Exam

Time: 10:20 – 13:10

12/05, 2024 (in class)

Name: ______________ Student ID: ______________

Policy: (Read before You Start to Work)

- The exam is **closed book**. However, you are allowed to bring **four A4-size cheat sheets (single-sheet, two-sided)**.
- If you access to any other materials such as books, computing devices, internet connected devices, etc., it is regarded as cheating, and the exam will not be graded. Moreover, we will file the case to the University Office.
- No discussion is allowed during the exam. Everyone has to work on his/her own.
- Please turn in this copy (exam sheets) when you submit your solution sheets.
- Please follow the seat assignment when you are seated.
- Only those written on the solution sheets will be graded. Those written on the exam sheets will not be graded.
- You can use Mandarin or English to write your solutions.

Note: (Read before You Start to Work)

- Part of the points will be given even if you cannot solve the problem completely. Write down your derivation and partial solutions in a clear and systematic way.
- You can make any additional reasonable assumptions that you think are necessary in answering the questions. Write down your assumptions clearly.
- You should express your answers as explicit and analytic as possible.
- You can reuse any known results from our lectures (**restricted to materials from the lecture slides L0–L5**) and homework problems (**HW1–HW4**) without reproving them. Other than those, you need to provide rigorous arguments, unless the problem mentions specifically.

Total Points: 100. Good luck!

1. (Extremal information measures) [24]

a) Let X be a continuous random variable with a probability density function. Moreover, $X \in [a,b]$ with probability 1. Find the maximum value of $\mathrm{h}(X)$ and a maximizing distribution. [12]

b) Let $\mathcal{P}(\mathbb{Z}^+)$ denote the collection of all probability distributions over $\mathbb{Z}^+$, the set of non-negative integers, and $\mathrm{Pois}(\lambda) \in \mathcal{P}(\mathbb{Z}^+)$ be the Poisson distribution with parameter $\lambda > 0$:

$$X \sim \mathrm{Pois}(\lambda) \iff \mathsf{Pr}\{X = n\} = \frac{\lambda^n}{n!}\mathrm{e}^{-\lambda},\ n \in \mathbb{Z}^+ = \{0,1,2,\ldots\}.$$

For a given distribution $\mathsf{P} \in \mathcal{P}(\mathbb{Z}^+)$ with $\mathsf{E}_{\mathsf{P}}[X] = \mu < \infty$, find the value of λ that minimizes $\mathrm{D}(\mathsf{P}\|\mathrm{Pois}(\lambda))$. [12]

a) Let f be the density of X, we have $\int_a^b f(x)\mathrm{d}x = 1$.
By definition,

$$\begin{aligned}
\mathrm{h}(X) &= \mathsf{E}\left[-\log f(X)\right] \\
&= \mathsf{E}\left[\log \frac{\frac{1}{b-a}}{f(X)}\right] + \mathsf{E}\left[-\log \frac{1}{b-a}\right] \\
&\le \log \mathsf{E}\left[\frac{\frac{1}{b-a}}{f(X)}\right] + \log(b-a) \qquad \text{(by concavity of log)} \\
&= \log \int_a^b \frac{f(x)}{f(x)} \frac{1}{b-a}\mathrm{d}x + \log(b-a) = \log(b-a)
\end{aligned}$$

And note that the equality can be achieved when $f(x) = \frac{1}{b-a}\mathbb{1}\{x \in [a,b]\}$.

b) By definition,

$$\begin{aligned}
\mathrm{D}(\mathsf{P}\|\mathrm{Pois}(\lambda)) &= \mathsf{E}_{X\sim\mathsf{P}}\left[\log \frac{\mathsf{P}(X)}{\frac{\lambda^X}{X!}e^{-\lambda}}\right] \\
&= \sum_{n=0}^{\infty} \mathsf{P}(n) \log \frac{\mathsf{P}(n)}{\frac{\lambda^n}{n!}e^{-\lambda}} \\
&= \sum_{n=0}^{\infty} \mathsf{P}(n)\left(\log \mathsf{P}(n) + \log n! - n\log\lambda + \lambda\log e\right) \\
&= \mathrm{H}(\mathsf{P}) + \sum_{n=0}^{\infty} \mathsf{P}(n)\log n! - \mu\log\lambda + \lambda\log e
\end{aligned}$$

Thus, to minimize $\mathrm{D}(\mathsf{P}\|\mathrm{Pois}(\lambda))$, it's equivalent to minimize $g(\lambda) := -\mu\log\lambda + \lambda\log e$. Observe that

$$\begin{aligned}
g'(\lambda) &= -\frac{\mu}{\lambda}\log e + \log e \quad \text{and} \\
g''(\lambda) &= \frac{\mu}{\lambda^2}\log e > 0,\ \forall\lambda
\end{aligned}$$

We can conclude that the minimizer $\lambda^* = \mu$.

Grading Policy

a) Definition of $\mathrm{h}(X)$ [2] Optimality [4] Maximum value [3] Achieving distribution [3]

b) Definition of KL divergence [4] Optimality [4] Minimizer [4]

Lack of 2nd derivative check [-0.5]

2. (An optical communication system) [24]

In an optical communication system, the receiver can count the number of received photons during a time interval. The arrival of photons at the receiver is modeled as a homogeneous Poisson process. The transmitter sends information by changing the rate of the Poisson process.

Recall the following important facts about a Poisson process with rate $\lambda > 0$:

- The number of photons that arrive during an interval $\mathcal{I}$ of length $t > 0$, denoted as $N_{\mathcal{I}}$, is a Poisson random variable with parameter λt. See Problem 1b) for the probability mass function of a Poisson random variable.
- The numbers of photons that arrive during *disjoint* intervals $\mathcal{I}_1, \mathcal{I}_2, \ldots$ are independent Poisson random variables.
- The inter-arrival times are i.i.d. exponential random variables with parameter λ.

a) (Warm-up) Let $X \sim \text{Pois}(\alpha)$ be a Poisson random variable with parameter λ. Recall that the moment generating function (MGF) of X

$$\mathsf{E}_{X\sim\text{Pois}(\alpha)}\left[\mathrm{e}^{uX}\right] = \mathrm{e}^{\alpha(\mathrm{e}^u - 1)}.$$

Consider n mutually independent random variables $X_1, X_2, \ldots, X_m$ with $X_i \sim \text{Pois}(\alpha_i), i = 1, 2, \ldots, m$. Use the MGF approach to show that $\sum_{i=1}^m X_i$ remains a Poisson random variable, and its parameter is $\sum_{i=1}^m \alpha_i$. [4]

Now, suppose the transmitter wants to send one bit to the receiver. For the receiver, to determine the rate, it is faced with the following binary detection (binary hypothesis testing) problem:

$$\begin{aligned} \mathcal{H}_0 :& \quad \lambda = \lambda_0 \\ \mathcal{H}_1 :& \quad \lambda = \lambda_1 \end{aligned}$$

where $0 < \lambda_0 < \lambda_1 < \infty$. λ_0 and λ_1 are the rates of the Poisson process corresponding to bit 0 and 1 respectively, and they are both known to the decision maker. In the following, consider two strategies of decision making.

b) Suppose the decision maker waits up until n photons arrive and makes the decision based on the sequence if inter-arrival times. Let $\varpi^*_{0|1}(n,\epsilon)$ denote the minimum type-II error probability subject to the constraint that the type-I error probability is not greater than ϵ, $0 < \epsilon < 1$. Does $\lim_{n\to\infty} \frac{1}{n} \log \frac{1}{\varpi^*_{0|1}(n,\epsilon)}$ exist? If so, find it. Otherwise, show that the limit does not exist. [10]

c) Now suppose it makes the decision based on the total number of received photons up to time τ. Let $\varpi^*_{0|1}(\tau,\epsilon)$ denote the minimum type-II error probability subject to the constraint that the type-I error probability is not greater than ϵ, $0 < \epsilon < 1$. Does $\lim_{\tau\to\infty} \frac{1}{\tau} \log \frac{1}{\varpi^*_{0|1}(\tau,\epsilon)}$ exist? If so, find it. Otherwise, show that the limit does not exist. (You may assume τ is an integer for simplicity, while the answer will be the same even if τ is not an integer.) [10]

a) The MGF of $\sum_{i=1}^m X_i$ is $\phi_{\sum_{i=1}^m X_i}(u) = \prod_{i=1}^m \phi_{X_i}(u) = \prod_{i=1}^m e^{\alpha_i(e^u-1)} = e^{\left(\sum_{i=1}^m \alpha_i\right)(e^u-1)}$, where the first equality is by independence.
Since MGF determines distribution, we have $\sum_{i=1}^m X_i \sim \mathrm{Pois}(\sum_{i=1}^m \alpha_i)$.

b) By independence of inter-arrival times of photons, what the decision maker observes are just i.i.d. exponential random variables. So, it is equivalent to test

$$
\begin{aligned}
\mathcal{H}_0 &: \ T_1, T_2, ..., T_n \sim \mathrm{Exp}^{\otimes n}(\lambda_0) \\
\mathcal{H}_1 &: \ T_1, T_2, ..., T_n \sim \mathrm{Exp}^{\otimes n}(\lambda_1)
\end{aligned}
$$

Therefore, Chernoff-Stein lemma can be applied and hence,

$$
\begin{aligned}
\lim_{n\to\infty} -\frac{1}{n}\log \varpi^*_{0|1}(n,\epsilon) &= \mathrm{D}\left(\mathrm{Exp}(\lambda_0)\|\mathrm{Exp}(\lambda_1)\right) \\
&= \mathsf{E}_{T\sim\mathrm{Exp}(\lambda_0)}\left[\log\frac{\lambda_0 e^{-\lambda_0 T}}{\lambda_1 e^{-\lambda_1 T}}\right] \\
&= \int_0^\infty \lambda_0 e^{-\lambda_0 t}\left(\log\frac{\lambda_0}{\lambda_1} + \log e\cdot(\lambda_1-\lambda_0)t\lambda_0 e^{-\lambda_0 t}\right)\mathrm{d}t \\
&= \log\frac{\lambda_0}{\lambda_1} + \log e\cdot\frac{\lambda_1-\lambda_0}{\lambda_0}
\end{aligned}
$$

c) For simplicity, let τ to be an integer.
Different from (a), the decision maker in this question may interpret the observation differently by uniformly slicing each unit time into $m \in \mathbb{N}$ intervals, each with length $\frac{1}{m}$. And since all $m\tau$ intervals are disjoint and by the property of Poisson process, the numbers of photons in the intervals, denoted by $\{N_i : i = 1, ..., m\tau\}$ are i.i.d. Poisson distributed with parameter $\frac{\lambda_0}{m}$ or $\frac{\lambda_1}{m}$. Thus, it is equivalent to test

$$
\begin{aligned}
\mathcal{H}_0 &: \ N_1, ..., N_{m\tau} \sim \mathrm{Pois}^{\otimes m\tau}(\frac{\lambda_0}{m}) \\
\mathcal{H}_1 &: \ N_1, ..., N_{m\tau} \sim \mathrm{Pois}^{\otimes m\tau}(\frac{\lambda_1}{m})
\end{aligned}
$$

Hence, Chernoff-Stein lemma can be applied again and we have

$$
\begin{aligned}
\lim_{\tau\to\infty} -\frac{1}{m\tau}\log \varpi^*_{0|1,m-\mathrm{sliced}}(m\tau,\epsilon) &= \mathrm{D}\left(\mathrm{Pois}(\frac{\lambda_0}{m})\middle\|\mathrm{Pois}(\frac{\lambda_1}{m})\right) \\
&= \sum_{k=0}^\infty \frac{\left(\frac{\lambda_0}{m}\right)^k}{(k)!}e^{-\frac{\lambda_0}{m}}\log\frac{\frac{\left(\frac{\lambda_0}{m}\right)^k}{(k)!}e^{-\frac{\lambda_0}{m}}}{\frac{\left(\frac{\lambda_1}{m}\right)^k}{(k)!}e^{-\frac{\lambda_1}{m}}} \\
&= \sum_{k=0}^\infty \frac{\left(\frac{\lambda_0}{m}\right)^k}{(k)!}e^{-\frac{\lambda_0}{m}}\left(k\log\frac{\lambda_0}{\lambda_1} + \frac{\log e}{m}(\lambda_1-\lambda_0)\right) \\
&= \frac{\lambda_0}{m}\log\frac{\lambda_0}{\lambda_1} + \frac{\log e}{m}(\lambda_1-\lambda_0)
\end{aligned}
$$

Therefore,

$$\lim_{\tau\to\infty} -\frac{1}{\tau}\log \varpi^*_{0|1,m-\text{sliced}}(m\tau,\epsilon) = \lambda_0 \log\frac{\lambda_0}{\lambda_1} + \log e\cdot(\lambda_1-\lambda_0)\,, \quad \text{for all } m$$

Finally, we have $\varpi^*_{0|1}(\tau,\epsilon) = \inf_m \varpi^*_{0|1,m-\text{sliced}}(m\tau,\epsilon)$. Thus, we have

$$\lim_{\tau\to\infty} -\frac{1}{\tau}\varpi^*_{0|1}(\tau,\epsilon) = \lambda_0 \log\frac{\lambda_0}{\lambda_1} + \log e\cdot(\lambda_1-\lambda_0)$$

Grading Policy

a) MGF calculation [3.5] State that MGF determines distribution [0.5]

b) State that the observations are i.i.d. samples [2] Apply Chernoff-Stein lemma [5] Error exponent calculation [3]

c) State that the observations are i.i.d. samples [1] Cover the case for all kinds of slicing [2] Apply Chernoff-Stein lemma [5] Error exponent calculation [2]

3. (Lossless source coding) [26]

a) Consider three binary stationary and ergodic random sources $\{X_i \,|\, i \in \mathbb{N}\}$, $\{Y_i \,|\, i \in \mathbb{N}\}$, and $\{Z_i \,|\, i \in \mathbb{N}\}$, satisfying

$$X_{i+1} = X_i \oplus Z_i,\ Y_{i+2} = Y_{i+1} \oplus Y_i \oplus Z_i,\ i = 1, 2, \ldots$$

Here $\oplus$ is the XOR operator: $a \oplus b = \mathbb{1}\{a \neq b\}$.

Show that the three sources have the same entropy rate. [12]

b) In the fixed-to-fixed lossless source coding problem, there is a slight probability that the source sequence cannot be correctly reconstructed. One possible way to improve the situation so that the source sequence is *always* reconstructed correctly is to allow *variable-length* codewords. For a given variable-length encoder enc, let $l_{\text{enc}}(s^n)$ denote the length of the codeword (the bit sequence representing s^n). Note that to achieve *exact* reconstruction for all source sequences $s^n \in \mathcal{S}^n$, different s^n's should be mapped to different codewords. To see the effect of data compression, one considers the *expected codeword length*

$$\mathsf{E}_{S^n}[l_{\text{enc}}(S^n)] = \sum_{s^n \in \mathcal{S}^n} \mathsf{Pr}\{S^n = s^n\} l_{\text{enc}}(s^n),$$

and let us denote the minimum of the expected codeword length over all possible variable-length encoders as

$$k^*_{\mathsf{variable}}(n) := \min_{\text{enc}} \mathsf{E}_{S^n}[l_{\text{enc}}(S^n)]\,.$$

The compression rate can then be defined as

$$\mathrm{R}^*_{\mathsf{variable}} = \lim_{n\to\infty} \frac{1}{n} k^*_{\mathsf{variable}}(n)$$

if the limit exists. In the following you may assume the limit exists.

For a discrete memoryless source $\{S_i\}$, use the fixed-to-fixed lossless source coding theorem in the lecture to directly show that

$$\mathrm{R}^*_{\mathsf{variable}} \le \epsilon + \mathrm{H}(S) \quad \forall \epsilon \in (0,1).$$

Hence, $\mathrm{R}^*_{\mathsf{variable}} \le \mathrm{H}(S)$. [14]

a)

$$\begin{aligned}
\mathcal{H}(\{X_i\}) := \lim_{n\to\infty} \frac{1}{n}\mathrm{H}(X^n) &= \lim_{n\to\infty} \frac{1}{n}\mathrm{H}(X_1, Z_1, ..., Z_{n-1}) \quad \text{(by the given relation)}\\
&= \lim_{n\to\infty} \frac{1}{n}\left(\mathrm{H}\left(X_1 \middle| Z^{n-1}\right) + \mathrm{H}(Z_1, ..., Z_{n-1})\right) \quad \text{(chain rule)}\\
&= 0 + \mathcal{H}(\{Z_i\}) \quad (\mathrm{H}\left(X_1 \middle| Z^{n-1}\right) \text{ is bounded by 1})
\end{aligned}$$

$$\begin{aligned}
\mathcal{H}(\{Y_i\}) := \lim_{n\to\infty} \frac{1}{n}\mathrm{H}(Y^n) &= \lim_{n\to\infty} \frac{1}{n}\mathrm{H}(Y_1, Y_2, Z_1, ..., Z_{n-1}) \quad \text{(by the given relation)}\\
&= \lim_{n\to\infty} \frac{1}{n}\left(\mathrm{H}\left(Y_1, Y_2 \middle| Z^{n-1}\right) + \mathrm{H}(Z_1, ..., Z_{n-1})\right) \quad \text{(chain rule)}
\end{aligned}$$

$$= 0 + \mathcal{H}(\{Z_i\}) \quad (\mathrm{H}\left(Y_1, Y_2 \middle| Z^{n-1}\right) \text{ is bounded by 2})$$

b) In this problem, we want to show the achievability.
For any given $\epsilon > 0$, by the fixed-to-fixed lossless source coding theorem, there exists a lossless $\left(n, k, \frac{\epsilon}{\log|\mathcal{S}|}\right)$ source code. Let $\mathcal{B}^{(n)}$ denote the range of decoding function.

For $S^n \in \mathcal{B}^{(n)}$, we use k bits to uniquely encode them.
For $S^n \in \mathcal{S}^n \setminus \mathcal{B}^{(n)}$, we may use $\lceil n \log|\mathcal{S}| \rceil$ to uniquely encode them.

For this variable-length source code, we now analyze its expected codeword length

$$\begin{aligned}\mathsf{E}_{S^n}\left[l_{\text{enc}}(S^n)\right] &= \mathsf{Pr}\left\{S^n \in \mathcal{B}^{(n)}\right\} \cdot k + \mathsf{Pr}\left\{S^n \in \mathcal{S}^n \setminus \mathcal{B}^{(n)}\right\} \cdot n \log|\mathcal{S}| \\ &\leq k + \epsilon n\end{aligned}$$

Thus, by definition of R^*_{Var}, we have

$$R^*_{\text{Var}}(n) \leq \frac{k}{n} + \epsilon$$

Finally, $\lim\limits_{n\to\infty} R^*_{\text{Var}}(n) \leq \lim\limits_{n\to\infty} \frac{k}{n} + \epsilon \leq \mathrm{H}(S) + \epsilon$

Granding Policy

a) Definition of entropy rate [4] Relate the information measures of $\{X_i\}$, $\{Y_i\}$, $\{Z_i\}$ [4]
Finalize the proof [4]

b) Invoke the fixed-to-fixed lossless source coding theorem [4]
Propose a variable length scheme that can guarantee exact recovery [4]
Finalize the proof [6]

4. (Coding with erasures) [26]

a) (Warm up) Let (A, B) be a pair of jointly distributed discrete random variables in $\mathcal{A} \times \mathcal{B}$ with joint PMF $\mathsf{P}_{A,B}$. Consider another pair of jointly distributed random variables $(G, H) \in (\mathcal{A} \cup \{*\}) \times \mathcal{B}$ where $* \notin \mathcal{A}$ is a special symbol that deserves further attention, and (G, H) follows the joint PMF

$$\mathsf{P}_{G,H}(g, h) = \begin{cases} p\mathsf{P}_B(h), & \text{if } g = * \\ (1-p)\mathsf{P}_{A,B}(g, h), & \text{if } g \in \mathcal{A} \end{cases}$$

for some $p \in (0, 1)$. Let $E := \mathbb{1}\{G = *\}$. Show that E and H are independent. [4]

b) Following Part a), show that $\mathrm{I}(G; H) = (1-p)\mathrm{I}(A; B)$. [6]

c) A discrete memoryless channel with input $X \in \mathcal{X}$, output $Y \in \mathcal{Y}$, and channel law $\mathsf{P}_{Y|X}$ has capacity C bits per channel use. Consider another discrete memoryless channel with input $U \in \mathcal{X}$, output $V \in \mathcal{Y} \cup \{*\}$, and channel law

$$\mathsf{P}_{V|U}(v|u) = \begin{cases} p, & \text{if } v = * \\ (1-p)\mathsf{P}_{Y|X}(v|u), & \text{if } v \in \mathcal{Y} \end{cases}.$$

Note that $* \notin \mathcal{Y}$ is a special symbol indicating that the channel output is erased.

Find the capacity of this new channel. Express your answer in terms of p and C. [6]

d) The rate distortion function for a discrete memoryless source $S \in \mathcal{S}$, $S \sim \mathsf{P}_S$ and distortion measure $d : \mathcal{S} \times \mathcal{S} \to [0, \infty)$ is $\mathrm{R}(\mathsf{D})$. Consider another discrete memoryless source $X \in \mathcal{S} \cup \{*\}$ with

$$\mathsf{P}_X(x) = \begin{cases} p, & \text{if } x = * \\ (1-p)\mathsf{P}_S(x), & \text{if } x \in \mathcal{S} \end{cases}.$$

Note that $* \notin \mathcal{S}$ is a special "don't care" symbol for the following distortion measure $\tilde{d} : \mathcal{S} \cup \{*\} \times \mathcal{S} \to [0, \infty)$ and

$$\tilde{d}(x, \hat{x}) = \begin{cases} 0, & \text{if } x = * \\ d(x, \hat{x}), & \text{if } x \in \mathcal{S} \end{cases}.$$

Find the rate distortion function for this new source with the above distortion measure. Express your answer in terms of p, D, and the rate distortion function $\mathrm{R}(\cdot)$. [10]

a) Let's show the independence by calculating the joint PMF of E and H.

$$\mathsf{P}_{E,H}(1, h) = \mathsf{Pr}\{G = *, H = h\} = \mathsf{P}_{G,H}(*, h) = p\mathsf{P}_B(h)$$
$$\mathsf{P}_{E,H}(0, h) = \mathsf{Pr}\{G \neq *, H = h\} = \sum_{g \in \mathcal{A}} \mathsf{P}_{G,H}(g, h) = (1-p)\mathsf{P}_B(h)$$
$$\mathsf{P}_H(h) = \sum_{g \in \mathcal{A}} \mathsf{P}_{G,H}(g, h) + \mathsf{P}_{G,H}(*, h) = (1-p)\mathsf{P}_B(h) + p\mathsf{P}_B(h) = \mathsf{P}_B(h)$$
$$\mathsf{P}_E(e) = \begin{cases} p & \text{if } e = 1 \\ 1-p & \text{if } e = 0 \end{cases}$$

So E and H are independent.

b) By chain rule and independence of E and H, we can get

$$\mathrm{I}(G;H) = \mathrm{I}(G,E;H) = \mathrm{I}(G;H|E) + \mathrm{I}(E;H) = \mathrm{I}(G;H|E)$$

By further decomposing,

$$\begin{aligned}\mathrm{I}(G;H|E) &= \mathsf{Pr}\{E=1\}\,\mathrm{I}(G;H|E=1) + \mathsf{Pr}\{E=0\}\,\mathrm{I}(G;H|E=0)\\ &= \mathsf{Pr}\{G=*\}\,\mathrm{I}(G;H|G=*) + \mathsf{Pr}\{G\neq *\}\,\mathrm{I}(G;H|G\neq *)\\ &= (1-p)\mathrm{I}(G;H|G\neq *) = (1-p)\mathrm{I}(A;B)\end{aligned}$$

c) Take $\mathsf{P}_U(u) = \mathsf{P}_X(u)$. Then we have $\mathsf{P}_{U,V}(u,v) = \begin{cases} p\cdot \mathsf{P}_X(u) & \text{if } v = * \\ (1-p)\cdot \mathsf{P}_{X,Y}(u,v) & \text{if } v\in\mathcal{Y}\end{cases}$.
By a) and b), we have $\mathrm{I}(U;V) = (1-p)\mathrm{I}(X;Y)$.
Now, we compute the capacity directly by

$$\max_{\mathsf{P}_U\in\mathcal{P}(\mathcal{X})} \mathrm{I}(U;V) = \max_{\mathsf{P}_X\in\mathcal{P}(\mathcal{X})} (1-p)\mathrm{I}(X;Y) = (1-p)C$$

d) First, we lower bound $\mathrm{I}\left(X;\hat{X}\right)$ by

$$\mathrm{I}\left(X;\hat{X}\right) = \mathrm{I}\left(X,E;\hat{X}\right) = \mathrm{I}\left(X;\hat{X}\middle|E\right) + \mathrm{I}\left(E;\hat{X}\right) \overset{(1)}{\geq} \mathrm{I}\left(X;\hat{X}\middle|E\right),$$

where $E := \mathbb{1}\{X=*\}$.
Further,

$$\mathrm{I}\left(X;\hat{X}\middle|E\right) = p\cdot 0 + (1-p)\mathrm{I}\left(X;\hat{X}\middle|X\neq *\right) = (1-p)\mathrm{I}\left(X;\hat{X}\middle|X\neq *\right)$$

On the other hand, we also need to deal with the distortion.
Note that

$$\mathsf{E}\left[\hat{d}(X,\hat{X})\right] = (1-p)\mathsf{E}\left[d(X,\hat{X})\middle|X\neq *\right].$$

Hence, the distortion constraint

$$\mathsf{E}\left[\hat{d}(X,\hat{X})\right] \leq D \Longleftrightarrow \mathsf{E}\left[d(X,\hat{X})\middle|X\neq *\right] \leq \frac{D}{1-p}$$

Also note that $\mathsf{P}_{X|X\neq *}(x) = \frac{(1-p\mathsf{P}_S(x))}{1-p} = \mathsf{P}_S(x)$. Therefore,

$$\begin{aligned}\mathrm{I}\left(X;\hat{X}\middle|X\neq *\right) &\overset{(2)}{\geq} \min_{\mathsf{P}_{X,\hat{X}|X\neq *}:\mathsf{P}_{X|X\neq *}\sim\mathsf{P}_S,\mathsf{E}\left[d(X,\hat{X})|X\neq *\right]\leq\frac{D}{1-p}} \mathrm{I}\left(X;\hat{X}\middle|X\neq *\right)\\ &= \min_{\mathsf{P}_{\hat{S}|S}:\mathsf{E}\left[d(S,\hat{S})\right]\leq\frac{D}{1-p}} \mathrm{I}\left(S;\hat{S}\right) = R(\frac{D}{1-p})\end{aligned}$$

To this place, we derive that $(1-p)R(\frac{D}{1-p})$ is a lower bound of the rate distortion function of the modified source. The remaining is to argue that the bound is indeed achievable, which can be traced by holding the equalities in each $\geq$.

The equality in (1) holds iff $\hat{X}$ and $\mathbb{1}\{X = *\}$ are independent.
The equality in (2) holds iff $\mathsf{P}_{X,\hat{X}|X\neq *} = \mathsf{P}^*_{S,\hat{S}}$.

Note that (2) can be achieved by $\mathsf{P}_{\hat{X}|X}(\hat{x}|x) = \mathsf{P}^*_{S}(\hat{x}|x)$ for all $x \neq *$.
And (1), by a), can also be achieved by

$$\mathsf{P}_{\hat{X}|X}(\hat{x}|*) = \mathsf{P}^*_{\hat{S}}(\hat{x}) \Longrightarrow \mathsf{P}_{X,\hat{X}}(x,\hat{x}) = \begin{cases} p\cdot \mathsf{P}^*_{\hat{S}}(\hat{x}) & \text{if } x = * \\ (1-p)\mathsf{P}^*_{S,\hat{S}}(x,\hat{x}) & \text{if } x \neq * \end{cases},$$

where $\mathsf{P}^*_{\hat{S}}$ is the marginal of $\mathsf{P}_S\mathsf{P}_{\hat{S}|S}$.

Grading Policy

a) Joint PMF [2] Equality to product distibution [2]

b) Arrive $\mathrm{I}(G;H) = \mathrm{I}(G;H|E)$ by chain rule [2] and independence of G, H [2] Finalizing the proof [2]

c) Recognizing (U, V) meet condition in a) and call b) [4] Finalizing the proof [2]

d) Lower bound $\mathrm{I}\left(X;\hat{X}\right)$ by $\mathrm{I}\left(X;\hat{X}\middle|E\right)$ [2] and by $(1-p)\mathrm{I}\left(X;\hat{X}\middle|X\neq *\right)$ [2]
Show the equivalence of $\min \mathrm{I}\left(X;\hat{X}\middle|X\neq *\right)$ and $\min \mathrm{I}\left(S;\hat{S}\right)$ under the two respective distortion constraints [4]
Show the lower bound is achievable [2]