EE276 Information Theory Lecture 16 - 02/29/2024

Lecture 16: Strong, Conditional, & Joint Typicality

Lecturer: Tsachy Weissman

In this lecture, we will continue developing tools that will be useful going forward, in particular in the context of lossy compression.[1] We will introduce the notions of **Strong**, **Conditional**, and **Joint Typicality**.

1 Notation

A quick recap of the notation:

1. **Random variables**: i.e. X
2. **Alphabet:** i.e. $\mathcal{X}$
3. **Specific values:** i.e. x
4. **Sequence of values:** i.e. x^n
5. **Set of all probability mass functions on alphabet $\mathcal{X}$:** $\mathcal{M}(\mathcal{X})$
6. **Empirical distribution of a sequence x^n:** $P_{x^n}(a) := \frac{N(a|x^n)}{n}$ [$N(a|x^n)$ is # of times symbol a appears in x^n]

2 Typicality

2.1 Strong Typicality

Definition 1. *A sequence $x^n \in \mathcal{X}^n$ is **strongly δ-typical** with respect to a probability mass function $P \in \mathcal{M}(\mathcal{X})$ if*

$$|P_{x^n}(a) - P(a)| \leq \delta \cdot P(a), \qquad \forall a \in \mathcal{X} \tag{1}$$

In words, a sequence is strongly δ-typical with respect to P if its empirical distribution is close to the probability mass function P. [δ is some fixed number, typically small.]

Definition 2. *The **strongly δ-typical set** [or simply strongly typical set] of p, $T_\delta(P)$, is defined as the set of all sequences that are strongly δ-typical with respect to P, i.e.*

$$T_\delta(P) = \{x^n : |P_{x^n}(A) - P(a)| \leq \delta \cdot P(a), \forall a \in \mathcal{X}\} \tag{2}$$

Recall: the *weakly* ϵ-typical set of an IID source P is defined as $A_\epsilon(P) := \{x^n : \left|-\frac{1}{n}\log P(x^n) - H(P)\right| \leq \epsilon\}$. **Note**: The condition for inclusion in the weakly ϵ-typical set is indeed weaker than the condition to be in the strongly δ-typical set. $-\frac{1}{n}\log P(x^n) = \frac{1}{n}\log\frac{1}{\prod_{i=1}^n P(x_i)} = \frac{1}{n}\sum_{i=1}^n \log\frac{1}{P(x_i)} = \frac{1}{n}\sum_{a\in\mathcal{X}} N(a|x^n)\log\frac{1}{P(a)} = \sum_{a\in\mathcal{X}} P_{x^n}(a)\log\frac{1}{P(a)}$. This is $\approx \sum_{a\in\mathcal{X}} P(a)\log\frac{1}{P(a)} = H(P)$ if $P_{x^n} \approx P$, i.e. if the empirical distribution induced by x^n is "close" to P, i.e. if the sequence is strongly typical. Thus, $P(x^n) \approx P \Rightarrow -\frac{1}{n}\log P(x^n) \approx H(P)$, i.e. strong typicality implies weak typicality. In the homework, we will show more precisely that

[1] *Optional Reading:* Chapter 2 in El Gamal and Kim, Network Information Theory.

$$T_\delta(P) \subseteq A_\epsilon(P)$$

for $\epsilon = \delta \cdot H(P)$.

Example: Here is an example of a sequence that is weakly typical but not strongly typical. Let P be the uniform distribution over $\mathcal{X}$, i.e. $P(a) = \frac{1}{|\mathcal{X}|}\ \forall a \in \mathcal{X}$. Then $P(x^n) = \frac{1}{|\mathcal{X}|^n} \Rightarrow -\frac{1}{n}\log p(x^n) = \log|\mathcal{X}| = H(P)\ \forall x^n \in \mathcal{X}^n$. Thus, $A_\epsilon(P) = \mathcal{X}^n$, while $T_\delta(P) = \{x^n : \left|P_{x^n}(a) - \frac{1}{|\mathcal{X}|}\right| \leq \frac{\delta}{|\mathcal{X}|}, \forall a \in \mathcal{X}\}$. In other words, the weakly typical set is the set of all sequences over $\mathcal{X}$, whereas the strongly typical set is the set of all sequences such that each symbol appears roughly the same number of times along the sequence.

We have already shown that the probability of a particular sequence being in $A_\epsilon(P)$ approaches 1 as $n \to \infty$. In the homework, we will investigate the probability of a particular sequence being in $T_\delta(P)$, i.e. $P(T_\delta(P))$. In fact, this also approaches 1 as $n \to \infty$.

$$\lim_{n\to\infty} P(T_\delta(P)) = 1$$

This is also a manifestation of the law of large numbers, which tells us that for every symbol a, the fraction of times that it appears in a sequence will approach its true probability under the source P, with probability close to 1. Finally, we will show that the size of the set of strongly δ-typical sequences $|T_\delta(P)|$ is roughly $2^{nH(P)}$; more precisely, that for all sufficiently large n:

$$2^{n[H(P)-\epsilon(\delta)]} \leq |T_\delta(P)| \leq 2^{n[H(P)+\epsilon(\delta)]} \tag{3}$$

where $\epsilon(\delta) \to 0$ as $\delta \to 0$. The lower bound follows from the previously shown fact that any set with size smaller than $2^{nH(P)}$ has vanishing probability. The upper bound simply follows from the fact that $T_\delta(P) \subseteq A_\epsilon(P)$.

2.2 Joint Typicality

In the following, we refer to the sequences $x^n = (x_1, x_2, \ldots, x_n)$, $x_i \in \mathcal{X}$ and $y^n = (y_1, y_2, \ldots, y_n)$, $y_i \in \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are finite alphabets.

Definition 3. *The **joint empirical distribution** of (x^n, y^n) is:*

$$P_{x^n,y^n}(x,y) = \frac{1}{n}N(x,y|x^n,y^n) \tag{4}$$

where $N(x,y|x^n,y^n) := \sum_{i=1}^{n} \mathbb{1}_{\{x_i=x, y_i=y\}}$

Definition 4. *(x^n, y^n) is **jointly δ-typical** with respect to $P \in \mathcal{M}(\mathcal{X}\times\mathcal{Y})$ if*

$$|P_{x^n,y^n}(x,y) - P(x,y)| \leq \delta \cdot P(x,y), \qquad \forall x \in \mathcal{X}, y \in \mathcal{Y} \tag{5}$$

where $N(x,y|x^n,y^n) := \sum_{i=1}^{n} \mathbb{1}_{\{x_i=x, y_i=y\}}$

Definition 5. *The **jointly δ-typical set** with respect to $P \in \mathcal{M}(\mathcal{X}\times\mathcal{Y})$ is*

$$T_\delta(P) = \{(x^n, y^n) : (x^n, y^n) \text{ is jointly } \delta\text{-typical with respect to } P\} \tag{6}$$

where $N(x,y|x^n,y^n) := \sum_{i=1}^{n} \mathbb{1}_{\{x_i=x, y_i=y\}}$

Observe that these definitions are just special cases of the definitions of the empirical distribution, strong δ-typicality, and the strongly δ-typical set, since a pair of a sequence in $\mathcal{X}$ and a sequence in $\mathcal{Y}$ is simply a sequence in the alphabet of pairs $\mathcal{X} \times \mathcal{Y}$.

Notation: For convenience, we will sometimes write $T_\delta(X)$ in place of $T_\delta(P)$,when $X \sim P$, or $T_\delta(X, Y)$ in place of $T_\delta(P)$ when $(X, Y) \sim P$.

In the homework, we will show that $\forall g : \mathcal{X} \to \mathbb{R}, x^n \in T_\delta(X)$,

$$(1-\delta)E[g(X)] \leq \frac{1}{n}\sum_{i=1}^{n} g(x_i) \leq (1+\delta)E[g(X)]$$

In other words, for strongly typical sequences, the average value of g computed on the components of the sequence is "close to" the expected value of $g(X)$. Observe that $\frac{1}{n}\sum\limits_{i=1}^{n} g(x_i) = \sum\limits_{a\in\mathcal{X}} P_{x^n}(a) \cdot g(a)$; the latter is the expectation of $g(X)$ when X is distributed according to the empirical distribution of P_{x^n}. But since $x^n \in T_\delta(x)$, P_{x^n} is close to the true PMF of X [i.e. P], which is why this expectation is close to the true expectation $E[g(X)]$. This property will be important for the rate distortion theorem where g will be replaced by the distortion function. In the homework, you will find cases where this does not hold for weak typicality.

2.3 Conditional Typicality

Definition 6. *Fix x^n. The **conditional δ-typical set** is*

$$T_\delta(Y|x^n) = \{y^n : (x^n, y^n) \in T_\delta(X, Y)\} \tag{7}$$

In other words, it is the set of all sequences y^n such that the pair (x^n, y^n) is jointly δ-typical.

Observe that if $x^n \notin T_\delta(X)$, then $T_\delta(Y|x^n) = \emptyset$, because for a sequence (x^n, y^n) to be jointly typical, each individual sequence must be typical with respect to P_X and P_Y, respectively (shown in homework).
In the homework, we will show that, assuming $x^n \in T_{\delta'}(X)$,

$$(1-\delta)2^{n[H(Y|X)-\epsilon(\delta)]} \leq |T_\delta(Y|x^n)| \leq 2^{n[H(Y|X)+\epsilon(\delta)]}$$

for all $0 < \delta' < \delta$ and n sufficiently large, where $\epsilon(\delta) = \delta \cdot H(Y|X)$.
In short, for a sequence x^n that is typical, the number of sequences y^n that are jointly typical with x^n is approximately $2^{nH(Y|X)}$. A starting point of the proof will be the "Conditional Typicality Lemma."

Lemma 7 (***Conditional Typicality Lemma***)**.** *For $0 < \delta' < \delta$, $x^n \in T_{\delta'}(X)$ and $Y^n \sim P(y^n|x^n) = \prod\limits_{i=1}^{n} P_{Y|X}(y_i|x_i)$, then*

$$\lim_{n\to\infty} P(Y^n \in T_\delta(Y|x^n)) = 1 \tag{8}$$

In other words, we fix an individual sequence x^n, and generate the sequence Y^n stochastically and independently according to the distribution conditioned on x^n, i.e. we generate $Y_i \sim P_{Y|X=x_i}$, [according to the joint probability mass function $P_{X,Y}$, which gives rise to the conditional probability mass function $P_{Y|X}$]. One can think of this in communication terminology: the sequence Y^n is generated is by taking the individual sequence x^n and passing it through the memoryless channel $P(Y|X)$. The probability that the sequence Y^n thus generated is conditionally typical approaches 1 as n becomes large.
To prove the conditional typicality lemma, we will employ the fact [to be proved earlier in the homework] that $P(T_\delta(P)) \overset{n\to\infty}{\to} 1$. Fix some $a \in \mathcal{X}$, and consider the subsequence of all components x_i in x^n that

are equal to a. Consider the subsequence of y_i's corresponding to the same indices. This subsequence is generated IID from the PMF $P_{Y|X=a}$. We will apply the aforementioned result separately to each such subsequence corresponding to a symbol in $a \in \mathcal{X}$.
To prove the bounds on the size of $|T_\delta(Y|x^n)|$, we will take a similar approach: we will use Equation (3) [which will also be proved earlier in the homework] and apply it to each subsequence associated with a symbol $a \in \mathcal{X}$.
We can interpret the Conditional Typicality Lemma qualitatively with the help of the following pictures:

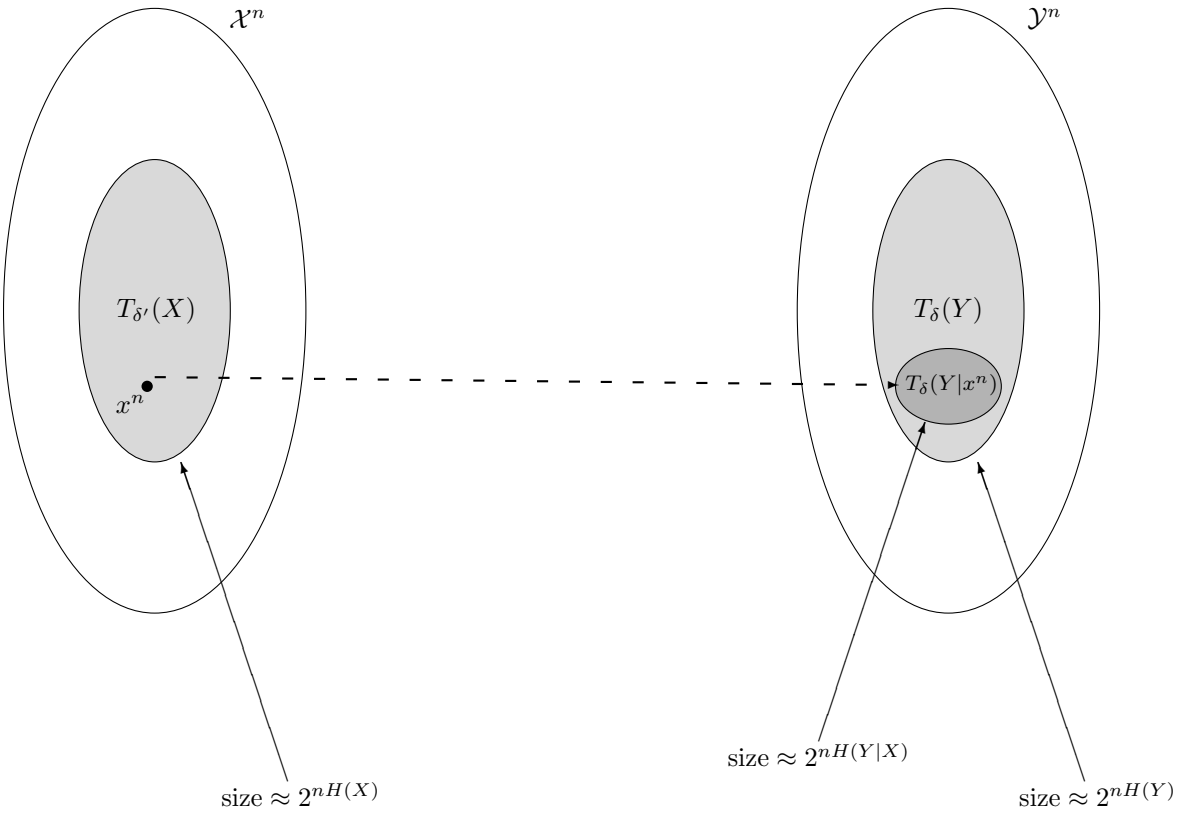

Figure 1: Illustration of the relationships between strongly δ-typical and conditionally δ-typical sets

The dashed line denotes that, given channel input x^n, the channel output will fall within the dark gray set $T_\delta(Y|x^n)$ with high probability. $T_\delta(Y|x^n)$ can be thought of the "noise ball" around the particular channel input sequence x^n. Recall that in lecture 11, we used this to give intuition for the channel coding converse.

Lemma 8 (***Joint Typicality Lemma***)**.** $\forall\, 0 < \delta' < \delta$, *if* $\tilde{Y}_i$ *IID* $\sim Y$, *then for all* n *sufficiently large and* $x^n \in T_{\delta'}(X)$,

$$2^{-n[I(X;Y)+\tilde{\epsilon}(\delta)]} \leq P(\tilde{Y}^n \in T_\delta(Y|x^n)) \leq 2^{-n[I(X;Y)-\tilde{\epsilon}(\delta)]} \tag{9}$$

where $\tilde{\epsilon}(\delta) \to 0$ *as* $\delta \to 0$.

The proof of the Joint Typicality Lemma will also be a homework problem. Intuitively speaking, since the sequence $\tilde{Y}^n$ is generated IID with respect to Y, on an exponential scale it is roughly uniformly distributed over the set $T_\delta(Y)$. Thus, the probability that the sequence falls within $T_\delta(Y|x^n)$ for some particular x^n is, on an exponential scale, roughly the ratio of the size of this set to the size of $T_\delta(Y)$, since $T_\delta(Y|x^n) \subseteq T_\delta(Y)$.

Again, refer to Figure 1 for a visual aid. So, $P(\tilde{Y}^n \in T_\delta(Y|X^n)) \approx \frac{2^{nH(Y|X)}}{2^{nH(Y)}} = 2^{-nI(X;Y)}$. So, the probability that a randomly generated sequence $\tilde{Y}^n$ "looks" jointly typical with a particular sequence x^n is exponentially unlikely.

In the next lecture, we will see why these notions are significant in the context of lossy compression. We will use them to prove the main achievability result of lossy compression.

3 Recap of the Strongly δ-Typical Set

An informal recap of previously discussed terms:

1. $T_\delta(X)$ = set of sequences x^n whose empirical distribution is close to pmf of X
2. $T_\delta(X,Y)$ = set of pairs of sequences (x^n, y^n) whose joint empirical distribution is close to the joint pmf of (X,Y)
3. $T_\delta(Y|x^n)$ = set of sequences y^n whose joint empirical distribution with x^n is close to joint pmf of (X,Y)

And respective sizes of these sets:

1. $|T_\delta(X)| \approx 2^{nH(X)}$
2. $|T_\delta(X,Y)| \approx 2^{nH(X,Y)}$
3. $|T_\delta(Y|x^n)| \approx 2^{nH(Y|X)}$ for $x^n \in T_\delta(X)$

Now we can look at the probability of randomly generated, iid sequences being in each of these sets. If we generate X_i iid $\sim X$, then the random sequence X^n is typical by the Law of Large Numbers,

$$Pr\left(X^n \in T_\delta(X)\right) \approx 1. \tag{10}$$

If a specific, δ-typical x^n is fed into a memoryless channel characterized by $P_{Y|X}$ to generate the stochastic channel output sequence Y^n, ie. $x^n \longrightarrow P(Y|X) \longrightarrow Y^n$, then Y^n is in the conditional δ-typical set $T_\delta(Y|x^n)$,

$$Pr\left(Y^n \in T_\delta(Y|x^n)\right) \approx 1, \forall x^n \in T_\delta(X). \tag{11}$$

Joint-Typicality Lemma: Finally we saw that for $\tilde{Y}_i$ iid $\sim Y$, the probability of the sequence $\tilde{Y}^n$ falling into the conditional δ-typical set given the input x^n is exponentially unlikely. That is, it is unlikely that any iid randomly generated sequence will look like the response of a channel to a particular input x^n. The probability can be described as a function of the mutual information between X and Y,

$$Pr\left(\tilde{Y}^n \in T_\delta(Y|x^n)\right) \approx 2^{-nI(X;Y)}. \tag{12}$$

By (10), $\tilde{Y}^n \in T_\delta(Y)$, so then the probability that it falls into the smaller subset $T_\delta(Y|x^n)$ of that region is small. Furthermore, we can express this approximation as a ratio:

$$2^{-nI(X;Y)} = \frac{2^{nH(Y|X)}}{2^{nH(Y)}} \approx \frac{|T_\delta(Y|x^n)|}{|T_\delta(Y)|} \tag{13}$$

Recall from Lecture 10, that given $\tilde{X}_i$ iid $\sim X$ and $\tilde{Y}_i$ iid $\sim Y$ generated independently, the probability they look jointly typical according to the notion of weak typicality is,

$$Pr\left((\tilde{X_i}^n, \tilde{Y_i}^n) \in A_e^{(n)}(X,Y)\right) \approx 2^{-nI(X;Y)} \tag{14}$$

This result is also true for the notion strong typicality and follows from Sanov's theorem. The Method of Types tells us that the probability that for $(\tilde{X_i}^n, \tilde{Y_i}^n)$ generated iid $\sim Q_{X,Y}$ looks like the joint empirical distribution P is $2^{-nD(P||Q)}$ (in this case, P is the joint distribution and Q is the product of the marginals). Thus:

$$Pr\left((\tilde{X_i}^n, \tilde{Y_i}^n) \in T_\delta(X,Y)\right) \approx 2^{-nD(P_{XY}||P_X \times P_Y)} \tag{15}$$

$$= 2^{-nI(X;Y)} \tag{16}$$

An alternative way to get this result without using Sanov's:

$$Pr\left((\tilde{X_i}^n, \tilde{Y_i}^n) \in T_\delta(X,Y)\right) \approx Pr\left(\tilde{X_i}^n \in T_\delta(X)\right) \times Pr\left(\tilde{Y_i}^n \in T_\delta(Y|\tilde{X}^n) \,|\, \tilde{X}^n \in T_\delta(X)\right) \tag{17}$$

$$\approx 1 \times 2^{-nI(X;Y)} \tag{18}$$

$$= 2^{-nI(X;Y)} \tag{19}$$

The idea is that the first requirement $Pr(\tilde{X_i}^n \in T_\delta(X))$ will cost nothing, being about 1 according to (10).

4 δ-Typicality in the Compression Setting

In the compression setting, let U be a random variable according to the source distribution and let V be the reconstruction random variable that is associated with mutual information minimization that characterizes the rate distortion function $R(D)$. Suppose (U, V) are generated according to their a joint pmf $P_{U,V}$. In this section, we apply the results we got from the previous section.

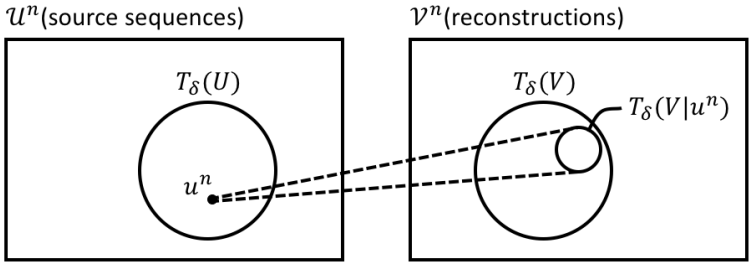

Figure 2: Conditionally typical set $T_{delta}(V|u^n)$

In Figure 2, $\mathcal{U}^n$ denotes the set of all possible source sequences of length n and $\mathcal{V}^n$ denotes the set of all possible reconstructions. For a particular source sequence $u^n \in T_\delta(U)$ from the set of typical source sequences, the conditionally typical set $T_\delta(V|u^n)$ is the set of all typical sequences v^n that are jointly typical with u^n. According to the Joint Typicality Lemma (3), if we generate an iid sequence $V_i \sim V$, then the probability that it belongs to the conditional typical set is,

$$Pr(V^n \in T_\delta(V|u^n)) \approx 2^{-nI(U;V)} \tag{20}$$

which is exponentially small. However, if we independently generate $2^{nI(U;V)}$ random V^n's, at least one will fall in $T_\delta(V|u^n)$ with high probability.

Note: If $(u^n, v^n) \in T_\delta(U,V)$, then

$$\frac{1}{n}\sum_{i_1}^{n} d(u_i, v_i) \approx \mathbb{E}[d(U,V)]. \tag{21}$$

We will use this argument to guarantee that for any source sequence you care about, there is some sequence in a randomly generated codebook of the appropriate size that is jointly typical with it. Therefore the distortion between the source and reconstruction sequences will be roughly the distortion between the generic pair (U, V).

5 Lossy Compression and $R(D)$

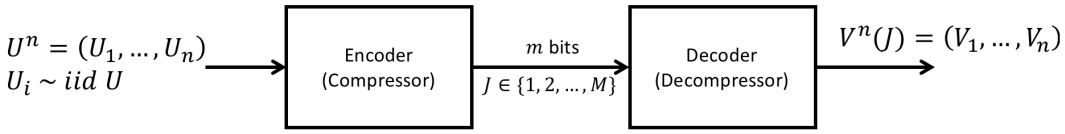

Figure 3: Scheme

A scheme is characterized by: $(n, M, \text{Encoder}, \text{Decoder})$ where

- n is the length of the source sequence
- $M = 2^m$ is the size of the index set or the number of bits you will use to represent a source n-tuple.

Communicating m bits is equivalent to 2^m possible messages so using m bits to represent the data is equivalent to conveying an index set of $M = 2^m$ indices. We can think of the encoder as having an output $J \in \{1, 2, \ldots, M\}$. Then the rate of the scheme is

$$\text{Rate} = \frac{\log M}{n} \frac{\text{bits}}{\text{source sequence}}. \tag{22}$$

We use the notation

$$d(u^n, v^n) = \frac{1}{n} \sum_{i=1}^{n} d(u_i, v_i). \tag{23}$$

Note 1. The decoder maps an index to a reconstruction. Therefore, specifying a decoder is equivalent to specifying a codebook $c_n = \{v^n(1), \ldots, v^n(M)\}$.

Note 2. Without loss of optimality, we can assume

$$d(U^n, V^n(J)) = \min_{v^n \in c_n} d(U^n, v^n). \tag{24}$$

i.e., the encoder is the optimal encoder for the given codebook. The encoder will output the index in the codebook that is closest under the relevant distortion criteria to the source sequence. That will lead to the smallest distortion with an optimal expected per-symbol distortion of

$$\text{expected\_dist}(c_n) = \mathbb{E}\left[\min_{v^n \in c_n} d(U^n, v^n)\right]. \tag{25}$$

6 Rate Distortion Theory

Reviewing the key definitions for rate distortion theory:

- (R, D) is achievable if $\forall \epsilon\ \exists\, n, c_n$ such that $|c_n| \leq 2^{n(R+\epsilon)}$ and $\text{expected\_dist}(c_n) \leq D + \epsilon$.

- $R(D) = \inf\{R' : (R', D) \text{ is achievable}\}$
- Theorem: $R(D) = \min_{E[d(U,V)] \leq D} I(U;V) = R^{(I)}(D)$

The above theorem is equivalent to:

- Converse: $R(D) \geq R^{(I)}(D)$
- Direct: $R(D) \leq R^{(I)}(D)$

The Direct Part of the the theorem can then be reframed as follows:

$$\text{If } U, V \text{ are such that } E[d(U,V)] \leq D \text{ and } R > I(U;V), \text{ then } (R, D) \text{ is achievable.} \tag{26}$$

If U,V is a feasible set for minimization, then any value for $I(U;V)$ in the feasible set (defined as $E[d(U,V)] \leq D$) is achievable and any rate $R > I(U;V)$ is such that (R, D) is achievable. Therefore the minimum of $I(U;V)$ in the feasible set is achievable and that minimizing pair (U, V) can be chosen.

6.1 Sketch of the Proof for the Direct Part

A rigorous proof of the following is in the class notes from 2016 on page 62.

The setup is as follows:

- Fix U, V such that $E[d(U,V)] \leq D$
- Fix $R > I(U;V)$
- Take $M = 2^{nR}$, where $M = |C_n|$ and is therefore $\gg 2^{nI(U;V)}$
- Generate a random codebook $C_n = \{V^n(1), V^n(2), ..., V^n(M)\}$, with $V^n(i)$ generated iid $\sim V$
- Fix u^n for any $u^n \in T_\delta(U)$

Recall that for a δ-typical u^n the probability that V^n is jointly typical is given by the Jointly Typical Lemma (Equation 12),

$$Pr\left((u^n, V^n(j)) \in T_\delta(U,V)\right) \approx 2^{-nI(U;V)} \, \forall \, 1 \leq j \leq M. \tag{27}$$

Since there are $M = 2^{nR} \gg 2^{nI(U;V)}$ j's, then with high probability one of the j's is jointly typical with u^n. This leads to the following results with high probability:

$$Pr\left((u^n, V^n(j)) \in T_\delta(U,V) \text{ for some } 1 \leq j \leq M\right) \approx 1 \tag{28}$$

$$\Rightarrow Pr\left(d(u^n, V^n(j)) \leq D \text{ for some } 1 \leq j \leq M\right) \approx 1 \tag{29}$$

$$\Rightarrow Pr\left(\min_{V^n \in C_n} d(u^n, V^n) \leq D\right) \approx 1 \tag{30}$$

This is all true conditioned on a fixed $u^n \in T_\delta(U)$, but in all likelihood $U^n \in T_\delta(U)$. This leads to the conclusions that:

$$Pr\left(\min_{v^n \in C_n} d(U^n, v^n) \leq D\right) \approx 1 \tag{31}$$

$$E[\min_{v^n \in C_n} d(U^n, v^n)] \leq D \tag{32}$$

Therefore, we can extract one particular codebook, ie. $\exists c_n$ such that:

$$|c_n| = M = 2^{nR} \tag{33}$$

$$\text{expected\_dist}(c_n) = E[\min_{v^n \in C_n} d(U^n, v^n)] \leq D \tag{34}$$

$$\Rightarrow (R, D) \text{ is achievable} \tag{35}$$

In conclusion, if we generate C_n randomly and generate $> 2^{nI(U;V)}$ V^n reconstructions randomly, then with high probability one of the reconstructions will be jointly typical with the input and hence consistent with the distortion criterion.

7 Recap

A quick recap of our current setting.

$$U_1 \dots U_n, \text{ iid} \sim U \to \boxed{\text{Encoder}} \xrightarrow{J \in \{1 \dots M\}} \boxed{\text{Decoder}} \to V_1 \dots V_n$$

$$\text{Rate is } \frac{\log M}{n} \frac{\text{bits}}{\text{symbol}}$$

$$\text{Distortion is } \mathrm{d}(U^n, V^n) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{d}(U_i, V_i)$$

We say a pair (R, D) is achievable if for any $\epsilon > 0$, there exists a scheme (encoder, decoder pair) with rate less than or equal to $R + \epsilon$, and expected distortion $\mathrm{E}[d(U^n, V^n)] \leq D + \epsilon$. In this setting, we define

$$R(D) = \inf \{R' : (R', D) \text{ is achievable}\}$$

$R(D)$ can be thought of as the minimum number of bits per symbol needed to achieve expected distortion D. From this, we have presented our main theorem for this section before, which states

$$R(D) = \min_{\mathrm{E}[d(U,V)] \leq D} I(U; V) \triangleq R^{(I)}(D)$$

In order to prove this main theorem, we split it into the two following parts, which together are equivalent to the theorem.

$$\text{Direct part: } R(D) \leq R^{(I)}(D) \text{ (Proven already)}$$

$$\text{Converse part: } R(D) \geq R^{(I)}(D) \text{ (Remains to prove)}$$

We will recap the idea for the proof of the direct part (sometimes called achievability). First, we generate a codebook $\{V^n(1), \dots, V^n(M)\}$ iid $\sim V$. Then, for a given $1 \leq i \leq m$:

$$\mathrm{P}((U^n, V^n(i)) \text{ is jointly typical}) \approx 2^{-nI(U;V)} \tag{36}$$

We have proved that a direct implication from (1) is:

$$\mathrm{P}((U^n, V^n(i)) \text{ is jointly typical for some } 1 \leq i \leq M) \approx 1, \text{ provided } M = 2^{nR}, \text{ for } R > I(U; V)$$

It is then clear that in order to cover the typical set $T_\delta(U)$, we need a number of points greater than or equal to $\frac{2^{nH(U)}}{2^{nH(U|V)}} = 2^{nI(U;V)}$.

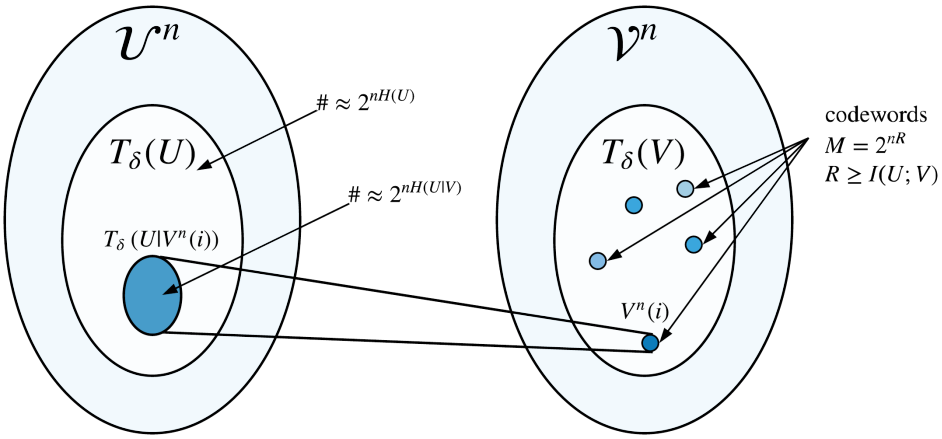

Figure 4: Typical sets and codewords - intuition. To cover $T_\delta(U)$ with conditionally typical balls, we need roughly $T_\delta(U)/T_\delta(U|V^n) \approx 2^{nI(U;V)}$ reconstruction sequences.

8 Proof of the converse part

Fix a scheme satisfying

$$\mathbb{E}\left[d(U^n, V^n)\right] \leq D$$

Then the entropy of the reconstruction under the scheme is no more than the log-size of the codebook, i.e.

$$H(V^n) \leq \log M$$

since the reconstruction takes values in a set of size at most M. Hence

$$\begin{aligned}
\log M &\geq H(V^n) \\
&\geq H(V^n) - H(V^n|U^n) \\
&= I(U^n; V^n) \\
&= H(U^n) - H(U^n|V^n) \\
&= \sum_{i=1}^{n} H(U_i) - H(U_i|U^{i-1}, V^n) \\
&\geq \sum_{i=1}^{n} H(U_i) - H(U_i|V_i) \quad \text{because conditioning reduces entropy} \\
&= \sum_{i=1}^{n} I(U_i; V_i) \\
&\geq \sum_{i=1}^{n} R^{(I)}\left(\mathbb{E}\left[d(U_i, V_i)\right]\right) \quad \text{from the definition of } R^{(I)}(D) \\
&\geq nR^{(I)}\Bigg(\underbrace{\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[d(U_i, V_i)\right]}_{\mathbb{E}[d(U^n,V^n)]\leq D}\Bigg) \quad \text{from the convexity of } R^{(I)}(D) \text{ (homework 5)} \\
&\geq nR^{(I)}(D) \quad \text{from the monotonicity of } R^{(I)}(\cdot)
\end{aligned}$$

thus

$$\text{rate} = \frac{\log M}{n} \geq R^{(I)}(D)$$

which finishes the proof of the converse part

$$R(D) \geq R^{(I)}(D)$$

□