EE376A Information Theory Lecture 1 - 02/26/2015

Lecture 15: Conditional and Joint Typicaility

Lecturer: Kartik Venkat *Scribe: Max Zimet, Brian Wai, Sepehr Nezami*

1 Notation

We always write a sequence of symbols by small letter. For example x^n is an individual sequence without any probability distribution assigned to it. We use capital letter for random variables, e.g. X^n i.i.d. according to some distribution. Throughout, $\mathcal{X}$ will denote the set of possible values a symbol can take.

2 Strong δ - typical sets

Before going forward, we define empirical distribution.

Definition 1. *Emperical Distribution:*

For any sequence x^n, empirical distribution is the probability distribution derived for letters of the alphabet based on frequency of appearance of that specific letter in the sequence. More precisely:

$$p_{x^n}(a) = \frac{1}{n}\sum 1(x_i = a), \quad \forall a \in \mathcal{X} \tag{1}$$

2.1 Typical Set (again)

A rough intuition for typical sets is that if one picks a sequence from an i.i.d distribution, $p \sim X^n$, then the typical set $\mathcal{T}(X)$ is a set of length n sequences with the following properties:
1. A sequence chosen at random will be in the typical set with probability almost one.
2. All the elements of the typical set have (almost) equal probabilities.

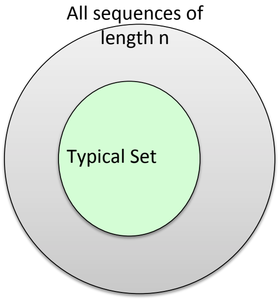


Figure 1: Space of all sequences and typical set inside.

More precisely, if $\mathcal{T}(X)$ is the typical set, $|\mathcal{T}(X)| \approx 2^{nH(X)}$ and probability of each sequence inside the typical set is $\sim 2^{-nH(X)}$. So a random sequence chosen from the set looks like one chosen uniformly from the typical set.

3 δ- strongly typical set (recap from last lecture)

Definition 2. *A sequence x^n is δ - typical set with respect to pmf p if:*

$$\forall a \in \mathcal{X} : |p_{x^n}(a) - p(a)| \leq \delta p(a) \tag{2}$$

Where $\mathcal{X}$ is the support of the distribution.

Note that this "strongly typical" set is different from the other "(weakly) typical" set defined in first lectures. This new notion is stronger, but as we will see it retains all the desirable properties of typicality, and more. The following example illustrates difference of this strong notion and weak notion defined earlier:
Example 3. Suppose that alphabet is $\mathcal{X} = \{a, b, c\}$ with the probabilities $p(a) = 0.1$, $p(b) = 0.8$ and $p(c) = 0.1$. Now consider two strings of length 1000:

$$x^{1000}_{strong} = (100a, 800b, 100c)$$

$$x^{1000}_{weak} = (200a, 800b)$$

In this example, these two sequences have same probability, so they are both identical in the weak notion of typical set $A^{(\epsilon)}_{1000}$ (for some ϵ). But it is not hard to see that x^{1000}_{strong} is a δ- strong typical set while the other is not (for sufficiently small δ). The strong notion is sensitive to frequency of different letters and not only the total probability of sequence.

It was shown in homework 6 that that we preserve important properties of typical sets in the new strong notion. We have the following results from homework:

1. $\mathcal{T}_\delta(X) \subseteq A_\epsilon(X)$ (i.e. strong typical sets are inside weak typical sets.)
2. Empirical probability p_{X^n} is almost equal to the probability distribution p. Therefore $p(x^n) \approx 2^{-nH(X)}$ for x^n in the strong typical set.
3. $|\mathcal{T}_\delta(X)| \approx 2^{nH}$.
4. $P(x^n \in \mathcal{T}_\delta(X)) \to 1$ as $n \to \infty$

4 δ - jointly typical set

In this section we extend the notion of δ - typical sets to pair of sequences $x^n = (x_1, ..., x_n)$ and $y^n = (y_1, ..., y_n)$ from alphabets $\mathcal{X}$ and $\mathcal{Y}$.

Definition 4. *A pair of sequences (x^n, y^n) is said to be δ - jointly typical with respect to a pmf p_{XY} on $\mathcal{X} \times \mathcal{Y}$ if:*

$$|p_{x^n,y^n}(x, y) - p(x, y)| \leq \delta p(x, y)$$

Where $p_{x^n,y^n}(x, y)$ is the empirical distribution.

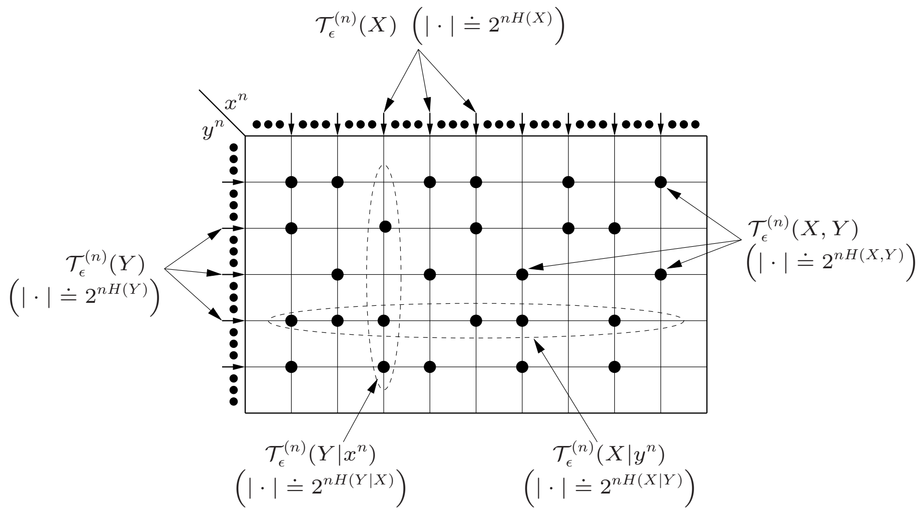
Similarly we define δ - jointly typical set $\mathcal{T}_\delta(X, Y)$:

$$\mathcal{T}_\delta(X, Y) = \mathcal{T}_\delta(P_{XY}) \triangleq \{(x^n, y^n) \,|\, p_{x^n,y^n}(x, y) - p(x, y)| \leq \delta p(x, y), \forall x \in \mathcal{X}, y \in \mathcal{Y}\}$$

If we look carefully, nothing is really very new. We just require that empirical distribution of pair of sequences be δ - close to the pmf p_{XY}.

It is easy to see that size of typical set is:

$$|\mathcal{T}_\delta(X, Y)| \approx 2^{nH(X,Y)}$$



$\mathcal{T}_\epsilon^{(n)}(X)$ $\left(|\cdot| \doteq 2^{nH(X)}\right)$
x^n
y^n
$\mathcal{T}_\epsilon^{(n)}(Y)$
$\left(|\cdot| \doteq 2^{nH(Y)}\right)$
$\mathcal{T}_\epsilon^{(n)}(X,Y)$
$\left(|\cdot| \doteq 2^{nH(X,Y)}\right)$
$\mathcal{T}_\epsilon^{(n)}(Y|x^n)$
$\left(|\cdot| \doteq 2^{nH(Y|X)}\right)$
$\mathcal{T}_\epsilon^{(n)}(X|y^n)$
$\left(|\cdot| \doteq 2^{nH(X|Y)}\right)$

Figure 2: A useful diagram depicting typical sets, from El Gamal and Kim (Chapter 2)

and

$$P((x^n, y^n) \in \mathcal{T}_\delta(X,Y)) \approx 2^{-nH(X,Y)}$$

In Figure 2we have depicted the sequences of the length n from alphabet $\mathcal{X}$ on the x axis and sequences from alphabet $\mathcal{Y}$ on the y axis.

Then we can look inside the table and any point corresponds to a pair of sequences. We have marked the jointly typical sets with dots. It is easy to see that if a set is jointly typical then both of the sequences in the set are typical as well. Also we will see in the next section that number of dots in the column corresponding to a typical x^n sequence is approximately $2^{nH(Y|X)}$.(Also similar statement is correct for rows). We can quickly check that this is consistent by a counting argument: We know that number of typical x^n sequences is $2^{nH(X)}$ and each column there are $2^{nH(Y|X)}$ jointly typical pairs. So in total there are $2^{nH(X)}.2^{nH(Y|X)}$ number of typical pairs. But $2^{nH(X)}.2^{nH(Y|X)} = 2^{n(H(X)+H(Y|X))} = 2^{nH(X,Y)}$ which is consistent to the fact that total number of jointly typical pairs is equal to $2^{nH(X,Y)}$.

5 δ-conditional typicality

In many applications of typicality, we know one sequence (say, the input to a channel) and want to know typical values for some other sequence. In this case, a useful concept is conditional typicality.

Definition 5. *Fix $\delta > 0, x^n \in \mathcal{X}^n$. Then, the conditional typical set $\mathcal{T}_\delta(Y|x^n)$ is defined by*

$$\mathcal{T}_\delta(Y|x^n) \triangleq \{y^n \in \mathcal{Y}^n : (x^n, y^n) \in \mathcal{T}_\delta(X,Y)\}.$$

As usual, it is useful to have bounds on the size of this set. For all $\delta' < \delta$ (and n large enough, as usual), $x^n \in \mathcal{T}_{\delta'}(X) \Rightarrow |\mathcal{T}_\delta(Y|x^n)| \doteq 2^{nH(Y|X)}$ (below, we'll be a bit more careful, and we'll see that this exponent should depend on a function $\epsilon(\delta)$ which vanishes as $\delta \to 0$). Note that this asymptotic behavior is does not depend on the specific sequence x^n, as long as x^n is typical! This is all in accordance with the intuition we have developed: all typical sequences behave roughly similarly. If $x^n \notin \mathcal{T}_\delta(X)$, then $|\mathcal{T}_\delta(Y|x^n)| = 0$, as (x^n, y^n) cannot be jointly typical if x^n is not typical.

To illustrate the methods used to describe the asymptotic behavior of $|T_\delta(Y|x^n)|$, we find an upper bound on this value. If $x^n \in \mathcal{T}_{\delta'}(X) \subseteq \mathcal{T}_\delta(X)$ and $y^n \in \mathcal{T}_\delta(Y|x^n)$, then by definition of the conditional typical set, $(x^n, y^n) \in \mathcal{T}_\delta(X,Y)$. Since strong typicality implies weak typicality,

$$(1-\delta)H(X,Y) \leq -\frac{1}{n}\log p(x^n, y^n) \leq (1+\delta)H(X,Y), \tag{3}$$

$$(1-\delta)H(X) \leq -\frac{1}{n}\log p(x^n) \leq (1+\delta)H(X). \tag{4}$$

So, fix some $x^n \in \mathcal{T}_{\delta'}(X)$. Then,

$$\begin{aligned}1 &\geq \sum_{y^n \in \mathcal{T}_\delta(Y|x^n)} p(y^n|x^n) = \sum_{y^n \in \mathcal{T}_\delta(Y|x^n)} \frac{p(x^n,y^n)}{p(x^n)} \\ &\geq |\mathcal{T}_\delta(Y|x^n)| \cdot 2^{-n(H(X,Y)-H(X)+\epsilon(\delta))} = |\mathcal{T}_\delta(Y|x^n)| \cdot 2^{-n(H(Y|X)+\epsilon(\delta))} \\ &\Rightarrow |T_\delta(Y|x^n)| \leq 2^{n(H(Y|X)+\epsilon(\delta))}.\end{aligned}$$

6 Conditional Typicality, and Encoding / Decoding Schemes for Sending Messages

We now explain the idea of how these concepts relate to sending messages over channels. Consider a discrete memoryless channel (DMC), described by conditional probability distribution $p(y|x)$, where x is the input to the channel and y is the output. Then,

$$p(y^n|x^n) = \prod_{i=1}^{n} p(y_i|x_i).$$

Say we want to send one of M messages over a channel. We encode each $m \in \{1, \ldots, M\}$ into a codeword $X^n(m)$. We then send the codeword over the channel, obtaining Y^n. Finally, we use a decoding rule $\hat{M}(Y^n)$ which yields $\hat{M} = m$ with high probability.

The conditional typicality lemma (to be proved in homework 7) characterizes the behavior of this channel for large n: if we choose a typical input, then the output is essentially chosen uniformly at random from $\mathcal{T}_\delta(Y|x^n)$. More precisely, for all $\delta' < \delta$, $x^n \in \mathcal{T}_{\delta'}(X)$ implies that

$$P(y^n \in T_\delta(Y|x^n)) = P((x^n, Y^n) \in \mathcal{T}_\delta(X,Y)) \to 1,$$

as $n \to \infty$; furthermore, the probability of obtaining each element of $\mathcal{T}_\delta(Y|x^n)$ is essentially $\dfrac{1}{|\mathcal{T}_\delta(Y|x^n)|} \approx 2^{-nH(Y|X)}$.

This lemma limits the values of M for which there is an encoding / decoding scheme that can succeed with high probability. As illustrated in Figure 3, what we are doing is choosing a typical codeword $x^n \in \mathcal{T}_\delta(X)$, and receiving some element $Y^n \in T_\delta(Y|x^n)$. We can think of this latter set as a "noise ball": it is the set of outputs y^n that we could typically expect to receive, given that our input is x^n. If these noise balls corresponding to different inputs overlap significantly, then we have no hope for being able to obtain m from Y^n with high probability, as multiple inputs give indistinguishable outputs. Since, for any input, the output will (with high probability) be typical – that is, $Y^n \in T_\delta(Y)$, the number of messages we can send is limited by the number of noise balls we can fit inside of $T_\delta(Y)$. Since the number of elements of $T_\delta(Y)$ is (approximately) $2^{nH(Y)}$ and the number of elements of $T_\delta(Y|x^n)$ is (approximately) $2^{nH(Y|X)}$, it follows that the number of messages we can send over this channel is at most

$$\frac{2^{nH(Y)}}{2^{nH(Y|X)}} = 2^{n(H(Y)-H(Y|X))} = 2^{nI(X;Y)} \leq 2^{nC},$$

where C is the channel capacity. Note that this argument does not give a construction that lets us attain this upper bound on the communication rate. The magic of the direct part of Shannon's channel coding theorem is that random coding lets us attain this upper bound.

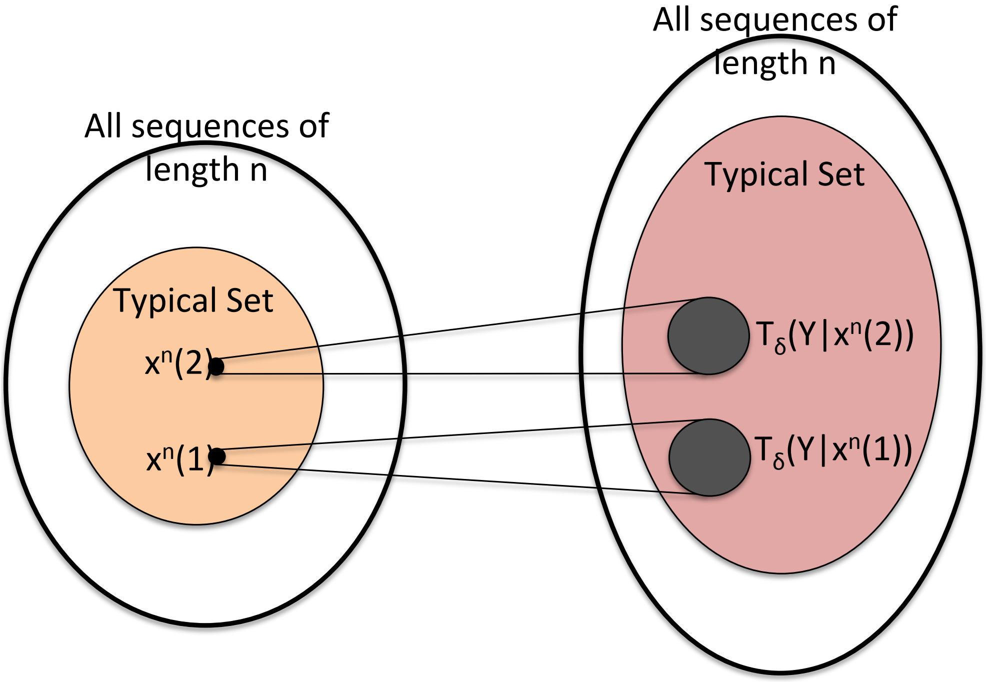


Figure 3: The typical sets $T_\delta(X), T_\delta(Y)$, and $T_\delta(Y|x^n)$.

7 Joint Typicality Lemma

In this final section, we discuss the joint typicality lemma, which tells us how well we can guess the output without knowing the input. Intuitively, if X and Y are strongly correlated, then we might expect that not knowing the input could strongly impair our ability to guess the output, and if X and Y are independent then not knowing the input should not at all impair our ability to guess the output. So, say that the actual input is x^n. We'll look for a bound on the probability that an output will be in the conditional typical set $\mathcal{T}_\delta(Y|x^n)$ – that is, the probability that we'll guess that x^n was the input – in terms of the mutual information $I(X;Y)$.

Fix any $\delta > 0$ and $\delta' < \delta$, and fix $x^n \in T_{\delta'}(X)$. Choose $\tilde{Y}^n \in \mathcal{Y}^n$ by choosing each $\tilde{Y}_i$ i.i.d. according to the marginal distribution $p(y)$. (So, intuitively we've forgotten what we sent as input to the channel, and are simulating the output). Then, noting that

$$y^n \in \mathcal{T}_\delta(Y|x^n) \Rightarrow y^n \in \mathcal{T}_\delta(Y) \Rightarrow p(y^n) \leq 2^{-n(H(Y)-\epsilon(\delta))},$$

where $\epsilon(\delta)$ is a function that approaches 0 as $\delta \to 0$, we have

$$\begin{aligned}
P(\tilde{Y}^n \in \mathcal{T}_\delta(Y|x^n)) &= P((x^n, \tilde{Y}^n) \in \mathcal{T}_\delta(X,Y)) \\
&= \sum_{y^n \in \mathcal{T}_\delta(Y|x^n)} p(y^n) \\
&\leq |T_\delta(Y|x^n)| \cdot 2^{-n(H(Y)-\epsilon(\delta))} \\
&\leq 2^{nH(Y|X)+\epsilon(\delta)} \cdot 2^{-n(H(Y)-\epsilon(\delta))} \\
&= 2^{-n(H(Y)-H(Y|X)-\tilde{\epsilon}(\delta))} \\
&= 2^{-n(I(X;Y)-\tilde{\epsilon}(\delta))},
\end{aligned}$$

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

Intuitive argument for joint typicality lemma The joint typicality lemma asserts that the probability of observing two random x^n and y^n sequences is roughly $2^{-nI(X;Y)}$. Observe that there are roughly $2^{nH(X)}$ typical x^n sequences, and $2^{nH(Y)}$ typical y^n sequences. The total number of jointly typical sequences is $2^{nH(X,Y)}$. Thus, what is the probability that two randomly chosen sequences are jointly typical?

$$\approx \frac{2^{nH(X,Y)}}{2^{nH(X)} \times 2^{nH(Y)}} = 2^{-nI(X;Y)} \tag{5}$$

8 Next time

Next lecture will be on lossy compression.