n | Y_1, \ldots, Y_n) \\
\geq H(X_1, \ldots, X_n) - H(Z_1, \ldots, Z_n) \\
\geq H(X_1, \ldots, X_n) - \sum H(Z_i) \\
= H(X_1, \ldots, X_n) - nH(p) \\
\geq H(X_1, \ldots, X_n) - nH(p) = nC.
\]

if \( X_1, \ldots, X_n \) are chosen i.i.d. \( \sim \text{Bern}(1/2) \). The capacity of the channel with memory over \( n \) uses of the channel is

\[
nC^{(n)} = \max_{p(x_1, \ldots, x_n)} I(X_1, \ldots, X_n; Y_1, \ldots, Y_n) \\
\geq I(X_1, \ldots, X_n; Y_1, \ldots, Y_n)_{p(x_1, \ldots, x_n) = \text{Bern}(1/2)} \\
\geq n(1 - H(p)) = nC.
\]

Hence channels with memory have higher capacity. The intuitive explanation for this result is that the correlation between the noise decreases the effective noise; one could use the information from the past samples of the noise to combat the present noise.

(b) (i) We will prove by induction that \( \forall i, \Pr(Z_i = 1) = \frac{1}{2} \). Notice that the result holds for \( i = 1 \) (from assumption). Assuming the result holds true for all \( Z_i, 1 \leq i \leq k - 1 \), we have that

\[
\Pr(Z_k = 1) = \Pr(Z_k = 1, Z_{k-1} = 0) + \Pr(Z_k = 1, Z_{k-1} = 1) \\
= \Pr(Z_k = 1 | Z_{k-1} = 0) \Pr(Z_{k-1} = 0) + \Pr(Z_k = 1 | Z_{k-1} = 1) \Pr(Z_{k-1} = 1) \\
= q \left( \frac{1}{2} \right) + (1 - q) \frac{1}{2} \\
= \frac{1}{2}
\]

(ii) The (a) follows from the fact that \( Z_i^n \) is a deterministic function of the \( X_i^n, Y_1^n \) and that \( Y_i^n \) is again a deterministic function of the \( X_i^n, Z_i^n \). Therefore we have

\[
H(Y_i^n, Z_i^n | X_i^n) = H(Y_i^n | X_i^n) + H(Z_i^n | X_i^n, Y_i^n) \\
= H(Y_i^n | X_i^n)
\]
Expanding the term $H(Y^n_1, Z^n_1 | X^n_1)$ once again in the other way

$$H(Y^n_1, Z^n_1 | X^n_1) = H(Z^n_1 | X^n_1) + H(Y^n_1 | X^n_1, Z^n_1) = H(Z^n_1 | X^n_1)$$

So $H(Y^n_1 | X^n_1) = H(Z^n_1 | X^n_1)$. Furthermore, since $Z^n_1$ is independent of the input sequence $X^n_1$, we have that $H(Y^n_1 | X^n_1) = H(Z^n_1)$.

The $(b)$ follows from the chain rule for expanding the joint entropy and from the Markov property of the $Z^n_1$ sequence.

$$H(Z^n_1) = H(Z_1) + \sum_{1}^{n} H(Z_{i+1}|Z_1, \ldots, Z_i)$$

$$= H(Z_1) + \sum_{1}^{n} H(Z_{i+1}|Z_i)$$

The $(c)$ follows since $H(Y_1, \ldots, Y_n) \leq \sum_{1}^{n} H(Y_i) \leq n$. Here, the last inequality follows since the $Y_i$’s are binary random variables. Note: The evaluation of the entropy is made in bits (base log 2).

The upper bound can be achieved if we find a distribution on $X^n_1$ that gives an i.i.d. $\sim$ Bern($\frac{1}{2}$) distribution on the output $Y^n_1$. This distribution is again the i.i.d. $\sim$ Bern($\frac{1}{2}$) distribution i.e., $p(X^n_1) = \prod_{1}^{n} p(X_1)$ where $p$ is the Bernoulli($\frac{1}{2}$) distribution.

**Problem 2.** To find the capacity of the product channel, we must find the distribution $p(x_1, x_2)$ on the input alphabet $\mathcal{X}_1 \times \mathcal{X}_2$ that maximizes $I(X_1, X_2; Y_1, Y_2)$. Since the joint distribution

$$p(x_1, x_2, y_1, y_2) = p(x_1, x_2)p(y_1|x_1)p(y_2|x_2),$$

$Y_1 \rightarrow X_1 \rightarrow X_2 \rightarrow Y_2$ forms a Markov chain and therefore

$$I(X_1, X_2; Y_1, Y_2) = H(Y_1, Y_2) - H(Y_1, Y_2|X_1, X_2)$$

$$= H(Y_1, Y_2) - H(Y_1|X_1, X_2) - H(Y_2|X_1, X_2) \quad (1)$$

$$= H(Y_1, Y_2) - H(Y_1|X_1) - H(Y_2|X_2) \quad (2)$$

$$\leq H(Y_1) + H(Y_2) - H(Y_1|X_1) - H(Y_2|X_2) \quad (3)$$

$$= I(X_1; Y_1) + I(X_2; Y_2) \quad (4)$$

where (2) and (3) follow from Markovity, and we have equality in (4) if $Y_1$ and $Y_2$ are independent. Equality occurs when $X_1$ and $X_2$ are independent. Hence

$$C = \max_{p(x_1, x_2)} I(X_1, X_2; Y_1, Y_2)$$

$$\leq \max_{p(x_1, x_2)} I(X_1; Y_1) + \max_{p(x_1, x_2)} I(X_2; Y_2)$$

$$= \max_{p(x_1)} I(X_1; Y_1) + \max_{p(x_2)} I(X_2; Y_2)$$

$$= C_1 + C_2.$$

with equality iff $p(x_1, x_2) = p^*(x_1)p^*(x_2)$ and $p^*(x_1)$ and $p^*(x_2)$ are the distributions for which $C_1 = I(X_1; Y_2)$ and $C_2 = I(X_2; Y_2)$ respectively.
PROBLEM 3. The assertion is clearly true with $n = 1$. To complete the proof by induction we need to show that the cascade of a BSC with parameter $q = \frac{1}{2}(1 - (1 - 2p)^n)$ with a BSC with parameter $p$ is equivalent to a BSC with parameter $\frac{1}{2}(1 - (1 - 2p)^{n+1})$. To do so, observe that for a cascade of a BSC with parameter $q$ and a BSC with parameter $p$, when a bit is sent, the opposite bit will be received if exactly one of the channels makes a flip, and this happens with probability $(1 - q)p + (1 - p)q$. Thus, the cascade is equivalent to a BSC with this parameter. For $q = \frac{1}{2}(1 - (1 - 2p)^n)$,

$$(1 - q)p + (1 - p)q = \frac{1}{2}(1 + (1 - 2p)^n)p + \frac{1}{2}(1 - (1 - 2p)^n)(1 - p) = \frac{1}{2}(1 - (1 - 2p)^{n+1}),$$

and the assertion is proved.

Alternate proof: the cascade makes flips the incoming bit if an odd number of the elements of the cascade flip. Thus the cascade is equivalent to a BSC with parameter

$$a = \sum_{k \text{ even}} \binom{n}{k} p^k (1 - p)^{n-k}.$$  

Let $b = \sum_{k \text{ odd}} \binom{n}{k} p^k (1 - p)^{n-k}$. Observe that

$$a + b = \sum_k \binom{n}{k} p^k (1 - p)^{n-k} = (p + (1 - p))^n = 1,$$

and

$$-a + b = \sum_k \binom{n}{k} (-p)^k (1 - p)^{n-k} = (-p + 1 - p)^n = (2 - 2p)^n.$$

Subtracting the two equalities and dividing by two, we get $a = \frac{1}{2}(1 + (1 - 2p)^n)$.

PROBLEM 4. Let $P'_{X,Y}(x, y) = P_{Y|X}(y|x)Q'(x)$, $P'_Y(y) = \sum_{x \in X} P'_{X,Y}(x, y)$ and $P_Y(y) = \sum_{x \in X} P_{Y|X}(y|x)Q(x)$. We then have for any $Q'$

$$\sum_{x \in X} Q'(x) \sum_{y \in Y} P_{Y|X}(y|x) \log \left( \frac{P_{Y|X}(y|x)}{\sum_{x' \in X} P_{Y|X}(y|x') Q(x')} \right) - I(Q')$$

$$= E_{P'_X,Y} \log \frac{P_{Y|X}}{P_Y} - I(Q')$$

$$= E_{P'_X,Y} \left( \log \frac{P_{Y|X}}{P_Y} - \log \frac{P'_{X,Y}}{Q_X P_Y} \right)$$

$$= E_{P'_X,Y} \log \frac{P_Y'}{P_Y}$$

$$= E_{P'_Y} \log \frac{P_Y'}{P_Y}$$

$$= D(P'_Y||P_Y) \geq 0$$

with equality if and only if $Q' = Q$. To prove (b), notice in the upper bound of part (a), that the inner summation is a function of $x$ and that the outer summation is an average of this function with respect to the distribution $Q'(x)$. The average of a function is upper bounded by the maximum value that the function takes, and hence (b) follows.