EE376A/STATS376A Information Theory Lecture 15 - 02/27/2018

Lecture 15: Strong, Conditional, & Joint Typicality

Lecturer: Tsachy Weissman Scribe: Nimit Sohoni, William McCloskey, Halwest Mohammad

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:

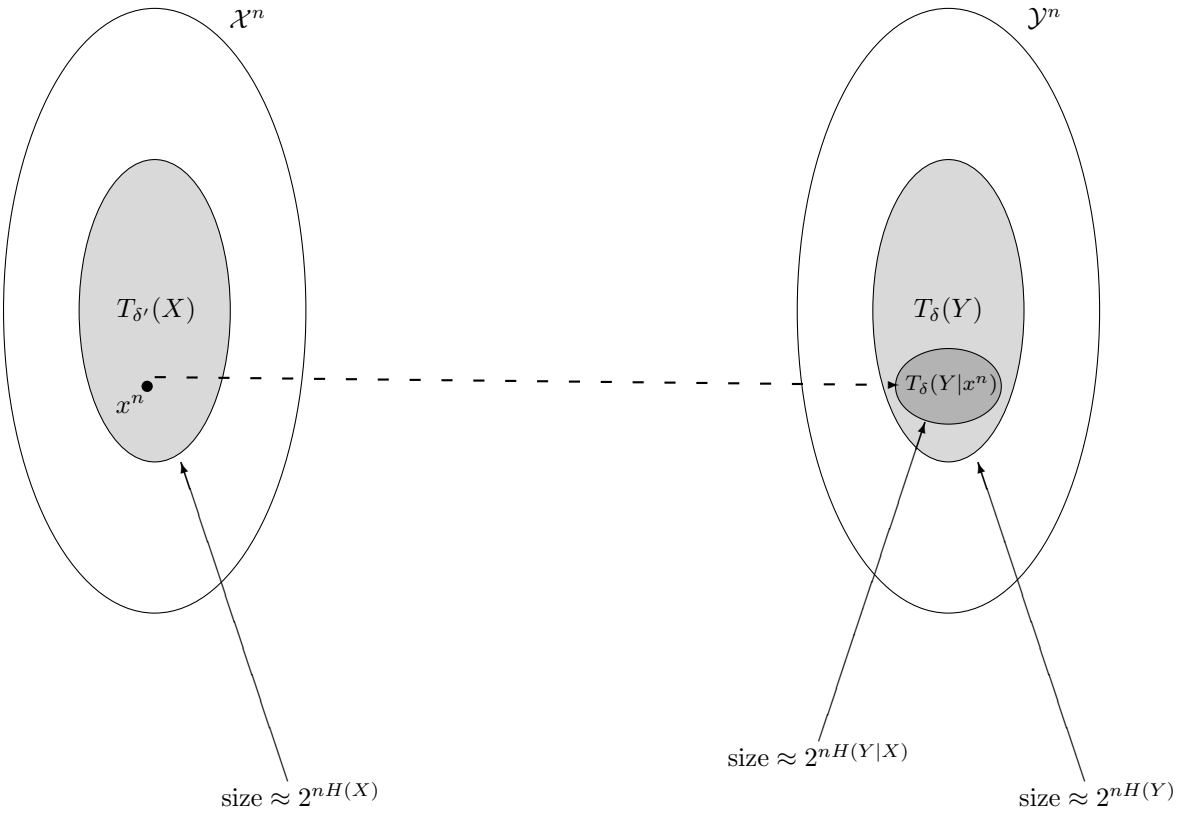

Figure 1: Illustration of the relationships between strongly δ-typical and conditionally δ-typical sets

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.