EE376A - Information Theory
Final, Monday March 16th Solutions

Instructions:

- You have **three hours**, 3.30PM - 6.30PM
- The exam has 4 questions, totaling 120 points.
- Please start answering each question on a new page of the answer booklet.
- You are allowed to carry the textbook, your own notes and other course related material with you. Electronic reading devices [including kindles, laptops, ipads, etc.] are allowed, provided they are used solely for reading pdf files already stored on them and not for any other form of communication or information retrieval.
- You are required to provide detailed explanations of how you arrived at your answers.
- You can use previous parts of a problem even if you did not solve them.
- As throughout the course, entropy (H) and Mutual Information (I) are specified in bits.
- log is taken in base 2.
- Throughout the exam 'prefix code' refers to a variable length code satisfying the prefix condition.
- Good Luck!

1. **Three Shannon Codes** *(25 points)*
   Let $\{U_i\}_{i\geq 1}$ be a stationary finite-alphabet source whose alphabet size is r. Note that the stationarity property implies that $P(u_i), P(u_i|u_{i-1})$ do not depend on i. Throughout this problem, assume that $-\log P(u_i)$ and $-\log P(u_i|u_{i-1})$ are integers for all (u_i, u_{i-1}). Recall the definition of a Shannon Code given in the lecture. Your TA's decided to compress this source in a lossless fashion using Shannon coding. However, each of them had a different idea:

   - Idoia suggested to code symbol-by-symbol, i.e., concatenate Shannon codes on the respective source symbols $U_1, U_2, \ldots$.
   - Kartik suggested to code in pairs. In other words, first code (U_1, U_2) with a Shannon code designed for the pair, then code (U_3, U_4), and so on.
   - Jiantao suggested to code each symbol given the previous symbol by using the Shannon code for the conditional pmf $\{P(u_i|u_{i-1})\}$. In other words, first code U_1, then code U_2 given U_1, then code U_3 given U_2, and so on.

   In this problem, you will investigate which amongst the three schemes is best for a general stationary source.

   (a) (10 points) If the source is memoryless (i.e. i.i.d.), compare the expected codeword length per symbol, i.e., $\frac{1}{n}E[l(U^n)]$, of each scheme, assuming $n > 2$ is even.

   (b) (15 points) Compare the schemes again, for the case where the source is no longer memoryless and, in particular, is such that U_{i-1} and U_i are not independent.

Solution:

We will first analyze each of the coding schemes for general stationary sources.
Idoia's scheme: Use codeword length $-\log P(u)$ for a symbol u.

$$\begin{aligned}
\bar{l_1} &= \frac{1}{n}\mathbb{E}[l_1(U^n)] \\
&= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} -\log P(U_i)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[-\log P(U_1)\right] \qquad \text{(stationarity, and linearity of expectation)} \\
&= H(U_1) \qquad \text{(definition of entropy)}
\end{aligned}$$

Kartik's scheme: Use codeword length $-\log P(u_i, u_{i+1})$ for each successive pair of symbols (u_i, u_{i+1}).

$$\bar{l_2} = \frac{1}{n}\mathbb{E}[l_2(U^n)]$$

$$= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n/2} - \log P(U_{2i-1}, U_{2i})\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n/2}\mathbb{E}\left[- \log P(U_1, U_2)\right] \qquad \text{(stationarity, and linearity of expectation)}$$

$$= \frac{1}{2}H(U_1, U_2) \qquad \text{(definition of entropy)}$$

Jiantao's scheme: Use codeword length $-\log P(u_1)$ for first symbol u_1, and then codeword length $-\log P(u_i|u_{i-1})$ for successive symbols.

$$\bar{l_3} = \frac{1}{n}\mathbb{E}[l_3(U^n)]$$

$$= \frac{1}{n}\mathbb{E}\left[- \log P(U_1) + \sum_{i=2}^{n} - \log P(U_i|U_{i-1})\right]$$

$$= \frac{1}{n}\mathbb{E}\left[- \log P(U_1) + \sum_{i=2}^{n} - \log P(U_2|U_1)\right] \qquad \text{(stationarity)}$$

$$= \frac{1}{n}[H(U_1) + (n-1)H(U_2|U_1)] \qquad \text{(linearity of expectation, definition of entropy)}$$

(a) Because the source is i.i.d., all three coding schemes have the same average codeword length, equal to the entropy $H(U_1)$. One can verify that when the source is i.i.d. $\bar{l_1} = \bar{l_2} = \bar{l_3} = H(U_1)$.

(b) Idoia's codeword length is longest because $H(U_2|U_1) \leq H(U_2) = H(U_1)$. For $n = 2$, the performance of Kartik's and Jiantao's coding schemes are identical. However for larger n, Jiantao's coding scheme has *smallest* codeword length, since $H(U_1, U_2) = H(U_1) + H(U_2|U_1) \geq 2H(U_2|U_1)$, with equality iff the source is memoryless. Therefore, in general, $\bar{l_1} \geq \bar{l_2} \geq \bar{l_3}$.

2. Channel coding with side information *(35 points)*

Consider the binary channel given by

$$Y_i = X_i \oplus Z_i, \tag{1}$$

where X_i, Y_i, Z_i all take values in $\{0,1\}$, and $\oplus$ denotes addition modulo-2. There are channel states S_i which determine the noise level of Z_i as follows.

- S_i is binary valued, taking values in the set $\{G, B\}$, distributed as

$$S_i = \begin{cases} G, & \text{with probability } \frac{2}{3} \\ B, & \text{with probability } \frac{1}{3} \end{cases}$$

- The conditional distribution of Z_i given S_i is characterized by

$$P(Z_i = 1 | S_i = s) = \begin{cases} \frac{1}{4}, & \text{if } s = G \\ \frac{1}{3}, & \text{if } s = B \end{cases}$$

  In other words, $Z_i | \{S_i = s\} \sim \mathit{Bernoulli}(p_s)$, where

$$p_s = \begin{cases} \frac{1}{4}, & \text{if } s = G \\ \frac{1}{3}, & \text{if } s = B \end{cases}$$

  $\{(S_i, Z_i)\}$ are i.i.d. (in pairs), independent of the channel input sequence $\{X_i\}$.

(a) (10 points) What is the capacity of this channel when *both* the encoder *and* the decoder have access to the state sequence $\{S_i\}_{i \geq 1}$?

(b) (10 points) What is the capacity of this channel when *neither* the encoder *nor* the decoder have access to the state sequence $\{S_i\}_{i \geq 1}$?

(c) (10 points) What is the capacity of this channel when *only the decoder* knows the state sequence $\{S_i\}_{i \geq 1}$?

(d) (5 points) Which is largest and which is smallest among your answers to parts (a), (b) and (c)? Explain.

Solution:

(a) The capacity of this channel is given by

$$\begin{aligned} C &= \max_{p(X|S)} I(X; Y|S) \\ &= \max_{p(X|S)} H(Y|S) - H(Y|X, S) \end{aligned}$$

$$
\begin{aligned}
&= \max_{p(X|S)} H(X \oplus Z|S) - H(Z|S) \qquad (Y = X \oplus S \oplus Z, \text{ and } Z \text{ is independent of } X) \\
&\leq 1 - H(Z|S) \qquad (\text{binary entropy is upper bounded by } 1) \\
&= 1 - P(S = G)H(Z|S = G) - P(S = B)H(Z|S = B) \\
&= 1 - \frac{2}{3}h_2(1/4) - \frac{1}{3}h_2(1/3),
\end{aligned}
$$

where $h_2(\cdot)$ is the binary entropy function. Note that the above bound is achieved by choosing input $X \sim Bern(0.5)$ regardless of the state S. Choosing this input gives the output a uniform distribution which maximizes the entropy. It is crucial to state the capacity achieving distribution to show that an upper bound on the mutual information can be achieved, which is why it is the capacity. Several students did not do this for the problem, and lost some points.

Alternate solution: Since both encoder and decoder know the state, they can use the corresponding capacity achieving codes for when the channel is "good" or "bad" respectively. The capacity is simply the weighted average of the capacities of the two binary symmetric channels, i.e.

$$
\begin{aligned}
C &= P(S = G)C_G + P(S = B)C_B \\
&= \frac{2}{3}(1 - h_2(1/4)) + \frac{1}{3}(1 - h_2(1/3)) \\
&= 1 - \frac{2}{3}h_2(1/4) - \frac{1}{3}h_2(1/3).
\end{aligned}
$$

(b) When neither encoder nor decoder has any state information, there is no way to use the state information. This results in an average BSC with equivalent crossover probability

$$
\begin{aligned}
p &= \frac{2}{3} \cdot \frac{1}{4} + \frac{1}{3} \cdot \frac{1}{3} \\
&= \frac{5}{18}
\end{aligned}
$$

Thus, the capacity of this channel is simply that of a BSC(p), i.e.

$$
C = 1 - h_2(5/18)
$$

(c) The decoder has access to the state, which means that equivalently the channel output can be viewed as the pair (Y, S). The capacity of the channel in this case is

$$
\begin{aligned}
C &= \max_{p(X)} I(X; Y, S) \\
&= \max_{p(X)} H(Y, S) - H(Y, S|X) \\
&= \max_{p(X)} H(S) + H(Y|S) - H(S|X) - H(Y|S, X) \\
&= \max_{p(X)} H(Y|S) - H(Y|S, X) \qquad (S \text{ and } X \text{ are independent})
\end{aligned}
$$

$$
\begin{aligned}
&= \max_{p(X)} H(Y|S) - H(Z|S) \qquad (Y = X \oplus S \oplus Z, \text{ and } Z \text{ is independent of } X) \\
&\le 1 - H(Z|S) \qquad \text{(binary entropy is upper bounded by 1)} \\
&= 1 - P(S = G)H(Z|S = G) - P(S = B)H(Z|S = B) \\
&= 1 - \frac{2}{3}h_2(1/4) - \frac{1}{3}h_2(1/3),
\end{aligned}
$$

where the upper bound can be achieved by choosing $X \sim Bern(0.5)$. This expression is exactly the same as (a). In other words, knowing the state at the decoder is just as useful as knowing the state at both encoder and decoder! This magic happens because the capacity achieving input for part (a) does not need to know the state of the channel.

(d) In this problem $C_a = C_c > C_b$. Clearly, knowing the state is advantageous since it reduces the uncertainty in the channel noise. I.e. $H(Z|S) < H(Z)$. The capacity of part (a) is largest because we have the entire state information available to both encoder and decoder. In this problem, the additional beauty is that just knowing the state at the decoder is sufficient to achieve the capacity of (a). The reason for this is stated above, and is due to the fact that $X \sim Bern(0.5)$ achieves the capacity in both (a) and (c).

3. Modulo-3 additive noise channel *(25 points)*

(a) (5 points) Consider the modulo-3 additive white noise channel given by

$$Y_i = X_i \oplus Z_i, \tag{2}$$

where X_i, Z_i, Y_i all take values in the alphabet $\{0, 1, 2\}$, $\oplus$ denotes addition modulo-3, and $\{Z_i\}$ are i.i.d. $\sim Z$ and independent of the channel input sequence $\{X_i\}$.

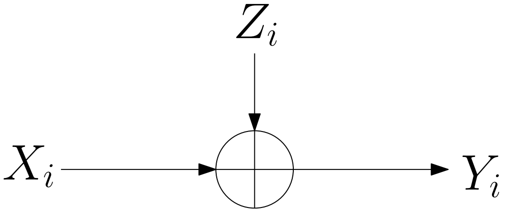

Figure 1: Ternary additive channel.

Show that the capacity of this channel is given by

$$C = \log 3 - H(Z). \tag{3}$$

(b) (7 points) For $\epsilon \geq 0$ define

$$\phi(\epsilon) = \max_{Z:Pr(Z\neq 0)\leq\epsilon} H(Z), \tag{4}$$

where the maximization is over ternary random variables Z that take values in $\{0, 1, 2\}$ (and that satisfy the indicated constraint). Obtain $\phi(\epsilon)$ explicity, as well as the distribution of the random variable, Z_ϵ, that achieves the associated maximum.

[Distinguish between the ranges $0 \leq \epsilon < 2/3$ and $\epsilon \geq 2/3$.]

(c) (5 points) Consider the problem of rate distortion coding of a memoryless source $U_i \sim U$, where the source and the reconstruction alphabets are both equal and ternary, i.e., $\mathcal{U} = \mathcal{V} = \{0, 1, 2\}$. Let the distortion measure be Hamming loss

$$d(u, v) = \begin{cases} 0, & \text{if } u = v \\ 1, & \text{otherwise.} \end{cases}$$

For U, V that are jointly distributed such that $E[d(U, V)] \leq D$, justify the following chain of equalities and inequalities

$$I(U; V) \overset{(i)}{=} H(U) - H(U|V)$$

$$\begin{aligned} &\overset{(ii)}{=} H(U) - H(U \ominus V|V) \\ &\overset{(iii)}{\geq} H(U) - H(U \ominus V) \\ &\overset{(iv)}{\geq} H(U) - \phi(D), \end{aligned}$$

where $\ominus$ denotes subtraction modulo-3 and $\phi(D)$ was defined in Equation (4). Argue why this implies that the rate distortion function of the source U is lower bounded as

$$R(D) \geq H(U) - \phi(D). \tag{5}$$

The above inequality is known as the 'Shannon lower bound' (specialized to our setting of ternary alphabets and Hamming loss).

(d) (8 points) Show that when U is uniform (on $\{0, 1, 2\}$), the Shannon lower bound holds with equality, i.e.,

$$R(D) = H(U) - \phi(D) = \log 3 - \phi(D), \quad 0 \leq D \leq 1. \tag{6}$$

[Hint: establish, by construction, existence of a joint distribution on U, V such that U is uniform and the inequalities in Part (c) hold with equalities]

Solution:

(a) Under any input distribution P_X,

$$\begin{aligned} I(X;Y) &\overset{(i)}{=} H(Y) - H(Y|X) \\ &\overset{(ii)}{=} H(Y) - H(Y \ominus X|X) \\ &\overset{(iii)}{=} H(Y) - H(Z|X) \\ &\overset{(iv)}{=} H(Y) - H(Z) \\ &\overset{(v)}{\leq} \log 3 - H(Z), \end{aligned}$$

where (i) follows from the definition of mutual information, (ii) is due to invariance of entropy to translation of the RV (or to any one-to-one transformation), (iii) is due to the channel model, (iv) to the independence of the additive channel noise component on the channel input, and (v) is because Y is ternary. On the other hand, when P_X is the uniform distribution, the distribution of Y is uniform as well, in which case (v) holds with equality.

(b) We have

$$\phi(\epsilon) = \max_{Z:Pr(Z \neq 0) \leq \epsilon} H(Z) = \max_{Z:E[\rho(Z)] \leq \epsilon} H(Z), \tag{7}$$

$$\rho(Z) = \begin{cases} 0, & \text{if } z = 0 \\ 1, & \text{otherwise.} \end{cases}$$

It can be shown that the maximum is attained by a distribution on Z of the form:

$$P_Z(z) = c(\lambda)e^{-\lambda\rho(z)}, \tag{8}$$

where $c(\lambda)$ is the normalization constant and $\lambda \geq 0$ is tuned so that the constraint is met with equality (when possible). In our case this boils down to the distribution

$$P_{Z_\epsilon}(z) = \begin{cases} 1-\epsilon, & \text{if } z = 0 \\ \epsilon/2, & \text{otherwise,} \end{cases}$$

for $0 \leq \epsilon \leq 2/3$. For $2/3 < \epsilon \leq 1$, the uniform distribution is in the constraint set, and is therefore the maximimizing distribution. Thus

$$\phi(\epsilon) = \begin{cases} H(Z_\epsilon), & \text{if } 0 \leq \epsilon \leq 2/3 \\ \log 3, & \text{if } /3 < \epsilon \leq 1 \end{cases}$$

(c) (i) From definition of mutual information.
(ii) From invariance of entropy to translation of the RV (or to any one-to-one transformation).
(iii) Conditioning reduces entropy.
(iv) Due to $P(U \ominus V \neq 0) = E[d(U,V)] \leq D$ and the definition of ϕ.
Thus, $H(U) - \phi(D)$ lower bounds any mutual information in the feasible set over which the minimum in the definition of $R(D)$ is taken, and therefore lower bounds $R(D)$.

(d) We need to find a distribution on (U, V) such that:

(a) U is uniform.

(b) $U \ominus V$ is independent on V (for equality in (iii)).

(c) $U \ominus V \sim Z_D$ (for equality in (iv)).

Taking V to be uniform, and $U = V \oplus Z_D$, for Z_D independent of V, satisfies these three conditions.

4. Gaussian source and channel *(35 points)*

- **Gaussian Channel**

  Consider the parallel Gaussian channel which has two inputs $X = (X_1, X_2)$ and two outputs $Y = (Y_1, Y_2)$, where

  $$\begin{aligned} Y_1 &= X_1 + Z_1 \\ Y_2 &= X_2 + Z_2, \end{aligned}$$

  and $Z_i \sim \mathcal{N}(0, \sigma_i^2), i = 1, 2$, are independent Gaussian random variables. We impose an average power constraint on the input X, which is

  $$\mathbb{E}[\|X\|^2] = \mathbb{E}[X_1^2 + X_2^2] \leq P$$

  (a) (10 points) Give an explicit formula for the capacity of this channel in terms of $P, \sigma_1^2, \sigma_2^2$.

  (b) (7 points) Suppose you had access to capacity-achieving schemes for the scalar AWGN channel whose capacity we derived in class. How would you use them to construct capacity-achieving schemes for this parallel Gaussian channel?

- **Gaussian Source**

  Consider a two-dimensional real valued source $U = (U_1, U_2)$ such that $U_1 \sim \mathcal{N}(0, \sigma_1^2)$, and $U_2 \sim \mathcal{N}(0, \sigma_2^2)$, and U_1 is independent of U_2. Let $d : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ be the distortion measure

  $$d(u, v) = \|u - v\|^2 = |u_1 - v_1|^2 + |u_2 - v_2|^2$$

  We wish to compress i.i.d. copies of the source U, with average per-symbol distortion no greater than D, i.e. the usual lossy compression setup discussed in class.

  (a) (10 points) Evaluate the rate-distortion function $R(D)$ explicitly in terms of the problem parameters $D, \sigma_1^2, \sigma_2^2$.

  (b) (8 points) Suppose you had access to good lossy compressors for the scalar Gaussian source whose rate-distortion function we derived in class. How would you use them to construct good schemes for this two-dimenstional Gaussian source?

Solution:

Gaussian Channel

By Shannon's channel coding theorem, the capacity of this parallel Gaussian channel is given by

$$\begin{aligned}
\max_{P_{X_1,X_2}:\mathbb{E}[X_1^2+X_2^2]\le P} I(X_1,X_2;Y_1,Y_2) &= \max_{P_{X_1,X_2}:\mathbb{E}[X_1^2+X_2^2]\le P} \left(h(Y_1,Y_2)-h(Y_1,Y_2|X_1,X_2)\right) && (9)\\
&= \max_{P_{X_1,X_2}:\mathbb{E}[X_1^2+X_2^2]\le P} \left(h(Y_1,Y_2)-h(Y_1|X_1)-h(Y_2|X_2)\right) && (10)\\
&\le \max_{P_{X_1,X_2}:\mathbb{E}[X_1^2+X_2^2]\le P} \left(h(Y_1)+h(Y_2)-h(Y_1|X_1)-h(Y_2|X_2)\right) && (11)\\
&= \max_{P_{X_1,X_2}:\mathbb{E}[X_1^2+X_2^2]\le P} \left(I(X_1;Y_1)+I(X_2;Y_2)\right) && (12)\\
&= \max_{P_1,P_2\ge 0,P_1+P_2\le P} \left(\max_{P_{X_1}:\mathbb{E}X_1^2\le P_1} I(X_1;Y_1)+\max_{P_{X_2}:\mathbb{E}X_2^2\le P_2} I(X_2;Y_2)\right) && (13)\\
&= \max_{P_1,P_2\ge 0,P_1+P_2\le P} \left(\frac{1}{2}\log\left(1+\frac{P_1}{\sigma_1^2}\right)+\frac{1}{2}\log\left(1+\frac{P_2}{\sigma_2^2}\right)\right). && (14)
\end{aligned}$$

Note that in the above chain of inequalities, if we take X_1 to be independent of X_2, then every inequality holds equality. Hence, it suffices to solve the last optimization problem, whose solution is not only an upper bound, but also a lower bound of the capacity of this parallel Gaussian channel.

Since the function $\frac{1}{2}\log(1+x)$ is increasing for $x \ge 0$, the maximum is attained when $P_1+P_2=P$. Define the following function of $P_1 \in [0,P]$:

$$f(P_1)=\frac{1}{2}\log\left(1+\frac{P_1}{\sigma_1^2}\right)+\frac{1}{2}\log\left(1+\frac{P-P_1}{\sigma_2^2}\right). \tag{15}$$

The function $f(P_1)$ is concave on $[0,P]$ and has only one maximum. Taking derivative with respect to P_1, we have

$$f'(P_1)=\frac{1/\sigma_1^2}{2(1+P_1/\sigma_1^2)}+\frac{-1/\sigma_2^2}{2(1+(P-P_1)/\sigma_2^2)}. \tag{16}$$

Setting it to zero, we have $P_1^*=\frac{P+\sigma_2^2-\sigma_1^2}{2}$. If $|\sigma_2^2-\sigma_1^2|\le P$, $P_1^*\in[0,P]$, the optimal power allocation is given by

$$\begin{aligned}
P_1^* &= \frac{P+\sigma_2^2-\sigma_1^2}{2} && (17)\\
P_2^* &= \frac{P+\sigma_1^2-\sigma_2^2}{2}. && (18)
\end{aligned}$$

If $|\sigma_1^2-\sigma_2^2|>P$, without loss of generality we assume $\sigma_1^2 <= \sigma_2^2$. Then we should set $P_1^*=P, P_2^*=0$. In other words, if the quality of the two Gaussian channels are very different (in the sense that $|\sigma_1^2 - \sigma_2^2| > P$), then we should allocate all the power to the stronger channel.

To sum up, we have the capacity of this parallel Gaussian channel equal to

$$\mathsf{C}_{\text{parallel}}(P) = \frac{1}{2}\log\left(1 + \frac{P_1^*}{\sigma_1^2}\right) + \frac{1}{2}\log\left(1 + \frac{P - P_1^*}{\sigma_2^2}\right), \tag{19}$$

where

$$P_1^* = \begin{cases} \frac{P+\sigma_2^2-\sigma_1^2}{2} & |\sigma_1^2 - \sigma_2^2| \le P \\ P & |\sigma_1^2 - \sigma_2^2| > P, \sigma_1^2 < \sigma_2^2 \\ 0 & |\sigma_1^2 - \sigma_2^2| > P, \sigma_1^2 > \sigma_2^2 \end{cases} \tag{20}$$

Suppose we now have the capacity achieving schemes for single Gaussian channel. In order to achieve the capacity of this parallel Gaussian channel, we first allocate power P_1^* to channel 1, power $P - P_1^*$ to channel 2. Then we take the codebook for Gaussian channel with rate $\frac{1}{2}\log\left(1 + \frac{P_1^*}{\sigma_1^2}\right)$ and power P_1^* for channel 1, and take codebook for Gaussian channel with rate $\frac{1}{2}\log\left(1 + \frac{P-P_1^*}{\sigma_2^2}\right)$ and power $P - P_1^*$ for channel 2. The joint codebook has rate $\mathsf{C}_{\text{parallel}}(P)$.

Gaussian Source

By Shannon's rate distortion theorem, the rate distortion function of this source is given by

$$\begin{aligned}
R_{\text{joint}}(D) &= \min_{P_{V_1,V_2|U_1,U_2}:\mathbb{E}[|U_1-V_1|^2+|U_2-V_2|^2]\le D} I(U_1,U_2;V_1,V_2) && (21)\\
&= \min_{P_{V_1,V_2|U_1,U_2}:\mathbb{E}[|U_1-V_1|^2+|U_2-V_2|^2]\le D} (h(U_1,U_2) - h(U_1,U_2|V_1,V_2)) && (22)\\
&= \min_{P_{V_1,V_2|U_1,U_2}:\mathbb{E}[|U_1-V_1|^2+|U_2-V_2|^2]\le D} (h(U_1) + h(U_2) - h(U_1|V_1,V_2) - h(U_2|U_1,V_1,V_2)) && (23)\\
&\ge \min_{P_{V_1,V_2|U_1,U_2}:\mathbb{E}[|U_1-V_1|^2+|U_2-V_2|^2]\le D} (h(U_1) + h(U_2) - h(U_1|V_1) - h(U_2|V_2)) && (24)\\
&= \min_{P_{V_1,V_2|U_1,U_2}:\mathbb{E}[|U_1-V_1|^2+|U_2-V_2|^2]\le D} (I(U_1;V_1) + I(U_2;V_2)) && (25)\\
&= \min_{D_1\ge 0,D_2\ge 0,D_1+D_2\le D}\left(\min_{P_{V_1|U_1}:\mathbb{E}[|U_1-V_1|^2]\le D_1} I(U_1;V_1) + \min_{P_{V_2|U_2}:\mathbb{E}[|U_2-V_2|^2]\le D_2} I(U_2;V_2)\right) && (26)\\
&= \min_{D_1\ge 0,D_2\ge 0,D_1+D_2\le D}\left(\max\{0,\frac{1}{2}\log\frac{\sigma_1^2}{D_1}\} + \max\{0,\frac{1}{2}\log\frac{\sigma_2^2}{D_2}\}\right) && (27)
\end{aligned}$$

We note that in the above chain of inequalities, if we take the joint test channel $P_{V_1,V_2|U_1,U_2}$ to be of form $P_{V_1|U_1}P_{V_2|U_2}$, all inequalities hold equality. Hence, it suffices to solves the last optimization problem and the resulting answer is not only a lower bound but also an upper bound on the rate distortion function of this two dimensional source.

Since the function $\frac{1}{2}\log(1/x)$ is non-increasing for $x > 0$, the minimum is attained when $D_1 + D_2 = D$. We define function

$$g(D_1) = \max\{0, \frac{1}{2}\log\frac{\sigma_1^2}{D_1}\} + \max\{0, \frac{1}{2}\log\frac{\sigma_2^2}{D - D_1}\}. \tag{28}$$

When $D_1 \le \sigma_1^2, D - D_1 \le \sigma_2^2$, taking derivatives of $g(D_1)$, we obtain

$$g'(D_1) = \frac{1}{2}\left(\frac{1}{D - D_1} - \frac{1}{D_1}\right). \tag{29}$$

Setting it to zero, we have $D_1 = D/2$. Hence, if $D \le 2\min\{\sigma_1^2, \sigma_2^2\}$, the optimal distortion allocation is

$$D_1^* = D/2, D_2^* = D/2. \tag{30}$$

When $\sigma_1^2 + \sigma_2^2 \ge D > 2\min\{\sigma_1^2, \sigma_2^2\}$, without loss of generality we assume $\sigma_1^2 \le \sigma_2^2$. Then we should use zero rate to describe U_1, and allocate distortion $D - \sigma_1^2$ to U_2. If $D > \sigma_1^2 + \sigma_2^2$, we simply use zero rate to describe both U_1 and U_2.

In other words, the joint rate distortion function is given by

$$R_{\text{joint}}(D) = \begin{cases} \frac{1}{2}\log\frac{2\sigma_1^2}{D} + \frac{1}{2}\log\frac{2\sigma_2^2}{D} & D \le 2\min\{\sigma_1^2, \sigma_2^2\} \\ \frac{1}{2}\log\frac{\max\{\sigma_1^2,\sigma_2^2\}}{D - \min\{\sigma_1^2,\sigma_2^2\}} & \sigma_1^2 + \sigma_2^2 \ge D > 2\min\{\sigma_1^2, \sigma_2^2\} \\ 0 & D > \sigma_1^2 + \sigma_2^2 \end{cases} \tag{31}$$

Suppose we now have good lossy compressors for Gaussian source. To achieve the rate distortion function of this two dimensional source, if $D \le 2\min\{\sigma_1^2, \sigma_2^2\}$, we use the rate distortion code for source U_1 and U_2 independently under distortion $D/2$ for each source. If $\sigma_1^2 + \sigma_2^2 \ge D > 2\min\{\sigma_1^2, \sigma_2^2\}$, we simply use a constant 0 to encode the source with smaller variance (say U_1), and use rate distortion code to encode another source with distortion $D - \min\{\sigma_1^2, \sigma_2^2\}$. If $D > \sigma_1^2 + \sigma_2^2$, we simply use constant 0 to describe both U_1 and U_2.