Lossy Source Coding with Side Information: Wyner Ziv & Generalizations

John MacLaren Walsh, Ph.D.
Multiterminal Information Theory, Spring Quarter, 2012

1 References


2 Strongly Typical Sequences and the Markov Lemma

Consider a sequence of iid random variables $X_n, n \in \{1, \ldots, N\}$ whose elements are I.I.D. according to $p_X(x)$ and define $X^N = (X_1, X_2, \ldots, X_N)$. The empirical distribution indicating the number of elements of this sequence that are equal to a particular $x \in \mathcal{X}$ will convergence to the true distribution $p_X(x)$ in probability as $N \to \infty$. That is

$$\frac{1}{N}N(x|X^N) \to p_X(x) \quad \text{in } P \text{ as } N \to \infty \quad (1)$$

To see this begin by observing that

$$\frac{1}{N}N(x|X^N) = \frac{1}{N} \sum_{n=1}^{N} 1_x(X_n), \text{ where } 1_a(x) = \begin{cases} 1 & \text{if } x = a \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

is a sample average of the $N$ IID random variables $1_x(X_n)$.

Since it is a sample average, the weak law of large numbers states that it must converge in probability to the expectation $\mathbb{E}[1_x(X_n)] = p_X(x)$. Motivated by this convergence, we can define the set of strongly typical sequences as

$$T^N_\epsilon(X) = \left\{ x \in \mathcal{X}^N \left| \frac{1}{N}N(x|x) - p_X(x) \leq \frac{\epsilon}{|\mathcal{X}|} \right\} \quad (3)$$

The convergence in probability (1) states precisely that

$$\mathbb{P}[X^N \in T^N_\epsilon(X)] \to 1 \text{ as } N \to \infty \quad (4)$$

The set $T^N_\epsilon(X)$ represents a convenient definition of the set of typical sequences because, among other reasons, for any function $f : \mathcal{X} \to \mathbb{R}$, and for any sequence $x \in T^N_\epsilon(X)$ the empirical average $\frac{1}{N} \sum_{n=1}^{N} f(x_n)$ is easily bounded as close to the expectation $\mathbb{E}[f(X)]$, in particular

$$\frac{1}{N} \sum_{n=1}^{N} f(x_n) = \sum_{x \in \mathcal{X}} f(x) \frac{1}{N}N(x|x) \leq \sum_{x \in \mathcal{X}} f(x)p_X(x) + \sum_{x \in \mathcal{X}} |f(x)| \frac{\epsilon}{|\mathcal{X}|} \leq \mathbb{E}[f(X)] + \epsilon f_{\max} \quad (5)$$

where $f_{\max} = \max_{x \in \mathcal{X}} |f(x)|$, and

$$\frac{1}{N} \sum_{n=1}^{N} f(x_n) = \sum_{x \in \mathcal{X}} f(x) \frac{1}{N}N(x|x) \geq \sum_{x \in \mathcal{X}} f(x)p_X(x) - \sum_{x \in \mathcal{X}} |f(x)| \frac{\epsilon}{|\mathcal{X}|} \geq \mathbb{E}[f(X)] - \epsilon f_{\max} \quad (6)$$
Putting these together we see that
\[ x \in T^N_e(X) \Rightarrow \left| \frac{1}{N} \sum_{n=1}^{N} f(x_n) - \mathbb{E}[f(X)] \right| \leq \epsilon_{\text{max}}. \]  

(7)

By selecting the function \( f(\cdot) = -\log_2(p_X(\cdot)) \) we can bound the probabilities \( p_{X^N}(x) \) of those \( x \in T^N_e(X) \) in a similar manner as we did for \( A^N_e(X) \). Indeed, we observe that for \( x \in T^N_e(X) \), since \( -\frac{1}{N} \log_2(p_{X^N}(x)) = \frac{1}{N} \sum_{n=1}^{N} \log_2(p_X(x_n)) \), applying (7) we have
\[ \left| -\frac{1}{N} \log_2(p_{X^N}(x)) - H(X) \right| \leq cc, \quad c = \max_{x \in X} -\log_2(p_X(x)) \]

(8)

Moving this around we observe that
\[ x \in T^N_e(X) \Rightarrow 2^{-N(H(X)+cc)} \leq p_{X^N}(x) \leq 2^{-N(H(X)-cc)}, \]

so that the probabilities of the sequences in \( T^N_e(X) \) are all roughly \( 2^{-N(H(X))} \), just as with \( A^N_e(X) \). We then can bound the number of elements in \( T^N_e(X) \) in the same manner as we did with \( A^N_e(X) \). Indeed, the chain
\[ 1 \geq \mathbb{P}[X^N \in T^N_e(X)] = \sum_{x \in T^N_e(X)} p_{X^N}(x) \geq \sum_{x \in T^N_e(X)} 2^{-N(H(X)+cc)} = |T^N_e(X)|2^{-N(H(X)+cc)} \]

(10)

shows that \( |T^N_e(X)| \leq 2^{N(H(X)+cc)} \). We can also easily underbound the number of elements in \( T^N_e(X) \). Indeed, because the probability of being in \( T^N_e \) converges to 1 as \( N \to \infty \) (4), for any \( \epsilon \), if we take \( N \) large enough, we have
\[ 1 - \epsilon \leq \mathbb{P}[X^N \in T^N_e(X)] = \sum_{x \in T^N_e(X)} p_{X^N}(x) \leq \sum_{x \in T^N_e(X)} 2^{-N(H(X)-cc)} = |T^N_e(X)|2^{-N(H(X)-cc)}. \]

(11)

Thus, for \( N \) large enough \( |T^N_e(X)| \geq (1 - \epsilon)2^{N(H(X)-cc)}. \)

### 2.1 Joint Strong Typicality

The set of jointly typical sequences for a pair of random variables, e.g. \( T^N_e(Y,Z) \), can be thought of as an instance of the set of typical sequences \( T^N_e(X) \) by selecting the random variable \( X = (Y,Z) \). Indeed, we observe that the inequalities
\[ \left| \frac{1}{N} N(y,z|y,z) - p_{Y,Z}(y,z) \right| \leq \frac{\epsilon}{|Y||Z|} \]

(12)

directly imply the inequalities
\[ \left| \frac{1}{N} N(y|y) - p_{Y}(y) \right| \leq \frac{\epsilon}{|Y|} \quad \text{and} \quad \left| \frac{1}{N} N(z|z) - p_{Z}(z) \right| \leq \frac{\epsilon}{|Z|} \]

(13)

via the summations
\[ \left| \frac{1}{N} N(y|y) - p_{Y}(y) \right| = \left| \frac{1}{N} \sum_{z \in Z} N(y,z|y,z) - \sum_{z \in Z} p_{Y,Z}(y,z) \right| \leq \sum_{z \in Z} \left| \frac{1}{N} N(y,z|y,z) - p_{Y,Z}(y,z) \right| \leq \sum_{z \in Z} \frac{\epsilon}{|Y||Z|} = \frac{\epsilon}{|Y|}, \]

(14)

\[ \left| \frac{1}{N} N(z|z) - p_{Z}(z) \right| = \left| \frac{1}{N} \sum_{y \in Y} N(y,z|y,z) - \sum_{y \in Y} p_{Y,Z}(y,z) \right| \leq \sum_{y \in Y} \left| \frac{1}{N} N(y,z|y,z) - p_{Y,Z}(y,z) \right| \leq \sum_{y \in Y} \frac{\epsilon}{|Y||Z|} = \frac{\epsilon}{|Z|} \]

(15)

Note that (13) are the inequalities defining \( T^N_e(Y) \) and \( T^N_e(Z) \) respectively. Hence we can take the definition of \( T^N_e(Y,Z) \) as
\[ T^N_e(Y,Z) = \left\{ (y,z) \in Y^N \times Z^N \left| \frac{1}{N} N(y,z|y,z) - p_{Y,Z}(y,z) \right| \leq \frac{\epsilon}{|Y||Z|} \right\} \]

(16)

and have the desired property that \( (y,z) \in T^N_e(Y,Z) \) implies \( y \in T^N_e(Y) \) and \( z \in T^N_e(Z) \).

Furthermore, we can directly apply the results we have about the set of typical sequences \( T^N_e(X) \) to the set of jointly strongly typical sequences simply by taking the random variable \( X \) to be the collection of random variables under consideration. For instance,

\footnote{Note that if we were dealing with the set \( A^N_e(Y,Z) \) we would have needed to introduce additional inequalities to the set \( A^N_e(X) \) when trying to view it as \( X = (Y,Z) \), but this is not necessary in the case of \( T^N_e \) due to the direct implication of (12) to (13).}
• \((x, y) \in T_e^N(X, Y) \Rightarrow 2^{−N(H(X,Y)+\epsilon c)} \leq p_{X,Y}(x, y) \leq 2^{−N(H(X,Y)−\epsilon c)}\)

• \(|T_e^N(X, Y)| \leq 2^{2N(H(X,Y)+\epsilon c)}\)

• \(|T_e^N(X, Y)| \geq (1−\epsilon)2^{2N(H(X,Y)−\epsilon c)}\) for \(N\) large enough

• \((x, y) \in T_e^N(X, Y) \Rightarrow \frac{1}{N} \sum_{n=1}^{N} f(x_n, y_n) − E[f(X,Y)] \leq \epsilon f_{\text{max}}\)

From this definition and fact, it is clear how to define the set of typical sequences \(T_e^N\) for an arbitrary number of random variables \((X_1, x_2, \ldots, X_k) \sim p_{X_1, X_2, \ldots, X_k}\). For instance, the set of typical sequences for \((X_n, Y_n, Z_n) \sim p_{X,Y,Z}\) is

\[
T_e^N(X, Y, Z) := \left\{(x, y, z) \in X^N \times Y^N \times Z^N \mid \frac{1}{N} \sum_{n=1}^{N} (x_n, y_n, z_n) - p_{X,Y,Z}(x, y, z) \leq \frac{\epsilon}{|X||Y||Z|}, \forall x, y, z\right\}
\]

(17)

### 2.2 Independent Sequences and the Jointly Typical Set

Also as we studied with \(A_e^N(X, Y)\), the probability of two independent sequences \(\tilde{X}_n, \tilde{Y}_n \sim p_X p_Y\) with the same marginal distributions \(p_X, p_Y\) as a joint distribution \(p_{X,Y}\) wind up in the set of typical sequences \(T_e^N(X, Y)\) with a probability dying exponentially at a rate based on the mutual information. Indeed,

\[
\mathbb{P}[(\tilde{X}_N, \tilde{Y}_N) \in T_e^N(X, Y)] = \sum_{(x, y) \in T_e^N(X, Y)} p_{\tilde{X}, \tilde{Y}}(x, y) = \sum_{(x, y) \in T_e^N(X, Y)} p_X(x) p_Y(y)
\]

\[
\leq |T_e^N(X, Y)|2^{-N(H(X)+\epsilon)} \leq 2^{-N(H(X)+H(Y)−\epsilon)} \leq 2^{-N(I(X;Y)+\epsilon)}
\]

\[
\geq (1−\epsilon)2^{-N(I(X;Y)+\epsilon)}
\]

(18)

More generally, assume \(\tilde{X}_N\) and \(\tilde{Y}_N\) are two independent sequences of random variables, and \(\tilde{Y}_N \sim p_Y\) is the marginal distribution for some joint distribution \(p_{XY}\) for a pair of random variables \((X, Y)\). Then, for any distribution \(p_X^N\) with \(\mathbb{P}[\tilde{X}_N \in T_e^N(X)] \neq 0\), we have

\[
\mathbb{P}[(\tilde{X}_N, \tilde{Y}_N) \in T_e^N(X, Y) \mid \tilde{X}_N \in T_e^N(X)] \leq 2^{-N(I(X;Y)−\epsilon)}
\]

(20)

and, for \(N\) large enough

\[
\mathbb{P}[(\tilde{X}_N, \tilde{Y}_N) \in T_e^N(X, Y) \mid \tilde{X}_N \in T_e^N(X)] \geq (1−\epsilon)2^{-N(I(X;Y)+\epsilon)}.
\]

(21)

Indeed, these inequalities follow from

\[
\mathbb{P}[(\tilde{X}_N, \tilde{Y}_N) \in T_e^N(X, Y) \mid \tilde{X}_N \in T_e^N(X)] = \sum_{x \in T_e^N(X)} \mathbb{P}[(\tilde{X}_N, \tilde{Y}_N) \in T_e^N(X, Y), \tilde{X}_N = x \mid \tilde{X}_N \in T_e^N(X)]
\]

(22)

\[
= \sum_{x \in T_e^N(X)} \mathbb{P}[(x, \tilde{Y}_N) \in T_e^N(X, Y) \mid \tilde{X}_N = x] \mathbb{P}[(\tilde{X}_N = x, \tilde{X}_N \in T_e^N(X)]
\]

(23)

and for \(x \in T_e^N(X)\)

\[
\mathbb{P}[(x, \tilde{Y}_N) \in T_e^N(X, Y)] = \sum_{y \mid (x, y) \in T_e^N(X, Y)} p_{Y|X}(y) \leq \frac{1}{|T_e^N(X, Y)|} \leq 2^{-N(H(Y)−\epsilon)} \leq 2^{-N(I(X;Y)−\epsilon')}
\]

(25)
and, for \( N \) large enough
\[ P \left[ (x, \tilde{Y}^N) \in T^N_e(X, Y) \right] = \sum_{y(x, y) \in T^N_e(X, Y)} p_{Y^N}(y) \leq \left| \{ y | (x, y) \in T^N_e(X, Y) \} \right| 2^{-N(H(Y) + \epsilon c)} \geq (1 - \epsilon) 2^{-N(H(Y) + \epsilon c)} \] (26)

Substituting these back bounds back into (24), and observing that \( \sum_{x \in T^N_e(X)} P \left[ \hat{X}^N = x | \hat{X}^N \in T^N_e(X) \right] = 1 \), we have proven (20) and (21).

Note that this proof made use of bounds on \( \left| \{ y | (x, y) \in T^N_e(X, Y) \} \right| \) which follow from
\[ 1 \geq P \left[ (x, Y^N) \in T^N_e(X, Y) | X^N = x \right] = \sum_{y(x, y) \in T^N_e(X, Y)} p_{Y^N | X^N}(y) \geq \sum_{y(x, y) \in T^N_e(X, Y)} 2^{-N(H(Y | X) + \epsilon c)} \] (27)

which shows (25), and the convergence with \( x \in T^N_e \) of
\[ \frac{1}{N} \sum_{n=1}^{N} 1_{x, y, z} \left( x_n, y_n, z_n \right) = \frac{1}{N} \sum_{n=1}^{N} 1_{y(x, y)} \sum_{n: x_n = x} 1_{y_n} \Rightarrow \lim_{N \to \infty} P \left[ (x, Y^N) | X^N = x \right] = 1 \] (28)

which implies that for \( N \) large enough
\[ 1 - \epsilon \leq P \left[ (x, Y^N) | X^N = x \right] = \sum_{y(x, y) \in T^N_e(X, Y)} p_{Y^N | X^N}(y) \leq \sum_{y(x, y) \in T^N_e(X, Y)} 2^{-N(H(Y | X) - \epsilon c)} \] (29)

proving (26).

### 2.3 The Markov Lemma

Suppose \( X \leftrightarrow Y \leftrightarrow Z \) form a Markov chain. In this special case of the joint distribution \( p_{X, Y, Z} = p_{X, Y} p_{Z | Y} \), checking pairwise strong typicality for \((x, y)\) and \((y, z)\) is enough to have strong typicality for all three \((x, y, z)\) as \( N \to \infty \).

Stated more precisely, if \( X \leftrightarrow Y \leftrightarrow Z \), and \( (x, y) \in T^N_e(X, Y) \), and we generate a sequence of random variables \( \tilde{Z}^N \) according to the distribution
\[ p_{\tilde{Z}^N}(z) = \prod_{n=1}^{N} p_{Z | Y}(z_n | y_n) \] (30)

Then the probability
\[ P[(x, y, \tilde{Z}^N) \in T^N_e(X, Y, Z)] \to 1 \text{ as } N \to \infty \] (31)

This follows from the fact that
\[ \frac{1}{N} \sum_{n=1}^{N} 1_{x, y, z} (x_n, y_n, z_n) = \frac{1}{N} \sum_{n=1}^{N} 1_{x, y, z} (x_n, y_n, \tilde{Z}_n) \] (32)

\[ = \frac{1}{N} \sum_{n=1}^{N} 1_{x, y, z} (x_n, y_n, \tilde{Z}_n) \frac{1}{N} \sum_{n: x_n = x} 1_{y_n} \] (33)

The term on the right involves IID random variables because for those \( n \) such that \( y_n = y, p_{\tilde{Z}_n}(z) = p_{Z | Y}(z | y) \). Furthermore, due to the joint typicality \((x, y) \in T^N_e(X, Y)\), we will have \( N(x, y | x, y) \to \infty \) as \( N \to \infty \), and hence we are taking a sample average of a number of IID terms going to infinity, so the WLLN implies
\[ \frac{1}{N} \sum_{n: x_n = x, y_n = y} 1_{z} (\tilde{Z}_n) \] (34)

When substituted into the expression (33), this convergence in probability then means that for all \( \epsilon_2 > 0 \)
\[ P \left[ \frac{1}{N} N(x, y, z | x, y, \tilde{Z}^N) - p_{X, Y}(x, y) p_{Z | Y}(z | y) \right] \leq \frac{\epsilon}{|X| |Y|} \epsilon_2 \] (35)

Selecting \( \epsilon_2 = \frac{1}{|Z|} \), then, we have shown that \( P[(x, y, \tilde{Z}^N) \in T^N_e(X, Y, Z)] \to 1 \text{ as } N \to \infty \).
3 Rate Distortion with Side Information (Wyner Ziv)

A rate distortion pair \((R, D)\) is said to be achievable if there exists a block length \(N\) and an encoder mapping \(f_N : \mathcal{X}^N \to \{1, \ldots, 2^{NR}\}\) and decoder mapping \(g_N : \{1, \ldots, 2^{NR}\} \times \mathcal{Y}^N \to \mathcal{X}^N\) such that

\[
\mathbb{E}[d(X^N, g_N(f_N(X^N), Y^N))] = \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}[d(X_n, g^N_n(f_N(X^N), Y^N))] \leq D
\]

(36)

The closure of the set of achievable \((R, D)\) pairs is called the rate distortion region \(\mathcal{R}\). The rate distortion function is infimum of all rates necessary to approach a distortion \(D\), that is

\[
R_{SI-D}(D) = \inf \{r \mid (r, d) \in \mathcal{R}, d \leq D\}
\]

(37)

\[
R_{SI-D}(D) = \min_{p(U|X), g} \left\{ \min_{U \leftrightarrow X \leftrightarrow Y} I(X;U) - I(Y;U) \right\} = \min_{p(U|X), g} \left\{ \min_{U \leftrightarrow X \leftrightarrow Y} I(X;U|Y) \right\}
\]

(38)

### 3.1 Achievability

We will in fact prove the existence of codes of rate \(R > R_{SI-D}(D)\) not only with expected distortion arbitrarily close to \(D\), as requested by (36), but also with the property that

\[
\text{for any } \delta > 0, \quad \mathbb{P} \left[ \frac{1}{N} \sum_{n=1}^{N} d(X_n, \hat{X}_n) \geq D + \delta \right] \to 0 \quad \text{as } N \to \infty.
\]

(39)

- **Codebook Generation**: Generate a \(2^{NR_1} \times N\) matrix with elements IID according to the distribution \(p_U(u) = \sum_{x \in X} p_{U|X}(u|x)p_X(x)\). Let the \(i\)th row of this matrix be denoted \(U^N(i)\). Share this matrix with the encoder and decoder. Assign each codeword index \(i \in \{1, \ldots, 2^{NR_1}\}\) to one of \(2^{NR_2}\) bins \(B(j)\), \(j \in \{1, \ldots, 2^{NR_2}\}\) by selecting an \(j\) uniformly at random from \(\{1, \ldots, 2^{NR_2}\}\) for each codeword index \(i\).

- **Encoder**: Select any \(i \in \{1, \ldots, 2^{NR_1}\}\) such that \((X^N, U^N(i)) \in \mathcal{T}_r^N(X, U)\). If there is no such \(i\), select \(i = 1\). Send the bin index \(j\) to which this \(i\) is assigned, i.e. send \(j\) such that \(i \in B(j)\).

- **Decoder**: Among those \(i \in B(j)\), find one such that \((Y^N, U^N(i)) \in \mathcal{T}_c^N(Y, U)\). If none can be found or there are several, select \(i = 1\). Output \(Z^N = g(Y^N, U^N(i))\).

We bound the distortion provided by this scheme by considering the following events

- \(E_1\): There is no \(i \in \{1, \ldots, 2^{NR_1}\}\) such that \((X^N, U^N(i)) \in \mathcal{T}_r^N(X, U)\).

- \(E_2\): There is some \(i \in \{1, \ldots, 2^{NR_1}\}\) such that \((X^N, U^N(i)) \in \mathcal{T}_r^N(X, U)\), but for this \(i\), \((X^N, Y^N, U^N(i))\) is not in \(\mathcal{T}_c^N(X, Y, U)\).

- \(E_3\): There is some \(i \in \{1, \ldots, 2^{NR_1}\}\) such that \((X^N, Y^N, U^N(i)) \in \mathcal{T}_c^N(X, Y, U)\) and \(i \in B(j)\), but there is some other \(i' \in B(j)\) such that \((Y^N, U^N(i')) \in \mathcal{T}_c^N(Y, U)\).
Let $\mathcal{E} = \mathcal{E}_1 \cup \mathcal{E}_2 \cup \mathcal{E}_3$. If $\mathcal{E}_c$ occurs, then there is an $i \in \{1, \ldots, 2^{NR_1}\}$ such that $(X^N, Y^N, U^N(i)) \in T^N(X, Y, U)$, $i$ is assigned to bin $B(j)$, and there is no other $i' \in B(j)$. Hence, given $\mathcal{E}_c$ has occurred, by the expectation property of the strongly typical set, the distortion is bounded as

$$\left| \frac{1}{N} \sum_{n=1}^{N} d(X_n, g(U_n(i), Y_n)) - \mathbb{E}[d(X, g(U, Y))] \right| \leq \epsilon d_{max}$$

(40)

which can be made arbitrary close to the desired distortion $\mathbb{E}[d(X, g(U, Y))] \leq D$ by selecting $\epsilon$ sufficiently small.

We now proceed to show that the rates $R_1, R_2$ can be selected such that the probability $\mathbb{P}[\mathcal{E}]$ can be made arbitrarily small. The union bound states that $\mathbb{P}[\mathcal{E}] \leq \mathbb{P}[\mathcal{E}_1] + \mathbb{P}[\mathcal{E}_2] + \mathbb{P}[\mathcal{E}_3]$. We proceed to show that each of these probabilities can be made arbitrarily small. For $N$ large enough

$$\mathbb{P}[\mathcal{E}_1] = \mathbb{P}[ (X^N, U^N(i)) \notin T^N_\epsilon(X, U) \ \forall i \in \{1, \ldots, 2^{NR_1}\} ] = \prod_{i=1}^{2^{NR_1}} \mathbb{P}[ (X^N, U^N(i)) \notin T^N_\epsilon(X, U) ]$$

(41)

$$= \prod_{i=1}^{2^{NR_1}} (1 - \mathbb{P}[ (X^N, U^N(i)) \in T^N_\epsilon(X, U) ]) = (1 - \mathbb{P}[ (X^N, U^N(i)) \in T^N_\epsilon(X, U) ])^{2^{NR_1}}$$

(42)

$$\leq \left( 1 - (1 - \epsilon)2^{-N(I(X;U) + \epsilon c_1 c_2 c_3)} \right)^{2^{NR_1}} \leq 1 - (1 - \epsilon) + \exp \left( -2^{NR_1} 2^{-N(I(X;U) + \epsilon c_1 c_2 c_3)} \right)$$

(43)

$$\leq \epsilon + \exp \left( -2^{NR_1} - I(X;U) - \epsilon c_1 c_2 c_3 \right)$$

(44)

Hence we observe that we can select $\epsilon$ to make $\mathbb{P}[\mathcal{E}_1]$ arbitrarily small when $R_1 > I(X;U)$ and $N \to \infty$.

Next consider the event $\mathcal{E}_2$. This is precisely the probability

$$\mathbb{P}[\mathcal{E}_2] = \mathbb{P} \left[ \exists i \text{ such that } (X^N, U^N(i)) \in \tau_\epsilon(X, U) \text{ but } (X^N, Y^N, U^N(i)) \notin \tau_\epsilon(X, Y, U) \right]$$

(45)

$$= \mathbb{P}[ \mathcal{E}_1^c \cap T^N_\epsilon(X, Y, U) ]$$

(46)

$$= \mathbb{P}[ (X^N, Y^N, U^N(i)) \notin T^N_\epsilon(X, Y, U) ]$$

(47)

$$\leq \mathbb{P}[ (X^N, Y^N, U^N(i)) \notin T^N_\epsilon(X, Y, U) \mid (X^N, U^N(i)) \in T^N_\epsilon(X, U) ]$$

(48)

The Markov lemma (35) showed that this probability $\to 0$ as $N \to \infty$.

Finally we pass to the third error event

$$\mathbb{P}[\mathcal{E}_3] = \mathbb{P} \left[ \exists i, i' \text{ such that } (X^N, Y^N, U^N(i)) \in T^N_\epsilon(X, Y, U) \text{ and } i, i' \in B(j), i \neq i' \text{ and } (Y^N, U^N(i')) \in T^N_\epsilon(Y, U) \right]$$

(49)

$$= \mathbb{P}[ (Y^N, U^N(i')) \in T^N_\epsilon(Y, U) \text{ some } i' \in B(j), i' \neq i \mid \mathcal{E}_1^c \cap \mathcal{E}_2^c ] \mathbb{P}[ \mathcal{E}_1^c \cap \mathcal{E}_2^c ]$$

(50)

$$\leq \mathbb{P}[ \mathcal{E}_1^c \cap \mathcal{E}_2^c ] \left( \sum_{i' = 1, i' \neq i}^{2^{2NR_1}} \mathbb{P}[ (Y^N, U^N(i')) \in T^N_\epsilon(Y, U) ] \mathbb{P}[ (Y^N, U^N(i')) \in T^N_\epsilon(Y, U) ] \right)$$

(51)

$$\leq 2^{2NR_1} \sum_{i' = 1, i' \neq i}^{2^{2NR_1}} \mathbb{P}[ i' \in B(j)] \leq 2^{2NR_1} 2^{-N(I(Y;U) - \epsilon d)}$$

(52)

(53)

This will go to 0 as $N \to \infty$ provided that $R_1 - R_2 - I(Y;U) < 0$. Putting this together with the requirement for $\mathbb{P}[\mathcal{E}_1], R_1 > I(X;U)$, we find that if $R_2 > I(X;U) - I(Y;U)$ for any $\delta > 0$ we can select an $\epsilon$ such that the probability

$$\mathbb{P} \left[ \frac{1}{N} \sum_{n=1}^{N} d(X_n, g(U_n(i), Y_n)) \geq D + \delta \right] \to 0 \ \text{as} \ N \to \infty$$

(54)

### 3.2 Convexity of $R_{SI-D}(D)$

The converse proof makes use of the convexity and monotone non-increasing nature of $R_{SI-D}(D)$ as defined in (38), which we now show. Indeed, consider random variables $U^1$ and $U^2$ with conditional distributions $p_{U^1|X}$ and $p_{U^2|X}$ with reconstructions
functions \( g^1 \) and \( g^2 \) attaining the minimum in (38) for the distortions \( D_1 \) and \( D_2 \), respectively. Define the random variable \( Q \in \{1, 2\} \), \( \mathbb{P}[Q = 1] = \lambda \), independent of everything else, and let \( U = (Q, U_Q) \), and define the reconstruction function \( g(U, Y) = g^Q(U_Q, Y) \). This random variable and reconstruction function achieves the distortion
\[
\mathbb{E}[d(X_n, g(U_Q, Y))] = \mathbb{E}[d(X_n, g^Q(U_Q, Y))|Q = 1]\mathbb{P}[Q = 1] + \mathbb{E}[d(X_n, g^Q(U_Q, Y))|Q = 2]\mathbb{P}[Q = 2] = \lambda D_1 + (1 - \lambda)D_2
\]
Hence, due to the minimization definition (38)
\[
R_{SI-D}(\lambda D_1 + (1 - \lambda)D_2) \leq I(X; U|Y) = H(X|Y) - H(X|U, Y) = H(X|Y) - H(X|Q, U_Q, Y) \quad (55)
\]
\[
= H(X|Y) - \mathbb{P}[Q = 1]H(X|U_1, Y) - \mathbb{P}[Q = 2]H(X|U_2, Y) \quad (56)
\]
\[
= \lambda(H(X|Y) - H(X|U_1, Y)) + (1 - \lambda)(H(X|Y) - H(X|U_2, Y) \quad (57)
\]
\[
= \lambda R_{SI-D}(D_1) + (1 - \lambda)R_{SI-D}(D_2) \quad (58)
\]
which establishes convexity. The mono-tone non-increasing nature is clear from the nested nature of the constraint set to optimize over for \( D \) and \( D' \) for \( D' > D \).

3.3 Converse

Suppose that there exists an encoder \( f_N : \mathcal{X}^N \rightarrow \{1, \ldots, 2^{NR}\} \) and a decoder \( g^N_n : \{1, \ldots, 2^{NR}\} \times \mathcal{Y}^N \rightarrow \mathcal{X}, n \in \{1, \ldots, N\} \) with
\[
\frac{1}{N} \sum_{n=1}^{N} \mathbb{E}[d(X_n, g_n^N(f_N(X_n), Y^N))] \leq D \quad (59)
\]
\[
NR \geq H(J) \geq H(J|Y^N) \geq H(J|X^N, Y^N) \quad (60)
\]
\[
= I(X; Y|J) = H(X|Y) - H(X|Y, J) \quad (61)
\]
\[
= \sum_{n=1}^{N} H(X_n|Y_n) - \sum_{n=1}^{N} H(X_n|X_{n-1}, Y_{n+1}, J) \geq \sum_{n=1}^{N} H(X_n|Y_n) - \sum_{n=1}^{N} H(X_n|Y_n, Y_{n+1}, J) \quad (62)
\]
\[
= \sum_{n=1}^{N} H(X_n|Y_n) - H(X_n|Y_n, U_n) = \sum_{n=1}^{N} I(X_n; U_n|Y_n) \quad (63)
\]
where we introduced the random variable \( U_n = (J, Y_{n-1}, Y_{n+1}) \). Observe that under this definition \( \hat{Z}_n = g^N_n(J, Y^N) = g^N_n(U_n, Y_n) \), and furthermore that \( U_n \leftrightarrow X_n \leftrightarrow Y_n \), defining \( D_n = \mathbb{E}[d(X_n, g_n^N(U_n, Y_n))] \) we have
\[
R \geq \frac{1}{N} \sum_{n=1}^{N} I(X_n; U_n|Y_n) \geq \frac{1}{N} \sum_{n=1}^{N} R_{SI-D}(D_n) \geq R_{SI-D} \left( \frac{1}{N} \sum_{n=1}^{N} D_n \right) \geq R_{SI-D}(D). \quad (64)
\]

3.4 Indirect Rate Distortion with Side Information

![Diagram](image)

Figure 2: Indirect variant of the rate distortion with side information problem.

In order to understand the result just proven in its full level of generality, it is instructive to observe that no aspect of the proof required that what the decoder is estimating in a lossy manner is the source being directly observed at the encoder! Naturally, there are a number of problems where the decoder is not trying produce estimates of the sequence \( X^N \), but is instead attempting to produce estimates of some quantity \( Z^N = (Z_1, \ldots, Z_N) \) statistically related to the observations at the encoder and the side information. For instance, \( X^N \) and \( Y^N \) may be noisy versions of some sequence \( Z^N \) that the decoder
would like to learn. If \((X_n, Y_n, Z_n) \sim iid p_{X,Y,Z}\), one can use a near identical achievability construction, and an identical converse to show that the rate distortion function is

\[
R_{SI-D}(D) = \min_{\min \mathbf{U}} \left| \begin{array}{c} p(U|X), g \\ U \leftrightarrow X \leftrightarrow (Y, Z) \end{array} \right| \mathbb{E}[d(Z, g(U, Y))] \leq D.
\]

The key point to realize is that given that we selected a codeword \(U^N(i)\) that is jointly typical with \(X^N\), i.e. \((X^N, U(i)) \in T^N(X, U)\), the Markov lemma applied to the Markov chain \(U \leftrightarrow X \leftrightarrow (Y, Z)\) implies that this will, with probability going to 1 as \(N \to \infty\), also satisfy \((U^N(i), X^N, Y^N, Z^N) \in T^N(U, X, Y, Z)\). This then implies, via the expectation property of the typical set \((7)\), that the distortion \(\frac{1}{N} \sum_{n=1}^N d(Z_n, g(U_n, Y_n))\) will be close to its expected value as required. The converse is identical, with the only exception being to replace \(\mathbb{E}[d(X_n, g^n_N(f_N(X^N), Y^N))]\) with \(\mathbb{E}[d(Z_n, g^n_N(f_N(X^N), Y^N))]\).

### 4 Rate Distortion with Multiple Degraded Side Informations

![Diagram](image_url)

A seminal paper by Heegard and Berger *Rate Distortion when Side Information May be Absent* (Trans. IT 6-1985) showed that the techniques discussed in the previous problem to find the rate distortion function when a common message must be capable of being decoded in the presence of multiple degraded side informations. Here, degraded means that the side informations form a Markov chain \(X \leftrightarrow Y_M \leftrightarrow Y_{M-1} \leftrightarrow \cdots \leftrightarrow Y_1\). The intuition is that decoders with better side information can determine more information about the source and achieve a lower distortion. The rate distortion function for this problem is

\[
R_{HB}(D) = \min_{\sum_{m=1}^M I(X; U_m|Y_m, U_{1:(m-1)})} \left| \begin{array}{c} p(U_{1:M}|X), g_m \\ U_{1:M} \leftrightarrow X \leftrightarrow Y_1 \leftrightarrow Y_2 \leftrightarrow \cdots \leftrightarrow Y_M \end{array} \right| \mathbb{E}[d_m(X, g_m(U_{1:m}, Y_m))] \leq D_m.
\]

#### 4.1 Converse

Start with the distortion constraints. We can rewrite them as

\[
D_m \geq \frac{1}{N} \sum_{n=1}^N \min g^n_N \mathbb{E}[d_m(X_n, g^n_{m,N}(J, Y^N_m))] = \frac{1}{N} \sum_{n=1}^N \min g^n_N \mathbb{E}[d_m(X_n, g^n_{m,N}(U_{1:m}, Y_m))] \quad (67)
\]

where we have introduced the relabeled variables \(U_{m,n} = Y_{m\setminus n}, \ m \in \{2, \ldots, M\}\) and \(U_{1,n} = (J, Y_{1\setminus n})\), with the notation \(Y_{m\setminus n} = (Y_{m,1}, \ldots, Y_{m,n-1}, Y_{m,n+1}, \ldots, Y_{m,N})\). Observe that these random variables obey the Markov Chain \((U_{1,n}, \ldots, U_{M,n}) \leftrightarrow X_n \leftrightarrow Y_{M,n} \leftrightarrow Y_{M-1,n} \leftrightarrow \cdots \leftrightarrow Y_{1,n}\). While it appears that the term on the far right in (67) has access to more information to form estimates of \(X_n\) than is available to decoder \(m\), this is not the case. This is because of the Markov chain \(X \leftrightarrow Y_M \leftrightarrow Y_{M-1} \leftrightarrow \cdots \leftrightarrow Y_1\); we could consider decoder \(m\) as having *all* of the side information with quality worse than...
it, as it will be no use in estimating $X^N$ given the conditional independence of $X^N$ and $Y^N_{m'}$ given $Y^N_m$ with $m' < m$. The convexity and monotone non-increasing nature of $R_{HB}(D)$, $D = (D_1, \ldots, D_M)$, shows

$$R_{HB}(D) \leq R_{HB} \left( \frac{1}{N} \sum_{n=1}^{N} D_n \right) \leq \frac{1}{N} \sum_{n=1}^{N} R_{HB}(D_n) \leq \frac{1}{N} \sum_{n=1}^{N} \sum_{m=1}^{M} I(X_n; U_{m,n}) Y_{m,n}, U_{1:(m-1),n}$$

(68)

where $D_n = (D_{1,n}, \ldots, D_{M,n})$, hence what remains to be proven is that this is a lower bound on the rate $R$.

$$\frac{1}{N} \sum_{n=1}^{N} \sum_{m=1}^{M} I(X_n; U_{m,n}) Y_{m,n}, U_{1:(m-1),n} = \frac{1}{N} \sum_{n=1}^{N} \sum_{m=1}^{M} H(U_{m,n}|Y_{m,n}, U_{1:(m-1),n}) - H(U_{m,n}|Y_{m,n}, U_{1:(m-1),n}, X_n)$$

$$= \frac{1}{N} \sum_{n=1}^{N} \sum_{m=1}^{M} H(U_{m,n}|Y_{m,n}, U_{1:(m-1),n}) - H(U_{m,n}|U_{1:(m-1),n}, X_n)$$

(69)

$$= \frac{1}{N} \sum_{n=1}^{N} \sum_{m=1}^{M} H(U_{m,n}|Y_{m,n}, U_{1:(m-1),n}) - H(U_{1:M,n} X_n)$$

(70)

$$\leq \frac{1}{N} \sum_{n=1}^{N} \sum_{m=1}^{M} H(U_{m,n}|U_{1:(m-1),n}) - H(U_{1:M,n}|X_n)$$

(71)

Here (69) reflected the Markov condition, (70) the chain rule, and (71) that conditioning reduces entropy. Moving on

$$\frac{1}{N} \sum_{n=1}^{N} \sum_{m=1}^{M} H(U_{m,n}|U_{1:(m-1),n}) - H(U_{1:M,n}|X_n) = \frac{1}{N} \sum_{n=1}^{N} H(U_{1:M,n} X_n) - H(U_{1:M,n}|X_n)$$

$$= \frac{1}{N} \sum_{n=1}^{N} H(Y_{1:M,n}, J) - H(Y_{1:M,n}, J|X_n) = \frac{1}{N} \sum_{n=1}^{N} H(Y_{1:M,n}) + H(J|Y_{1:M,n}) - H(Y_{1:M,n}|X_n) - H(J|Y_{1:M,n}, X_n)$$

$$= \frac{1}{N} \sum_{n=1}^{N} H(Y_{1:M,n}) + H(J|Y_{1:M,n}) - H(Y_{1:M,n}) - H(J|Y_{1:M,n}, X_n) = \frac{1}{N} \sum_{n=1}^{N} H(J|Y_{1:M,n}) - H(J|Y_{1:M,n}, X_n)$$

$$\leq \frac{1}{N} \sum_{n=1}^{N} H(J) \leq R$$

so $R \geq R_{SI-D}(D)$ as we needed to prove.

### 4.2 Achievability

- **Codebook Generation:** For each $m \in \{1, \ldots, M\}$, generate a $2^{N R_m} \times N$ dimensional matrix $C_m$ with elements IID according to $p_{U_m}(u_m) = \sum x \in X, u_m \in U_m$ p_{U_m}(x)p_X(x)$. Denote the $i$th row of this matrix by $U_m(i)$ for each $i \in \{1, \ldots, 2^{N R_m}\}$. Assign each $i \in \{1, \ldots, 2^{N R_m}\}$ to one of $2^{NR_m}$ bins $B_m(j)$, $j \in \{1, \ldots, 2^{NR_m}\}$ by selecting a $j \in \{1, \ldots, 2^{NR_m}\}$ uniformly at random. Distributed the matrices and the bins to the encoders. For each $m \in \{1, \ldots, M\}$, give the $m$th decoder the matrices $C_1, \ldots, C_m$ and the bins $B_1(\cdot), \ldots, B_m(\cdot)$.

- **Encoder:** Find $i = (i_1, \ldots, i_M)$, $i_m \in \{1, \ldots, 2^{N R_m}\}$, $m \in \{1, \ldots, M\}$ such that $(U_m(i_1), \ldots, U_m(i_M), X^N) \in T_{\epsilon}(U_1, \ldots, U_M, X)$. If no such $i$ exists, select $i = 1$. Send the bin indices $J = (j_1, \ldots, j_M)$ such that $i_m \in B_m(j_m)$ for each $m \in \{1, \ldots, M\}$ to all the decoders.

- **Decoders:** Decoder $m$ sequentially considers consecutive bins. First, it looks for a codeword with index $\hat{i}_{m,1} \in B_1(j)$ such that $(U_1^{N}(\hat{i}_{m,1}), Y_{m}^{N}) \in T_{\epsilon}(U_1, Y_m)$. Next it looks for a codeword with index $\hat{i}_{m,2} \in B_2(j)$ such that $U_2^{N}(\hat{i}_{m,1}), Y_2^{N} \in T_{\epsilon}(U_1, U_2, Y_m)$. The decoding continues consecutively, with $\hat{i}_{m,k} \in B_k(j_k)$ found such that $U_k^{N}(\hat{i}_{m,1}), \ldots, U_k^{N}(\hat{i}_{m,k-1}), Y_{m}^{N} \in T_{\epsilon}(U_1, \ldots, U_k, Y)$ until $k = m$. (Note it attempts only to decode the first $m$ descriptions.) If at any step $k \in \{1, \ldots, m\}$ in the sequence there is either no such $i_{m,k} \in B_k(j_k)$ or there are multiple $i_{m,k}$, then $\hat{i}_{m,k'} = 1$ is selected for all $k' \in \{k, k+1, \ldots, m\}$.

The distortion provided by this scheme can be bounded using the following events.
- $E_m$: There is an $i$ such that $(U_m^N(i), X^N) \in T_c^N(U_1, \ldots, U_m, X)$, but for this $i$, $(U_m^N(i), X^N, Y_{1:M}) \notin T_c^N(U_1, X, Y_{1:M}).$

- $E_{m,k}$: There is an $i$ such that $(U_m^N(i_1), \ldots, U_m^N(i_m), X^N) \in T_c^N(U_1, \ldots, U_m, X)$, and for this $i$, $(U_m^N(i), X^N, Y_{1:M}) \in T_c^N(U_1, X, Y_{1:M})$, and $i_{m,1} = i_1, \ldots, i_{m,k-1} = i_{k-1}$, but there is some $i_{m,k} \neq i_k$ such that $i_k \in B_k(j_k)$ and $(U_{m,k}^N(i_1, i_2, \ldots, i_m), Y_{m,k}^N) \in T_c^N(U_{1:k-1}, U_k, Y_m).$

Note that if none of these events occur, then $(U^N(i), X^N, Y_{1:M}) \in T_c^N(U_1, X, Y_{1:M})$ and $i_1, \ldots, i_m$ are reproduced faithfully at decoder $m$. This implies that the distortion can be made arbitrarily close to $D$ by selection of $\epsilon$ if none of these events $E_m^1, E_2^2, E_{m,k}, k \in \{1, \ldots, m\}, m \in \{1, \ldots, M\}$ occur. Hence we set out to show that each of these events can be made arbitrarily small through selection of $N$ and $\epsilon$.

Indeed, for $N$ large enough

$$P[E_m^1] \leq \prod_{i=1}^{2^N R_m} \mathbb{P}[(U_{m-1}(i_{m-1}), U_m^N(i_m), X^N) \notin T_c^N(U_{1:m-1}, U_m, X) \mid (U_{m-1}^N(i_{m-1}), X^N) \in T_c^N(U_{1:m-1}, X)]$$

$$\leq (1 - (1 - \epsilon)2^{-N(I(U_m; (X, U_{1:m-1})))})^{2^N R_m} \leq \epsilon + \exp(-2N(R_m - I(U_m; (X, U_{1:m-1}))) - \epsilon)$$

(72)

So that $P[E_m^1] \rightarrow 0$ as $N \rightarrow \infty$ if we select $R_m > I(U_m; (X, U_{1:m-1})).$

Next, because $U_{1:M} \leftrightarrow X \leftrightarrow Y_{1:M}$ the Markov lemma shows that $P[E_2^2] \rightarrow 0$ as $N \rightarrow \infty$.

Finally, we consider the last event

$$P[E_{m,k}] \leq \mathbb{P}[(U_k^N(i_k), U_{m,k}^N(i_{m,k-1}), Y_m^N) \in T_c^N(U_k, U_{1:k-1}, Y_m), i_k \neq i_k \in B_k(j_k) \mid (U_{1:M}^N(i), X^N, Y_{1:M}^N) \in T_c^N(U_1, X, Y_{1:M})]$$

as the binning is done independently of the codeword selection

$$P[E_{m,k}] \leq \sum_{i_k=1, i_k \neq i_k} \mathbb{P}[i_k \in B_k(j_k)] \mathbb{P}[(U_k^N(i_k), U_{m,k}^N(i_{m,k-1}, Y_m^N) \in T_c^N(U_k, U_{1:k-1}, Y_m) \mid (U_{1:M}^N(i), X^N, Y_{1:M}^N) \in T_c^N(U_1, X, Y_{1:M})]$$

$$\leq 2^{-N R_k} 2^{-N(I(U_k; (Y_m, U_{1:k-1}))) - \epsilon}$$

(73)

where the last inequality follows from applying (20). Hence $P[E_{m,k}] \rightarrow 0$ as $N \rightarrow \infty$ provided that $R_k > R_k - I(U_k; (Y_m, U_{1:k-1})).$

Plugging in the bound we needed for $R_k$, we need

$$R_k > I(U_k; (X, U_{1:k-1})) - I(U_k; (Y_m, U_{1:k-1})) = H(U_k | Y_m, U_{1:k-1}) - H(U_k | X, U_{1:k-1})$$

for all $k \leq m$, $m \in \{1, \ldots, M\}$

(74)

Due to the Markov chain $U_{1:M} \leftrightarrow X \leftrightarrow Y_{1:M}$ and the Markov chain $X \leftrightarrow Y_m \leftrightarrow Y_{m-1} \leftrightarrow \cdots \leftrightarrow Y_1$, for $k \leq m$

$$H(U_{1:k} | Y_k) \geq H(U_{1:k} | Y_m Y_k) = H(U_{1:k} | Y_m)$$

(75)

Hence, since the decoder with side information $Y_k$ only decodes messages $U_{1:k}$, the tightest lower bound for $R_k$ is

$$R_k > H(U_k | Y_k, U_{1:k-1}) - H(U_k | X, U_{1:k-1}) = I(U_k; X | Y_k, U_{1:k-1})$$

(76)

If $R_k$ satisfies this equation, then $P[E_{m,k}] \rightarrow 0$ for all $m \leq k$. Summing these rates (76) gives the expression in (66).