EE 276 - Information Theory
Final
March 20, 2026

1. You have 3 hours to take the exam. There are a total of 4 questions and 100 points, with potential 6 additional bonus points. Questions have different numbers of points as indicated before each sub-problem.
2. Write your SUID (8-digit number) in the box on the top right corner of each page.
3. Please write your answer in the designated box underneath each question. **Answers written outside the box will not be graded.**
4. If you need more space for your work, you may use pages 18-21. In the original problem space, **you must indicate that you have used this extra space in order for it to be graded.**
5. All answers should be justified, unless otherwise stated.
6. Even if you did not solve a problem/subproblem, you **may use** its result in later problems/subproblems.
7. The exam is closed book/notes/electronic devices/etc. but you are allowed **two** double-sided sheets of handwritten notes. No other materials/resources are allowed.
8. Calculators are not allowed.
9. You should use $\log = \log_2$ (as opposed to ln or $\log_{10}$).
10. Do not discuss the contents of the exam with anyone who has not yet taken it.

Good luck!

Full Name: _____

Email: _____

SUID: _____

I have abided with both the letter and spirit of the Stanford Honor Code. I have neither given nor received unpermitted aid on this examination.

Signature: _____

1. 37 total points **Channel with a Side Switch.**
Suppose we have a channel with binary input X, binary output Y, and a binary side switch with position represented by U. The output is the result of passing $V := X \oplus U$ (where $\oplus$ is addition mod 2) through the (memoryless) Z-channel with parameter p. In other words, $\mathcal{X} = \mathcal{U} = \mathcal{V} = \mathcal{Y} = \{0, 1\}$, and for all n we have $\Pr(Y^n = y^n | X^n = x^n, U^n = u^n) = \prod_{i=1}^n p_{Y|V}(y_i | x_i \oplus u_i)$. Recall the Z-channel is defined by

$$p_{Y|V}(y|v) = \begin{bmatrix} 1 & 0 \\ 1-p & p \end{bmatrix}. \tag{1}$$

As per the standard communication setting, the sender (encoder) chooses the input X^n based on the message bits to be communicated B^m, while the receiver (decoder) observes the output Y^n. We will vary who has knowledge and control of the switch U^n.

All subparts can be attempted independently, though you may need to refer to facts you are asked to establish in preceding subparts.

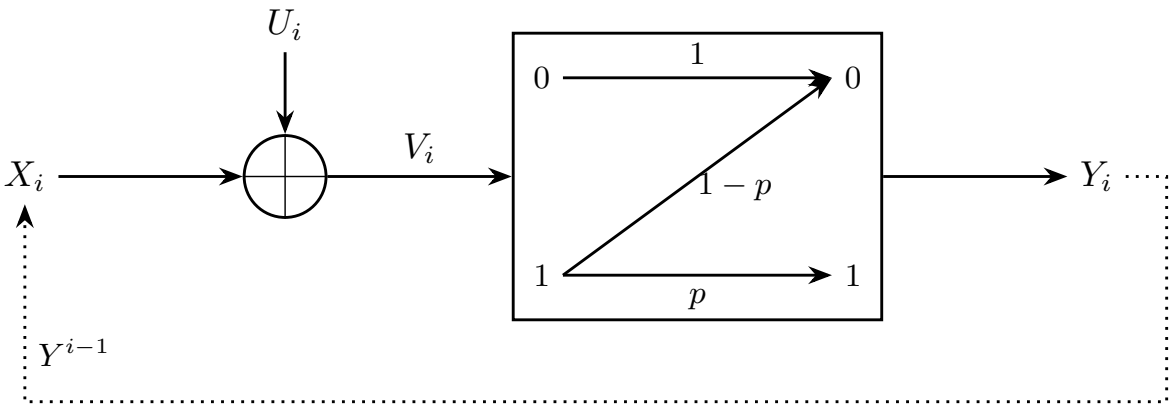


Figure 1: Dotted line is feedback occuring in part (e).

(a) [7 points] Show that the Z-channel from V to Y has capacity

$$C = \log\left(1 + p \cdot (1-p)^{(1-p)/p}\right). \tag{2}$$

Solution: Let $V = \text{Ber}(q)$.

$$\begin{aligned} I(V;Y) &= H(Y) - H(Y|V) && (1)\\ &= h_2(pq) - qh_2(p) && (2)\\ &= -pq\log(pq) - (1-pq)\log(1-pq) - qh_2(p) && (3) \end{aligned}$$

Since the binary entropy function h_2 is concave, so is $I(V;Y)$. Clearly, setting $q = 0, 1$ result in $I(V;Y) = 0$. Then the maximum is achieved when the derivative is equal to 0:

$$\begin{aligned} \frac{d}{dq}I(V;Y) = -p\log(pq) - p + p\log(1-pq) + p - h_2(p) = 0 && (4)\\ \iff -p\log(pq) + p\log(1-pq) = h_2(p) && (5)\\ \iff \frac{1}{pq} - 1 = 2^{h_2(p)/p} && (6)\\ \iff q = \frac{1}{p\left(2^{h_2(p)/p}+1\right)} && (7) \end{aligned}$$

Then

$$\begin{aligned} C &= h_2\left(\frac{1}{(2^{h_2(p)/p}+1)}\right) - \frac{h_2(p)}{p\left(2^{h_2(p)/p}+1\right)} && (8)\\ &= \frac{1}{(2^{h_2(p)/p}+1)}\log\left(2^{h_2(p)/p}+1\right) + \frac{2^{h_2(p)/p}}{2^{h_2(p)/p}+1}\log\frac{2^{h_2(p)/p}+1}{2^{h_2(p)/p}} - \frac{h_2(p)}{p\left(2^{h_2(p)/p}+1\right)} && (9)\\ &= \log\left(2^{-h_2(p)/p}+1\right) && (10)\\ &= \log\left(1 + p(1-p)^{(1-p)/p}\right) && (11) \end{aligned}$$

(b) [5 points] Suppose the U_i are i.i.d. $\sim$ Ber(1/2) (independent of the message bits B^m), and known to neither the encoder nor the decoder. Show that the capacity of the channel from X to Y is $C = 0$.

Solution: We note that we consider the channel $p(y|x) = \sum_{u=0,1} p(u|x)p(y|x,u) = \sum_{u=0,1} \frac{1}{2}p(y|x,u)$. Computing, we get the channel matrix

$$\begin{bmatrix} 1-\frac{p}{2} & \frac{p}{2} \\ 1-\frac{p}{2} & \frac{p}{2} \end{bmatrix} \tag{1}$$

Then Y is independent of X and thus the capacity is 0.

Alternate Solution: we can bound

$$C = \max_{p(x)} I(X;Y) \leq \max_{p(x)} I(X;V) = 0 \tag{2}$$

where the inequality follows by the data processing inequality for the mutual information and $I(X;V) = 0$ because V is the result of passing X through a BSC(1/2) channel. Since the capacity is nonnegative, we conclude it is 0.

(c) [5 points] Suppose now that the U_i are as in the previous part, but **known** to the sender and receiver. That is, the sender can choose $X^n = f(U^n, B^m)$ for some function f, and the receiver receives (Y^n, U^n). What is the maximum achievable rate of reliable communication? Your answer should be non-zero.

Solution: Consider $\overline{X}^n = X^n \oplus U^n$. Given U^n, it is possible to convert from X^n to $\overline{X}^n$ and vice versa. The channel from $\overline{X}^n$ to Y^n is a DMC and is in fact the Z channel, so the maximum achievable rate is the capacity of the Z channel from part (a). Note we have to argue reductions both ways to extend the converse (upper bound) for the Z-channel to this case.

Alternate Solution We note that given U^n, we have a DMC.

$$C = \max_{p(x|u)} I(X;Y|U) = \max_{p(x|u)} I(V;Y|U) = \max_{p(x|u)} I(V;Y) \tag{1}$$

The first equality is a slightly stronger version of the channel coding theorem, which was accepted for grading. The second equality follows from the fact that V_i and X_i are deterministic functions of each other given U_i. The third equality follows from the fact that we have the Markov chain $U \to V \to Y$. Finally, we note that we can achieve any distribution on V by choosing $p(x|u)$, so the capacity is the capacity of the Z-channel.

(d) [6 points] This part is **standalone**, and does not depend on the setting of the problem. Suppose $S \sim \text{Ber}(1/2)$. Suppose $T_1, T_2, \ldots, T_n$ are i.i.d. $\text{Ber}(q_0)$ if $S = 0$ and i.i.d. $\text{Ber}(q_1)$ if $S = 1$, where $0 \le q_0 < q_1 \le 1$. Show that it is possible to estimate S from T^n with probability of error that, for sufficiently large n, is no more than $\exp(-\alpha n)$, where $\alpha > 0$.

*Hint: It suffices to analyze the maximum likelihood estimate where $\hat{S}(t^n) = 0$ if $\Pr(T^n = t^n | S = 0) > \Pr(T^n = t^n | S = 1)$ and $\hat{S}(t^n) = 1$ otherwise. You may assert without proof that it **suffices** to bound the probability of error when $S = 0$. You may find it helpful to show that $\Pr(T^n = t^n | S = 0) < \Pr(T^n = t^n | S = 1)$ iff $\frac{1}{n}\sum_i t_i \ge q^*$ for some threshold q^* depending only on q_0, q_1.*

Solution: We proceed by the hint.

$$\Pr(T^n = t^n | S = 0) = q_0^{N(1|t^n)}(1-q_0)^{N(0|t^n)} \tag{1}$$
$$\Pr(T^n = t^n | S = 1) = q_1^{N(1|t^n)}(1-q_1)^{N(0|t^n)} \tag{2}$$

Then

$$\Pr(T^n = t^n | S = 0) < \Pr(T^n = t^n | S = 1) \tag{3}$$
$$\iff q_0^{N(1|t^n)}(1-q_0)^{N(0|t^n)} < q_1^{N(1|t^n)}(1-q_1)^{N(0|t^n)} \tag{4}$$
$$\iff N(1|t^n)\log\left(\frac{q_0}{q_1}\right) + N(0|t^n)\left(\frac{1-q_0}{1-q_1}\right) < 0 \tag{5}$$
$$\iff \sum_i t_i \log\left(\frac{q_0}{q_1}\right) + (m - \sum_i t_i)\log\left(\frac{1-q_0}{1-q_1}\right) < 0 \tag{6}$$
$$\iff \log\left(\frac{q_0(1-q_1)}{(1-q_0)q_1}\right)\frac{1}{m}\sum_i t_i \le \log\left(\frac{1-q_1}{1-q_0}\right) \tag{7}$$
$$\iff \frac{1}{n}\sum_i t_i \ge \frac{\log\left(\frac{1-q_1}{1-q_0}\right)}{\log\left(\frac{q_0(1-q_1)}{(1-q_0)q_1}\right)} \tag{8}$$

We note that the factor on the left is negative by assumption $q_0 < q_1$. We are done with the first part of the hint. To proceed, T^n is i.i.d. conditioned on the value of S. Then, we can apply Sanov's theorem. Then, $\Pr(\frac{1}{n}\sum_i T_i \ge q^* | S = 0) \le (n+1)^2 \exp(-D(q^*||q_0)n)$. We can set α to be the relative entropy which is nonnegative. The $(n+1)^2$ can be handled by decreasing α by some small $\epsilon > 0$, which is okay as the bound only needs to hold eventually in n.

(e) [6 points] Suppose someone has **locked** the switch into one random position. In other words, $U_1 = U_2 = \cdots = U_n = U$ and $U \sim \text{Ber}(1/2)$. As in part (b), neither the sender nor receiver has access to U. However, the sender does have access to **noiseless feedback** of the previous channel outputs. In other words, the sender can choose each input as a function of the message and past outputs $X_i = f_i(Y^{i-1}, B^m)$. Argue why the maximal rate of reliable communication for this setting is the same as that in part (c).
Hint: Let the sender send 0 *for the first* $m = \lceil \sqrt{n} \rceil$ *channel uses and apply Part (d).*

Solution: Fix an $R < C$, where C is the capacity of the Z channel as in part (c). Fix a sequence of schemes that achieve R for the setup in part (c). For a block of length n, we do the scheme as in the hint. Set $n' = \lceil \sqrt{n} \rceil$, $S = U$, $T^{n'} = Y^{n'}$, $q_0 = 0$, $q_1 = 1$. Applying part (d), we can find $\hat{U}$ with $\Pr(\hat{U} \neq U) \to 0$ as $n \to \infty$. Since we have feedback, both the sender and receiver can compute $\hat{U}$. For the remaining $n - \lceil \sqrt{n} \rceil$ channel uses, the sender and receiver follow the scheme for $n' = n - m$, acting as if $U_m = \cdots = U_n = \hat{U}$. The new scheme operates at rate $\frac{n - \lceil \sqrt{n} \rceil}{n} R \to R$. The probability of error is at most the probability $U \neq \hat{U}$ plus the probability of error for the scheme from part (c), both of which vanish. We have thus demonstrated that R is achievable.

Since we are free to ignore U, the capacity is higher when U is known. Any scheme for known $U_1 = U_2 = \cdots = U_n = U$ can be converted into one when U^n is known but random, simply by XORing with the sequence of U^n. Thus, a converse for the known U^n case with feedback is a converse for the unknown U case with feedback. We can apply the fact that the feedback capacity is the information capacity, which was computed in part (a).

(f) [4 points] Your friend has the following complaint about the result in the part (e): "In class we learned that feedback does not increase capacity, and we know from part (b) that, when U is available to neither the sender nor receiver, the capacity is 0. Yet you claim to achieve a positive rate with feedback." Explain briefly why there is **no** contradiction.

Solution: The theorem on feedback capacity holds only for discrete memoryless channels, but the channel is not memoryless. The fact that U is unchanged throughout the block means different channel uses are dependent.

Note that contrary to many answers, feedback does increase capacity in the locked U setting. Without knowledge of U, the sender is unable to ensure V has capacity achieving distribution for the Z-channel, which is in general not uniform.

(g) [4 points] Suppose the sender is allowed to toggle the switch for each channel use. In other words, the sender chooses $X^n = f(B^m), U^n = g(B^m)$ for some functions f, g. Note that this is a **no-feedback** setting. What is the maximum achievable rate of communication in bits per channel use?

Solution: Same as Z channel again. For achievability we can set $U_i = 0$ for all i set X^n to be the output of an encoder for the Z-channel, and for the converse we can apply the data processing inequality to upper bound the capacity by that of the Z-channel.

Alternate Solution By channel coding theorem

$$\begin{aligned} C &= \max_{p_{U,X}} I(X,U;Y) && (1)\\ &= \max_{p_{U,X}} I(X;Y) + I(U;Y|X) && (2)\\ &\geq \max_{p_{U,X}} I(X;Y) \geq \log(1+p(1-p)^{(1-p)/p} && (3)\\ C &= \max_{p_{U,X}} I(X,U;Y) && (4)\\ &\leq \max_{p_{U,X}} I(V;Y) = \log(1+p(1-p)^{(1-p)/p} && (5) \end{aligned}$$

2. [30 total points] **Rate-Distortion under Log Loss**
Recall that in rate-distortion theory, the reconstruction can take values in an alphabet that may differ from that of the source. In this problem, we study the case of a finite source alphabet and a reconstruction alphabet comprising the set of distributions over the source alphabet, under the **log loss** distortion criterion. Specifically, the log loss between a source symbol $x \in \mathcal{X}$ and a reconstruction, which is a PMF $q \in \mathcal{M}(\mathcal{X})$ (the space of distributions on $\mathcal{X}$), is given by

$$d(x, q) = \log \frac{1}{q(x)}$$

(a) [5 points] Suppose $X \sim \text{Ber}(\frac{1}{4})$. Consider a **fixed** reconstruction q parametrized by θ, given by $q(1) = \theta$, $q(0) = 1 - \theta$. Find the value θ^* that minimizes the expectation $\mathbb{E}[d(X, q)]$. Can you interpret the relation between the minimizing distribution and the source distribution?

Solution:

$$\begin{aligned}\mathbb{E}[d(X,q)] &= P(X=0) \cdot \log \frac{1}{q(0)} + P(X=1) \cdot \log \frac{1}{q(1)} \\ &= -\frac{1}{4} \log \theta - \frac{3}{4} \log(1-\theta)\end{aligned}$$

Differentiating with respect to θ and setting equal to zero:

$$\frac{d}{d\theta}\mathbb{E}[d(X,q)] \;=\; -\frac{1}{4\theta \ln 2} + \frac{3}{4(1-\theta)\ln 2} = 0 \implies 3\theta = 1-\theta \implies \boxed{\theta^* = \tfrac{1}{4}.}$$

The second derivative is positive, confirming this is a minimum. Substituting $\theta^* = 1/4$:

$$\mathbb{E}[d(X,q^*)] = \frac{1}{4}\log 4 + \frac{3}{4}\log\frac{4}{3} = -\frac{1}{4}\log\frac{1}{4} - \frac{3}{4}\log\frac{3}{4} = H(X).$$

Note that $\theta^* = 1/4 = p(1)$, i.e. the optimal fixed reconstruction is exactly the source distribution.

(b) [5 points] Let X have an arbitrary PMF p over $\mathcal{X}$, and let $q \in \mathcal{M}(\mathcal{X})$ be a **fixed** reconstruction distribution. Prove that

$$\mathbb{E}[d(X,q)] \;=\; H(X) + D(p \,\|\, q),$$

and use this to show that $\mathbb{E}[d(X,q)] \geq H(X)$, with equality if and only if $q = p$.

Solution:

Expanding the expectation directly:

$$\mathbb{E}[d(X,q)] = \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{q(x)}.$$

Add and subtract $\sum_x p(x) \log \frac{1}{p(x)}$:

$$= \sum_x p(x) \log \frac{1}{p(x)} + \sum_x p(x) \log \frac{p(x)}{q(x)} = H(X) + D_{\mathrm{KL}}(p\|q).$$

Since $D_{\mathrm{KL}}(p\|q) \geq 0$ with equality iff $p = q$, we conclude

$$\mathbb{E}[d(X,q)] \;\geq\; H(X), \qquad \text{with equality iff } q = p.$$

Therefore, even the best possible fixed reconstruction cannot achieve expected log loss below $H(X)$. The optimal fixed reconstruction is the source distribution p itself.

(c) [6 points] Now consider Q being a **random** element of $\mathcal{M}(\mathcal{X})$, jointly distributed with X. Write $p_{X|Q=q}$ for the conditional PMF of X given that the reconstruction random variable Q is a specific reconstruction distribution q.

Prove the following identity:

$$\mathbb{E}[d(X,Q)] \;=\; H(X \mid Q) \;+\; \mathbb{E}[f(Q)]\,,$$

where $f(q) = D(p_{X|Q=q}\|q) = \sum_{a \in \mathcal{X}} p(a|q) \log \frac{p(a|q)}{q(a)}$ with $p(a|q) = p_{X|Q=q}(a) \quad \forall a \in \mathcal{X}$.

Hint: Condition on Q, and apply the result from part (b) to the inner expectation using the tower property.

Solution:
Apply the tower property (law of total expectation) by conditioning on Q:

$$\mathbb{E}[d(X,Q)] \;=\; \mathbb{E}_Q\Big[\,\mathbb{E}_{X|Q}\Big[\log\frac{1}{Q(X)}\;\Big|\;Q\Big]\Big]\,.$$

For a fixed realization $Q=q$, the inner conditional expectation is

$$\mathbb{E}_{X|Q=q}\Big[\log\frac{1}{q(X)}\Big]=\sum_{a\in\mathcal{X}}p(a|q)\log\frac{1}{q(a)},$$

Splitting $\log\frac{1}{q(a)}=\log\frac{1}{p(a|q)}+\log\frac{p(a|q)}{q(a)}$ (exactly as in Part (b)):

$$\mathbb{E}_{X|Q=q}\Big[\log\frac{1}{q(X)}\Big] \;=\; H(X|Q=q)+D_{\mathrm{KL}}(p_{X|Q=q}\|q).$$

Taking the expectation over Q:

$$\mathbb{E}[d(X,Q)] \;=\; H(X|Q) \;+\; \mathbb{E}_Q\big[D_{\mathrm{KL}}(p_{X|Q}\|Q)\big]\,.$$

(d) [4 points] Using the identity from Part (c), prove that for **any** joint distribution between X and Q satisfying the distortion constraint $\mathbb{E}[d(X,Q)]\le D$, we have that

$$I(X;Q) \;\ge\; H(X)-D.$$

Solution: From the identity in Part (c), we have that $\mathbb{E}[d(X,Q)] \ge H(X \mid Q)$, by the non-negativity of KL-Divergence. Therefore, given the constraints of the problem, $H(X \mid Q)\le D$. Substituting in the expression for mutual information,

$$\begin{aligned} I(X;Q) &= H(X)-H(X\mid Q)\\ &\ge H(X)-D. \end{aligned}$$

(e) [4 points] Consider the following two joint distributions for (X, Q):

(i) The reconstruction $Q = \delta_X$, i.e., if $X = a$, the reproduction distribution $q(x) = \delta_a$ (the degenerate PMF assigning the symbol a probability 1).

(ii) The reconstruction $Q = p_X$, i.e., the fixed PMF of X.

Argue that the mutual information $I(X;Q)$ and distortion $\mathbb{E}[d(X,Q)]$ of Scheme (i) is $(H(X), 0)$ and that of Scheme (ii) is $(0, H(X))$.

Solution:

(i) The distortion is $d(x, \delta_x) = -\log \delta_x(x) = -\log 1 = 0$. Since Q reveals X exactly, $H(X|Q) = 0$, $I(X;Q) = H(X)$.

(ii) The expected distortion is $\mathbb{E}[d(X, p_X)] = \mathbb{E}[-\log p_X] = H(X)$. Since the reconstruction Q is fixed at p_X and independent of the value of X, $I(X;Q) = 0$.

(f) [6 points] Given the results of Part (d) and Part (e), argue why the rate distortion function for the source X under log loss is given, for **any** value of $D \in [0, H(X)]$, by

$$R(D) = \min_{p_{Q|X}: \mathbb{E}[d(X,Q)] \leq D} I(X;Q) = H(X) - D.$$

Hint: Consider time-sharing between the two extremes in Part (e).

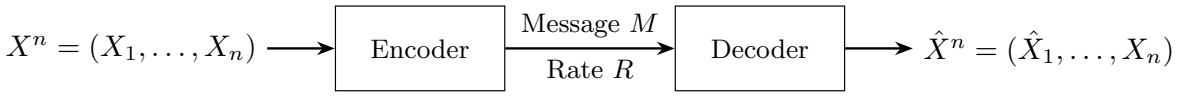
Solution:

Given a value of $D \in [0, H(X)]$, take $\alpha = 1 - \frac{D}{H(X)}$. Consider a scheme which, with probability α, uses Scheme (i), and with probability $1-\alpha$, uses Scheme (ii). This scheme has distortion $\alpha \cdot 0 + (1-\alpha) \cdot H(X) = D$ and rate $\alpha \cdot H(X) + (1-\alpha) \cdot 0 = H(X) - D$.

Therefore, this scheme satisfies the distortion constraint $\mathbb{E}[d(X,Q)] \leq D$ and has rate $H(X) - D$. Since the minimum value of the rate $I(X;Q)$ in Part (d) is achieved, this is also the value of the rate distortion function.

$$R(D) = \begin{cases} H(X) - D & D \leq H(X) \\ 0 & D > H(X) \end{cases}.$$

3. 18 total points **From Fano's Inequality to Source Coding Converses**
Suppose we have a sequence X^n whose components are generated i.i.d. from a source X with finite alphabet $\mathcal{X}$. We encode this sequence to a message M, which is one of 2^{nR} possible codewords. On decoding M, we obtain the sequence $\hat{X}^n$. Define $P_e^{(n)} = \mathbb{P}\{X^n \neq \hat{X}^n\}$.



(a) 5 points Suppose $R < H(X)$, i.e. $R = H(X) - \epsilon$ for $\epsilon > 0$.
Show that $H(X^n \mid M) \geq n\epsilon$.

Solution:

$$\begin{aligned} H(X^n \mid M) &= H(X^n) - I(X^n; M) \\ &\geq H(X^n) - H(M) \\ &\geq nH(X) - nR \\ &= n\epsilon \end{aligned}$$

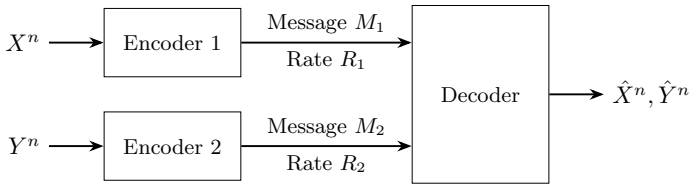
(b) 5 points For $R = H(X) - \epsilon$ as in Part (a), show that it **cannot** hold that $\lim_{n\to\infty} P_e^{(n)} = 0$, i.e., we cannot have (near) lossless compression.

Solution: From Fano's Inequality,

$$\begin{aligned} P_e^{(n)} &\geq \frac{H(X^n \mid M) - 1}{\log(|\mathcal{X}|^n)} \\ &\geq \frac{n\epsilon - 1}{n \log |\mathcal{X}|} \\ &= \frac{\epsilon}{\log |\mathcal{X}|} - \frac{1}{n \log |\mathcal{X}|} \xrightarrow{n\to\infty} \frac{\epsilon}{\log |\mathcal{X}|} > 0. \end{aligned}$$

Since the probability of error does not decay to 0, we do not have lossless compression. Therefore, we cannot achieve rates $R < H(X)$.

Now consider the distributed source coding problem. The pair of sequences (X^n, Y^n) is generated via i.i.d. drawings of pairs (X_i, Y_i) from the joint source (X, Y) with finite alphabets $\mathcal{X}$ and $\mathcal{Y}$, respectively. The sequence X^n is encoded to M_1, which is one of 2^{nR_1} codewords, and the sequence Y^n is separately encoded to M_2, which is one of 2^{nR_2} codewords. The pair (M_1, M_2) is jointly decoded to the pair of sequences $(\hat{X}^n, \hat{Y}^n)$. Define $P_e^{(n)} = \mathbb{P}\{(X^n, Y^n) \neq (\hat{X}^n, \hat{Y}^n)\}$.



(c) [8 points] Suppose $R_1 < H(X \mid Y)$, i.e., $R_1 = H(X \mid Y) - \epsilon$ for $\epsilon > 0$. Show that it **cannot** hold that $\lim_{n\to\infty} P_e^{(n)} = 0$, i.e., we cannot have (near) lossless compression. *Hint: Just as in part (a), try and lower bound $H(X^n, Y^n \mid M_1, M_2)$ away from 0. You might find it useful to lower bound $H(X^n \mid M_1, Y^n)$.*

Solution:

$$\begin{aligned}
H(X^n, Y^n \mid M_1, M_2) &= H(X^n \mid M_1, M_2, Y^n) + H(Y^n \mid M_1, M_2) \\
&\geq H(X^n \mid M_1, M_2, Y^n) \\
&= H(X^n \mid M_1, Y^n) \\
&= H(X^n \mid Y^n) - I(X^n; M_1 \mid Y^n) \\
&\geq H(X^n \mid Y^n) - H(M_1) \\
&\geq nH(X \mid Y) - nR_1 \\
&= n\epsilon
\end{aligned}$$

From Fano's Inequality,

$$\begin{aligned}
P_e^{(n)} &\geq \frac{H(X^n, Y^n \mid M_1, M_2) - 1}{\log(|\mathcal{X}|^n |\mathcal{Y}|^n)} \\
&\geq \frac{n\epsilon - 1}{n(\log |\mathcal{X}||\mathcal{Y}|)} \\
&= \frac{\epsilon}{\log |\mathcal{X}||\mathcal{Y}|} - \frac{1}{n \log |\mathcal{X}||\mathcal{Y}|} \xrightarrow{n\to\infty} \frac{\epsilon}{\log |\mathcal{X}||\mathcal{Y}|} > 0.
\end{aligned}$$

Therefore, we cannot achieve rates with $R_1 < H(X \mid Y)$. By a symmetrical argument, we can say that we cannot achieve rates with $R_2 < H(Y \mid X)$.

4. [15 total points] **Strong Data Processing Inequality**

Consider PMFs p_X, q_X on the finite alphabet $\mathcal{X}$ and a channel (or conditional PMF) $p_{Y|X}$ taking $\mathcal{X}$ to the finite alphabet $\mathcal{Y}$. Recall (from the midterm) that the data-processing inequality for KL divergence (relative entropy) tells us that

$$D(p_X||q_X) \geq D(p_Y||q_Y),$$

where p_Y, q_Y are the PMFs on Y at the output of the channel for respective input distributions p_X, q_X, i.e., $p_Y(y) = \sum_{x\in\mathcal{X}} p_X(x)p_{Y|X}(y|x)$ and $q_Y(y) = \sum_{x\in\mathcal{X}} q_X(x)p_{Y|X}(y|x)$.

You may use results from earlier parts without proving them again in later parts. No part of this question requires having solved an earlier part.

(a) [3 points] For a channel $p_{Y|X}$ define

$$\eta(p_{Y|X}) \triangleq \sup_{p_X, q_X, p_X \neq q_X} \frac{D(p_Y||q_Y)}{D(p_X||q_X)}. \tag{3}$$

Argue why $\eta(p_{Y|X}) \leq 1$.

Solution: By the data processing inequality (3) we have that

$$D(p(y)||q(y)) \leq D(p(x)||q(x)),$$

so dividing both sides by $D(p(x)||q(x))$ shows that for any $p(x), q(x)$ with $p(x) \neq q(x)$ we have that

$$\frac{D(p(y)||q(y)}{D(p(x)||q(x))} \leq 1.$$

Taking a supremum over all $p(x), q(x), p(x) \neq q(x)$ proves the desired inequality.

(b) [6 points] Consider some random variables U, X, Y on alphabets $\mathcal{U}, \mathcal{X}, \mathcal{Y}$ such that we have $U \leftrightarrow X \leftrightarrow Y$ is a Markov chain (e.g. $p_{Y|X,U}(y|x,u) = p_{Y|X}(y|x)$ for any u, x, y). Show that

$$\frac{I(U;Y)}{I(U;X)} \leq \eta(p_{Y|X}). \tag{4}$$

where we use the **same** $\eta(p_{Y|X})$ as defined in (3).

Hint: Construct a channel from the joint random variable (U, X) to (U, Y) using $p_{Y|X}$. Choose the two distributions on $\mathcal{U} \times \mathcal{X}$ in order to match the form of (4)*.*

Solution: Consider the channel from (U, X) to (U, Y) that copies the first component and applies the channel $p(y|x)$ to the second component. Then we apply (3) to the distributions $p(u, x)$ and $q(u, x) := p(u)p(x)$. By Markovity, we can rewrite

$$q(u, y) = \sum_x q(u, x, y) = \sum_x p(u)p(x)p(y|x) = p(u)p(y).$$

We then have that

$$\begin{aligned}\frac{D(p(u,y)||q(u,y))}{D(p(u,x)||q(u,x))} &= \frac{D(p(u,y)||p(u)p(y)}{D(p(u,x)||p(u)p(x))} \\ &= \frac{I(U;Y)}{I(U;X)}.\end{aligned}$$

On the other hand,

$$\begin{aligned}D(p(u,y)||q(u,y)) &= \mathbb{E}_u[D(p(y|u)||q(y|u)] \\ &\leq \mathbb{E}_u[\eta(p_{Y|X})D(p(x|u)||q(x|u)] \\ &= \eta(p_{Y|X})D(p(u,x)||q(u,y)).\end{aligned}$$

For the rest of the problem we specialize to the case of the binary symmetric channel with crossover probability parameter δ (denoted by BSC_δ), where $\delta \le \frac{1}{2}$.

(c) 6 points (bonus) Note that $p(x), q(x)$ on binary alphabets can be written as $p := \mathrm{Ber}(a), q := \mathrm{Ber}(b)$ for some $0 \le a, b \le 1$. For the sake of convenience, define the function $g(x) := \delta + (1-2\delta)x$. Fix b, and define the function

$$f(a) := D(\mathrm{Ber}(g(a))||\mathrm{Ber}(g(b))) - (1-2\delta)^2 D(\mathrm{Ber}(a)||\mathrm{Ber}(b)).$$

Use f and its derivatives to show that

$$\boxed{\eta(\mathrm{BSC}_\delta) \le (1-2\delta)^2}. \tag{5}$$

You may use the following identities without proof:

$$\frac{\partial}{\partial a} D(\mathrm{Ber}(a)||\mathrm{Ber}(b)) = \log_2(a/b) - \log_2((1-a)/(1-b)),$$
$$\frac{\partial^2}{\partial a^2} D(\mathrm{Ber}(a)||\mathrm{Ber}(b)) = \frac{\log_2(e)}{a(1-a)}.$$

Solution: We first notice that because the KL divergence between a distribution and itself is 0, we have $f(b) = 0$. Then

$$\frac{\partial}{\partial a} D(\mathrm{Ber}(g(a))||\mathrm{Ber}(g(b))) = (1-2\delta)\log_2(g(a)/g(b)) - (1-2\delta)\log_2((1-g(a))/(1-g(b))),$$

which evaluates to 0 at $a = b$. The derivative of the second term is also 0 at $a = b$ via a similar calculation.

Now it suffices to show that $f''(a) \le 0$. By chain rule, the second derivative of the first term in f is

$$\frac{\partial^2}{\partial a^2} D(\mathrm{Ber}(g(a))||\mathrm{Ber}(g(b))) = \frac{\log_2(e)(1-2\delta)^2}{g(a)(1-g(a))},$$

and so

$$f''(a) = \log_2(e)\left(\frac{(1-2\delta)^2}{g(a)(1-g(a))} - \frac{(1-2\delta)^2}{a(1-a)}\right).$$

Noting that $g(a)$ must be closer to $\frac{1}{2}$ than a is proves $f''(a) \le 0$. Therefore, we can conclude that $a = b$ maximizes f, and f has value 0 at $a = b$, so $f \le 0$.

Finally, we note that $f \le 0$ implies $D(\mathrm{Ber}(g(a))||\mathrm{Ber}(g(b))) \le (1-2\delta)^2 D(\mathrm{Ber}(a)||\mathrm{Ber}(b))$, which we can rearrange to derive the conclusion.

(d) [6 points] Let X be a random variable at the input of this channel, and let Y be the result of applying n BSC_δ channels successively. Using (5), show that

$$I(X;Y) \leq H(X)(1-2\delta)^{2n}.$$

Interpret this result—how does the ability to communicate scale with the number of channels we compose?
Hint: Recall that $H(X) = I(X;X)$.

Solution: We note that

$$X \leftrightarrow X_{k-1} \leftrightarrow X_k$$

is a Markov chain for any $k \geq 1$, where X_k is the random variable obtained after k applications of BSC_δ. (More generally, we have that $X \leftrightarrow X_1 \leftrightarrow \cdots \leftrightarrow X_n$.) Then we can write by combining parts (b) and (c) that

$$I(X;Y) = I(X;X_n) \leq \eta(\text{BSC}_\delta) I(X;X_{n-1}) \leq \cdots \leq \eta(\text{BSC}_\delta)^n I(X;X) \leq (1-2\delta)^{2n} H(X).$$

We can then interpret this result as saying that the channel capacity of n composed BSC_δ channels decays exponentially with n, as denoising becomes increasingly difficult as the crossover probability nears $\frac{1}{2}$.
Remark: We note that another way to see that the capacity decays exponentially is to explicitly compute the crossover probability of n BSC_δ channels chained together. Applying $x \mapsto (1-2x)^2$ results in the same ratio. However, this approach generalizes to other channels for which composition is not so easy to deal with.

Additional space: Clearly state the problem(s) for which you are using this space.

Additional space: Clearly state the problem(s) for which you are using this space.

Additional space: Clearly state the problem(s) for which you are using this space.

Additional space: Clearly state the problem(s) for which you are using this space.