EE376A/STATS376A Information Theory Lecture 16 - 03/01/2018

Lecture 16: Strongly Typical Sequences and Rate Distortion

Lecturer: Tsachy Weissman *Scribe: Evan Huang, Ariana Mann, Jae Hyuck Park*

In this lecture, we review the definition for strong typicality and the joint-typicality lemma. Returning to the compression setting, we apply the joint-typicality lemma to design a general optimal scheme for lossy compression, and use this to illustrate the proof for the direct part of the main theorem of rate distortion theory.

1 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,

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

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 $T_\delta(Y|x^n)$,

$$Pr\left(Y^n \in T_\delta(Y|x^n)\right) \approx 1, \forall x^n \in T_\delta(X). \tag{2}$$

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,

$$Pr\left(\tilde{Y}^n \in T_\delta(Y|x^n)\right) \approx 2^{-nI(X;Y)}. \tag{3}$$

By (1), $\tilde{Y}^n \in T_\delta(Y)$, 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:

$$2^{-nI(X;Y)} = \frac{2^{nH(Y|X)}}{2^{nH(Y)}} \approx \frac{|T_\delta(Y|x^n)|}{|T_\delta(Y)|} \tag{4}$$

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,

$$Pr\left((\tilde{X_i}^n, \tilde{Y_i}^n) \in A_\epsilon^{(n)}(X,Y)\right) \approx 2^{-nI(X;Y)} \tag{5}$$

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:

$$Pr\left((\tilde{X_i}^n, \tilde{Y_i}^n) \in T_\delta(X,Y)\right) \approx 2^{-nD(P_{XY}||P_X \times P_Y)} \tag{6}$$

$$= 2^{-nI(X;Y)} \tag{7}$$

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

$$Pr\left((\tilde{X_i}^n, \tilde{Y_i}^n) \in T_\delta(X,Y)\right) \approx Pr\left(\tilde{X_i}^n \in T_\delta(X)\right) \times Pr\left(\tilde{Y_i}^n \in T_\delta(Y|\tilde{X}^n) \,|\, \tilde{X}^n \in T_\delta(X)\right) \tag{8}$$

$$\approx 1 \times 2^{-nI(X;Y)} \tag{9}$$

$$= 2^{-nI(X;Y)} \tag{10}$$

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

2 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.

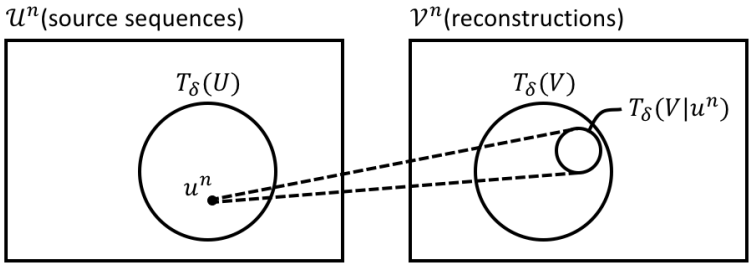

Figure 1: Conditionally typical set $T_{delta}(V|u^n)$

In Figure 1, $\mathcal{U}^n$ 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,

$$Pr(V^n \in T_\delta(V|u^n)) \approx 2^{-nI(U;V)} \tag{11}$$

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

$$\frac{1}{n}\sum_{i_1}^{n} d(u_i, v_i) \approx \mathbb{E}[d(U, V)]. \tag{12}$$

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).

3 Lossy Compression and $R(D)$

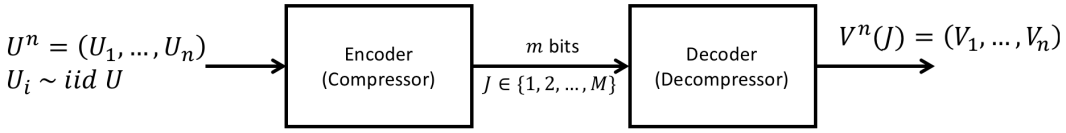

Figure 2: Scheme

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

$$\text{Rate} = \frac{\log M}{n} \frac{\text{bits}}{\text{source sequence}}. \tag{13}$$

We use the notation

$$d(u^n, v^n) = \frac{1}{n}\sum_{i=1}^{n} d(u_i, v_i). \tag{14}$$

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

$$d(U^n, V^n(J)) = \min_{v^n \in c_n} d(U^n, v^n). \tag{15}$$

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

$$\text{expected\_dist}(c_n) = \mathbb{E}\left[\min_{v^n \in c_n} d(U^n, v^n)\right]. \tag{16}$$

4 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 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:

$$\text{If } U, V \text{ are such that } E[d(U,V)] \leq D \text{ and } R > I(U;V), \text{ then } (R, D) \text{ is achievable.} \tag{17}$$

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.

4.1 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 3),

$$Pr\left((u^n, V^n(j)) \in T_\delta(U,V)\right) \approx 2^{-nI(U;V)} \; \forall \, 1 \leq j \leq M. \tag{18}$$

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:

$$Pr\left((u^n, V^n(j)) \in T_\delta(U,V) \text{ for some } 1 \leq j \leq M\right) \approx 1 \tag{19}$$

$$\Rightarrow Pr\left(d(u^n, V^n(j)) \leq D \text{ for some } 1 \leq j \leq M\right) \approx 1 \tag{20}$$

$$\Rightarrow Pr\left(\min_{V^n \in C_n} d(u^n, V^n) \leq D\right) \approx 1 \tag{21}$$

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:

$$Pr\left(\min_{v^n \in C_n} d(U^n, v^n) \leq D\right) \approx 1 \tag{22}$$

$$E[\min_{v^n \in C_n} d(U^n, v^n)] \leq D \tag{23}$$

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

$$|c_n| = M = 2^{nR} \tag{24}$$

$$\text{expected\_dist}(c_n) = E[\min_{v^n \in C_n} d(U^n, v^n)] \leq D \tag{25}$$

$$\Rightarrow (R, D) \text{ is achievable} \tag{26}$$

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.