Multi-User Information Theory 2 February 28th, 2013

Lecture 12

Lecturer: Haim Permuter *Scribe: Maxim Lvov*

I. Fixed-length lossless source coding with side information

In the first course in information theory, we've seen the next problem: We are given two sequences $\{X_k\}_{k=1}^{\infty}$ of i.i.d random variables ($X_k \sim P_X$) which we want to compress. An example can be seen in Fig. 1. where the sequence $\{Y_k\}_{k=1}^{\infty}$ is a side information ($Y_k \sim P_Y$) known both to the encoder and to the decoder, and (X_k, Y_k) are i.i.d random vectors. We've seen that the minimal achievable rate was given

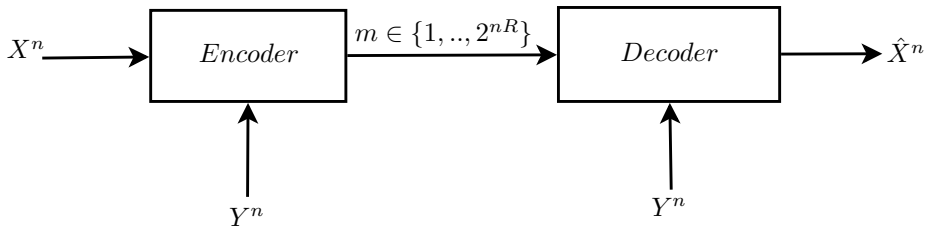

Fig. 1. Lossless source coding with side information system.

by the formula:

$$R \geq H(X|Y) \tag{1}$$

We've also seen that this rate is achievable by multiplexing.

Now we want to find, what is the minimal achievable rate when the side information is known only to the decoder? We know for sure that $H(X|Y)$ is a lower bound of the rate, since we cant compress more that we could when $\{Y_k\}_{k=1}^{\infty}$ was known to the encoder, but we'll prove now that this rate is also achievable.

Definition 1 (Source coding with side information) A lossless source code $(n, 2^{nR})$ with side information available only at the decoder, consists of:

- An encoding function $f : \mathcal{X}^n \to \{1, .., 2^{nR}\}$
- A decoding function $g : \{1, .., 2^{nR}\} \times \mathcal{Y}^n \to \hat{\mathcal{X}}^n$

A rate R is called achievable, if there exists a sequence of codes $(n, 2^{nR})$ such that

$$\lim_{n\to\infty} \Pr(X^n \neq g(f(X^n), Y^n)) = 0.$$

Theorem 1 Let $\{(X_k, Y_k)\}_{k=1}^{\infty}$ be an i.i.d sequence of random variables, distributed by the p.m.f $P_{X,Y}$. Suppose that we want to encode the sequence $\{X_k\}_{k=1}^{\infty}$ with a constant length lossless source coding with a rate R, while $\{Y_k\}_{k=1}^{\infty}$ is a side information known only to the decoder. Then the minimal rate is given by (1).

Proof: We only need to show achievability to prove the theorem.
Suppose we choose $R = H(X|Y) + 2\epsilon$. We construct the next code: We take all of the vectors $x^n \in T_\epsilon^n(X)$ and divide them randomly into 2^{nR} bins. There are not more than $2^{n(H(X)+\epsilon)}$ vectors in $T_\epsilon^n(X)$, so we'll have not more than $2^{n(I(X;Y)-\epsilon)}$ vectors in each bin.

The encoder: Gets a vector X^n, and looks for the bin containing it. The index of the bin M, is the compressed value representing the vector X^n. Pay attention that this index (in its binary representation) has exactly R bits, since there are exactly 2^{nR} bins.
The decoder: Gets the index M, looks into the bin associated with that index, and searches for a vector $\hat{x}^n$ that is jointly typical with the vector Y^n. If such $\hat{x}^n$ is found, it is declared as the decoded x^n, otherwise declared an error.
Error analysis: We have errors only in two cases:
E_1: X^n and Y^n aren't jointly typical (This probability goes to zero as $n \to \infty$ by L.L.N).
E_2: There is more than one vector x^n in the bin associated with the index M that is jointly typical with Y^n. For that event to happen, at least one of the other vectors in that bin (which were chosen randomly from $T_\epsilon^n(X)$ and independently of Y^n) should be jointly typical with Y^n. For every such vector, the probability for that to happen is not more than $2^{-n(I(X;Y)-0.5\epsilon)}$. Now by union bound, we have:

$$\Pr(E_2) \leq \sum_{X^n \neq x^n \in bin_M} P((x^n, Y^n) \in T_\epsilon^n(X,Y)) \tag{2}$$

$$\leq \sum_{X^n \neq x^n \in bin_M} 2^{-n(I(X;Y)-0.5\epsilon)} \tag{3}$$

$$\leq |bin_M| \cdot 2^{-n(I(X;Y)-0.5\epsilon)} \tag{4}$$

$$\leq 2^{n(I(X;Y)-\epsilon)} \cdot 2^{-n(I(X;Y)-0.5\epsilon)} \tag{5}$$

$$= 2^{-n\frac{\epsilon}{2}} \tag{6}$$

In (4), the term $|bin_M|$ represents the cardinality of the bin associated with the index M. The expression in (6) tends to zero as $n \to \infty$. ■

II. Lossless Multi user source coding (Slepian-Wolf)

Each encoder can encode its sequence into indexes at rates R_1 and R_2 respectively.
The decoder gets the indexes from both encoders, and need to reconstruct the sequences $\{X_k\}$ and $\{Y_k\}$,

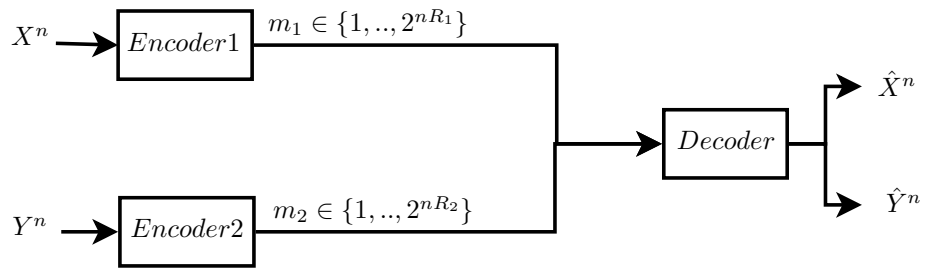

Fig. 2. Multi user lossless source coding system.

without error.

We are asked to find the achievable region, that is, the closer of the set:

$$\{(R_1, R_2) \in \mathbb{R}^2 : R_1 \; and \; R_2 \; are \; achievable\} \tag{7}$$

We first give a formal definition of the terms, and then find the acheivable region.

Definition 2 (Lossless multi source code) Given the sequences $\{X_k\}_{k=1}^{\infty}$ and $\{Y_k\}_{k=1}^{\infty}$ at encoders 1 and 2, a lossless multi source code $(n, 2^{nR_1}, 2^{nR_2})$ contains:

- Two encoding functions $f_1 : \mathcal{X}^n \to \{1,..,2^{nR_1}\}$ and $f_2 : \mathcal{Y}^n \to \{1,..,2^{nR_2}\}$
- Two decoding functions $(g_1, g_2) : \{1,..,2^{nR_1}\} \times \{1,..,2^{nR_2}\} \to \mathcal{X}^n \times \mathcal{Y}^n$.

The probability of error for a code $(n, 2^{nR_1}, 2^{nR_2})$ is defined by

$$P_e^n = \Pr((X^n, Y^n) \neq (g_1(f_1(X^n), f_2(Y^n)), g_2(f_1(X^n), f_2(Y^n)))) \tag{8}$$

The rates (R_1, R_2) are called achievable if for any $\epsilon > 0$ there exists a code $(n, 2^{nR_1}, 2^{nR_2})$ with a probability of error $P_e^n < \epsilon$.

The closure of all achievable rates (R_1, R_2) is called the achievable region.

The next theorem shows what the achievable region for this problem is. In order to prove it, we'll show that the achievable region is contained inside and contains the offered region.

Theorem 2 The achievable region for the problem stated above is the set of all $\{(R_1, R_2)\} \in \mathbb{R}^+ \times \mathbb{R}^+$ that satisfy (9)-(11)

$$R_1 + R_2 \geq H(X, Y) \tag{9}$$

$$R_1 \geq H(X|Y) \tag{10}$$

$$R_2 \geq H(Y|X) \tag{11}$$

Before we show a formal proof, we would explain intuitively why the above is correct. For any achievable rates (R_1, R_2), equation (9) must hold because when both of the encoders know both $\{X_k\}$ and $\{Y_k\}$

(which is equivalent to have only one encoder that has these sequences at its input), the best we can do is to compress at rate: $R_1 + R_2 = R = H(X,Y)$. In our case, we have a worse situation (each encoder has access to only one sequence), so we can't get a better result (a smaller rate).

Equation (10) must hold because when the sequence $\{Y_k\}$ is known both to the decoder, and to encoder 1, it serves as a side information, and cannot decrease R_1 below $H(X|Y)$.

Equation (10) must hold from the same considerations.

Therefore, we see that achievable region must be contained in the greyed region, as shown in Fig.3. We

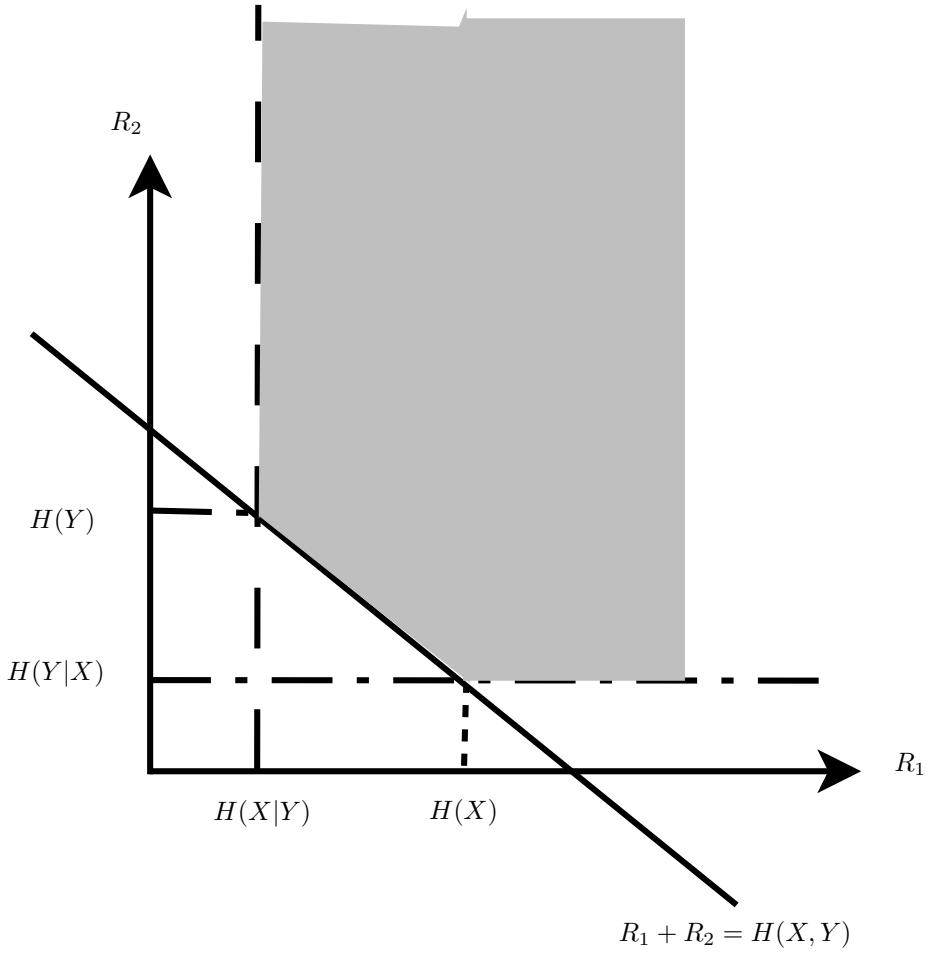

Fig. 3. Multi user lossless source coding system.

now explain why the greyed region is contained in the achievable region. The rates $(R_1 = H(X), R_2 = H(Y|X))$ are achievable by the next coding scheme: Encoder 1 is using the usual code with a rate $H(X)$ to achieve zero error probability. The decoder receives that sequence $\{X_k\}$, and uses it as side information to decode the sequence $\{Y_k\}$. This side information is known only to encoder 1 and to the decoder, but

we've seen in previous section that its enough for encoder 2 to achieve the rate $H(Y|X)$).

By the same considerations, the rates $(R_1 = H(X|Y), R_2 = H(Y))$ are also achievable. Now, by using the convexity of the achievable region, we get the desired result.

We now give a formal proof for the theorem stated above.

Proof: Achievability: We show the achievability of the rates $(R_1 = H(X), R_2 = H(Y|X))$. The achievability of the rates $(R_1 = H(X|Y), R_2 = H(Y))$ can be shown by the same way. The other greyed region will be achievable by considerations of convexity of the achievable region.

For a given $\epsilon > 0$, $R_1 > H(X)$ and $R_2 > H(Y|X)$ we know that there exists $N_1(\epsilon) \in \mathbb{N}$, such that for $n > N_1$ there exist functions $f : \mathcal{X}^n \to \{1, .., 2^{nR_1}\}$ and $g : \{1, .., 2^{nR_1}\} \to \mathcal{A} \subseteq \mathcal{X}^n$, such that $\Pr(g(f(X^n)) \neq X^n) < \frac{\epsilon}{2}$. We know that these function exist, because these functions serve in ordinary lossless source coding. We therefore choose our first encoding function f_1 to be $f_1 = f$, and $g_1 = g$.

We also know, that if the symbols X^n were reconstructed correctly at the decoder ($\forall n > N_1$), then exists $N_2(\epsilon) \in \mathbb{N}$, such that for $n > N_2 \geq N_1$ there exist functions $f : \mathcal{Y}^n \to \{1, .., 2^{nR_2}\}$ and $g : \{1, .., 2^{nR_2}\} \to \mathcal{B} \subseteq \mathcal{Y}^n$, such that $\Pr(g(f(Y^n)) \neq Y^n) < \frac{\epsilon}{2}$. We've given these functions in section I, and explained why the above holds. In fact, these functions exist even if X^n were not reconstructed correctly, but we cannot be sure that the above inequality will hold.

We therefore choose our second encoding function f_2 to be $f_2 = f$, and $g_2 = g$.

Error analysis: An error occurs if:

- E_1: X^n isn't reconstructed correctly at the decoder
- E_2 : X^n is reconstructed correctly at the decoder, but not Y^n.

Because of the way we've chosen our encoding and decoding functions, we have:

$$P_e^n = \Pr(E_1 \cup E_2) \tag{12}$$

$$\stackrel{(a)}{=} \Pr(E_1) + \Pr(E_2) \tag{13}$$

$$\leq \frac{\epsilon}{2} + \Pr(E_2|E_1^c)\Pr(E_1^c) \tag{14}$$

$$\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} \cdot 1 = \epsilon \tag{15}$$

where [(a)] follows from the fact that E_1 and E_2 are disjoint events.

Therefore, the rates (R_1, R_2) are achievable.

Converse: We finish this section with a converse proof for (10).

$$nR_1 \geq H(M_1) \tag{16}$$

$$\stackrel{(a)}{\geq} H(M_1|Y^n) \tag{17}$$

$$\stackrel{(b)}{=} H(M_1|Y^n, m_2) \tag{18}$$

$$\stackrel{(c)}{=} H(X^n, M_1|Y^n, m_2) - H(X^n|M_1, Y^n, m_2) \tag{19}$$

$$\stackrel{(d)}{=} H(X^n, M_1|Y^n, M_2) - H(X^n|M_1, Y^n, m_2, \hat{X}^n) \tag{20}$$

$$\stackrel{(e)}{\geq} H(X^n|Y^n) - (1 + n\epsilon log|\mathcal{X}|) \tag{21}$$

$$= nH(X|Y) - n\epsilon_n \tag{22}$$

where:

(a) follows the fact that conditioning reduces entropy.

(b) follow the fact that m_2 is deterministic given Y^n.

(c) is due to chain rule.

(d) follow the fact that $\hat{X}^n$ is deterministic given M_1, M_2.

(e) follows from Fano's inequality.

By dividing by n, and taking $n \to \infty$ we get the desired result. ■

III. Multi User Rate Distortion

We return now to the problem with lossy compressing. We are given the source coding scheme shown in Fig .4. The encoder receives the sequence $\{X_k\}_{k=1}^{\infty}$ of i.i.d random variables, with a p.m.f P_X, and encodes every n symbols X^n into an index $M \in \{1, .., 2^{nR}\}$. The decoder receives the index M, and reconstructs the n symbols $\hat{X}^n$. Pay attention that we don't demand that $X^n = \hat{X}^n$ with high probability. Instead of that, we have a "*distortion*" function that measures, how much $\hat{X}^n$ is "far" from X^n. We demand that the distortion won't be too big with high probability.

Here again, we're interested in the possible rates R that satisfy the given distortion.

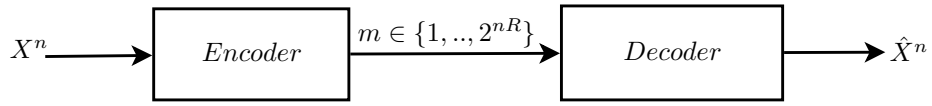

Fig. 4. Rate-distortion system.

Definition 3 (Distortion) For given alphabets $\mathcal{X}, \hat{\mathcal{X}}$, a distortion function is a function $d : (\mathcal{X}, \hat{\mathcal{X}}) \to \mathbb{R}^+$. For a given distortion function d, we also define: $d(X^n, \hat{X}^n) = \frac{1}{n}\sum_{k=1}^{n} d(X_k, \hat{X}_k)$

Example 1 There are many useful distortion functions:

- Hamming distortion: $d(x, \hat{x}) = \begin{cases} 1 & \text{if } x = \hat{x} \\ 0 & \text{if } x \neq x \end{cases}$
  For this distortion, $d(x^n, \hat{x}^n)$ measures the mean error rate for the reconstruction.

- Square error distortion: $d(x,\hat{x}) = (x-\hat{x})^2$. For this distortion, $d(x^n,\hat{x}^n)$ measures the mean square error of $\hat{x}^n$.

Definition 4 (Coding Scheme) A code $(n, 2^{nR})$ consists of:

- An encoding function $f : \mathcal{X}^n \to \{1,..,2^{nR}\}$
- A decoding function $g : \{1,..,2^{nR}\} \to \hat{\mathcal{X}}^n$

Definition 5 (Achievable rate) For a given distortion function $d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+$, and for a maximum allowed distortion D, a rate R is called achievable, if there a sequence of codes $(n, 2^{nR})$ such that

$$\lim_{n\to\infty} E[d(X^n, \hat{X}^n)] \leq D \tag{23}$$

In some places, instead of demanding (23) to hold, a bit different condition is used:
R is called achievable if for any given $\epsilon > 0$, there exists a sequence of codes $(n, 2^{nR})$ such that

$$\lim_{n\to\infty} \Pr(d(X^n, \hat{X}^n) > D + \epsilon) = 0 \tag{24}$$

The minimal achievable rate for a given maximum allowed distortion D is denoted by $R(D)$

We are interested in finding the minimal achievable rate $R(D)$. The main rate-distortion theorem states what this minimal rate is.

Theorem 3 The minimal achievable rate $R(D)$ for the problem described above, is given by (25)

$$R_I(D) = \min\{I(X;\hat{X})\} \tag{25}$$

where the minimum is taken over all of the functions $P_{\hat{X}|X}$ that satisfy

$$E[d(X,\hat{X})] \leq D \tag{26}$$

A. Rate-Distortion with side information

We are now interested in the system shown in Fig .1. The system is the same as described in section I, but this time, we're given a distortion function d, and a maximum allowable distortion D instead of demanding that $P_e^n \to 0$ as $n \to \infty$.

Definition 6 (Rate-Distortion Source coding with side information) A lossless source code $(n, 2^{nR})$ with side information available at both the encoder and the decoder, consists of:

- An encoding function $f : \mathcal{X}^n \times \mathcal{Y}^n \to \{1,..,2^{nR}\}$
- A decoding function $g : \{1,..,2^{nR}\} \times \mathcal{Y}^n \to \hat{\mathcal{X}}^n$

Achievable rate is defined as in Definition(5).

Theorem 4 For the problem of Rate-Distortion source coding with side information, available both at the encoder and at the decoder, the minimal rate $R(D)$ is given by

$$R_I(D) = \min\{I(X;\hat{X}|Y)\} \tag{27}$$

where the minimum is taken over all of the functions $P_{\hat{X}|X,Y}$ that satisfy equation (23).

But what if the encoder doesn't have the side information? In that case, the result is different, and is the subject of the next discussion.

B. Side information available only at the decoder

We're discussing a source coding system as described in Definition(1). Achievable rate is defined as in Definition(5). This time, the result will be different from $R_I(D)$ as given in (27), and is given in the next theorem.

Theorem 5 For a rate-distortion coding system with side information $\{Y_n\}$ known only at the decoder, the minimal achievable rate $R(D)$ is given by the next expression:

$$R_I(D) = \min\{I(X;U|Y)\} \tag{28}$$

where the minimum is taken over all of the functions $P_{U|X,Y}$ that satisfy the next conditions:

- (U, X, Y) form a Markov chain $U - X - Y$, or $P_{U|X,Y}(u|x, y)$ doesn't depend on y.
- There exists a function $f(u, y)$ such that if we let $\hat{X} = f(U, Y)$ then

$$E[d(X, \hat{X})] \leq D \tag{29}$$

We first prove the converse, and then the achievability.

The Converse:

Suppose that R is achievable. Then there exists a sequence of codes $(n, 2^{nR})$ that satisfies equation (23). We denote by M the compressed value of the vector X^n, and by $\hat{X}^n$ the reconstructed vector. For a given block of length n, we have:

$$\begin{aligned}
nR &\geq H(M) && (30)\\
&\geq H(M|Y^n) && (31)\\
&\geq I(X^n; M|Y^n) && (32)\\
&= H(X^n|Y^n) - H(X^n|Y^n, M) && (33)\\
&= \sum_{i=1}^{n} H(X_i|Y_i) - \sum_{i=1}^{n} H(X_i|X^{i-1}, Y^n, M) && (34)
\end{aligned}$$

$$\geq \sum_{i=1}^{n} H(X_i|Y_i) - \sum_{i=1}^{n} H(X_i|Y^n, M) \tag{35}$$

$$= \sum_{i=1}^{n} H(X_i|Y_i) - \sum_{i=1}^{n} H(X_i|Y_1, .., Y_{i-1}, Y_{i+1}, .., Y_n, m, Y_i) \tag{36}$$

We have used in equations (31,35) that coditioning reduces entropy, and we've used chain-rule in equation (34).

Denote by $U_i = (Y_1, .., Y_{i-1}, Y_{i+1}, .., Y_n, M)$. This R.V satisfies three interesting properties:

- The triple (U_i, X_i, Y_i) is a Markov chain: $U_i - X_i - Y_i$, because U_i has information about Y_i only through M which is a function of X_i.
- $\hat{X}_i$ is a function f of $(U_i, Y_i) = (Y^n, M)$
- $E[d(X_i, f(U_i, Y_i))] \leq E[d(X_i, \hat{X}_i)]$

Therefore, by definition of R_I, we have:

$$nR \geq \sum_{i=1}^{n} H(X_i|Y_i) - \sum_{i=1}^{n} H(X_i|Y_1, .., Y_{i-1}, Y_{i+1}, .., Y_n, M, Y_i) \tag{37}$$

$$= \sum_{i=1}^{n} H(X_i|Y_i) - \sum_{i=1}^{n} H(X_i|U_i, Y_i) \tag{38}$$

$$= \sum_{i=1}^{n} I(X_i; U_i|Y_i) \tag{39}$$

$$\overset{(a)}{\geq} \sum_{i=1}^{n} R_I(E[d(X_i, \hat{X}_i)]) \tag{40}$$

$$= n\left(\frac{1}{n}\sum_{i=1}^{n} R_I(E[d(X_i, \hat{X}_i)])\right) \tag{41}$$

$$\overset{(b)}{\geq} nR_I\left(\frac{1}{n}\sum_{i=1}^{n} E[d(X_i, \hat{X}_i)]\right) \tag{42}$$

$$= nR_I(E[d(X^n, \hat{X}^n)]) \tag{43}$$

where:

(a) follows from the definition of R_I

(b) follow the fact that $R_I(D)$ is convex in D (we state it without proof).

Now, by dividing by n, letting n tend to ∞, and using the facts that $R_I(D)$ is convex in D (and therefore continues), and is a decreasing function, we get the result: $R \geq R_I(D)$.

Achievability:

To show achievability, we suppose that there is some p.m.f $P_{U|X}$ and a function $f(u, y)$ such that

$$E[d(X, f(U, Y))] \leq D \tag{44}$$

The expectation value is taken with respect to $P_{X,Y,U} = P_{X,Y}P_{U|X}$.

We will also assume that

$$d(x,\hat{x}) < \infty, \forall (x,\hat{x}) \in \mathcal{X} \times \hat{\mathcal{X}} \tag{45}$$

Otherwise, we can have events with probability almost equal to zero that would affect $E[d(X,\hat{X})]$.

Let $R = I(X;U|Y) + 2\epsilon$. We'll show now that this rate is achievable. First note that:

$$R = I(X;U|Y) + 2\epsilon \tag{46}$$

$$= H(U|Y) - H(U|X,Y) + 2\epsilon \tag{47}$$

$$= H(U|Y) - H(U|X) + 2\epsilon \tag{48}$$

$$= I(X;U) - I(U;Y) + 2\epsilon \tag{49}$$

Code construction: We create $2^{n(I(X,U)+\epsilon)}$ codewords U^n independently and with respect to the p.m.f P_U. We divide these codewords randomly into 2^{nR} bins. Each bin will have $2^{n(I(U;Y)-\epsilon)}$ codewords.

The encoder: Gets X^n, and searches for a codeword u^n it has generated that would be jointly typical with X^n. If it finds such codeword, It declares its bin number to be the compressed data, otherwise it declares an error.

The decoder: Gets the bin number, and looks inside it for a codeword u^n that would be jointly typical with Y^n. If it finds such codeword, it declares $\hat{X}^n = f(u^n, Y^n)$ to be the decoded symbol. If it doesn't find, it declares an error.

Probability and distortion analysis:

- The encoder would declare an error with a very low probability because of Covering Lemma.
- The decoder would declare an error with a very low probability because of Markov Lemma: if $U - X - Y$ form a Markov chain, (U^n, X^n) are jointly typical, and Y^n is created with respect to $P_{Y^n|X^n}(y^n|x^n) = \prod_{i=1}^{n} P_{Y|X}(y_i|x_i)$ then with very high probability (tends to 1 as $n \to \infty$) the triple (U^n, X^n, Y^n) will be jointly typical.
- The probability that there is more than one u^n that is jointly typical with Y^n is very low, because there are only $2^{n(I(U;Y)-\epsilon)}$ codewords $\{u^n\}$ in each bin.

We'll now show why the distortion is almost equal to D (in case the errors described above didn't happened). The distortion is related to the type of the random vectors in the sense that its

$$d(X^n, \hat{X}^n) = \frac{1}{n}\sum_{i=1}^{n} d(X,\hat{X}) = \sum_{(x,\hat{x}) \in \mathcal{X} \times \hat{\mathcal{X}}} \tilde{P}(x,\hat{x}) d(x,\hat{x}) \tag{50}$$

where $\tilde{P}$ is the joint type of $(X^n, \hat{X}^n)$. The expression above can be understood as the "mean" or "empirical expectation" of the distortion.

If the joint type of (X^n, U^n, Y^n) is close to its real p.m.f then the above "mean" is close to the real expectation value. More precisely:

$$d(X^n, \hat{X}^n) = \sum_{(x,\hat{x})\in\mathcal{X}\times\hat{\mathcal{X}}} \tilde{P}(x,\hat{x})d(x,\hat{x}) \tag{51}$$

$$= \sum_{(x,\hat{x})\in\mathcal{X}\times\hat{\mathcal{X}}} (\tilde{P}(x,\hat{x}) - P(x,\hat{x}) + P(x,\hat{x})d(x,\hat{x}) \tag{52}$$

$$\leq E[d(X,\hat{X})] + |\mathcal{X}\times\hat{\mathcal{X}}| \max_{x,\hat{x}}\{d(x,\hat{x})|\tilde{P}(x,\hat{x}) - P(x,\hat{x})|\} \tag{53}$$

The first term is not more than D, and the second term tends to zero as $n \to \infty$.

REFERENCES

[1] Slepian, D and Wolf, J K (1973). Noiseless coding of correlated information sources. IEEE Transactions on information Theory 19: 471-480.

[2] Wyner, A.; Ziv, J.; , "The rate-distortion function for source coding with side information at the decoder," Information Theory, IEEE Transactions on , vol.22, no.1, pp. 1- 10, Jan 1976.

[3] Gallager, R. G., "Source Coding With Side Information and Universal Coding", M.I.T. LIDS-P-937, 1976 (revised 1979).