PHYS 7895: Quantum Information Theory Fall 2015

Lecture 4 — September 2, 2015

Prof. Mark M. Wilde *Scribe: Chenglong You*

1 Overview

In the last lecture we detailed the information processing task for channel coding, and we discussed an overview of Shannon's Channel Capacity Theorem.

In this lecture, we begin with the formal definition of a conditionally typical set and prove three important properties regarding it. Next we provide a detailed proof of Shannon's channel capacity theorem, discussing the decoding algorithm in detail and analyzing the expectation of the average error probability.

2 Conditionally Typical Set

Conditional typicality is a property that we expect to hold for any two random sequences—it is also a useful tool in the proofs of coding theorems. Suppose two random variables X and Y have respective alphabets $\mathcal{X}$ and $\mathcal{Y}$ and a joint distribution $p_{X,Y}(x,y)$. We can factor the joint distribution $p_{X,Y}(x,y)$ as the product of a marginal distribution $p_X(x)$ and a conditional distribution $p_{Y|X}(y|x)$, and this factoring leads to a particular way that we can think about generating realizations of the joint random variable. We can consider random variable Y to be a noisy version of X, where we first generate a realization x of the random variable X according to the distribution $p_X(x)$ and follow by generating a realization y of the random variable Y according to the conditional distribution $p_{Y|X}(y|x)$.

2.1 Definition of Conditionally Typical Set

Definition 1 (Conditional Sample Entropy)**.** *The conditional sample entropy $\overline{H}(y^n|x^n)$ of two sequences x^n and y^n is*

$$\overline{H}(y^n|x^n) = -\frac{1}{n}\log p_{Y^n|X^n}(y^n|x^n), \tag{1}$$

where

$$p_{Y^n|X^n}(y^n|x^n) \equiv p_{Y|X}(y_1|x_1)\cdots p_{Y|X}(y_n|x_n). \tag{2}$$

Definition 2 (Conditionally Typical Set)**.** *The δ-conditionally typical set $T_\delta^{Y^n|x^n}$ consists of all sequences y^n whose conditional sample entropy is δ-close to the true conditional entropy:*

$$T_\delta^{Y^n|x^n} \equiv \left\{y^n : \left|\overline{H}(y^n|x^n) - H(Y|X)\right| \leq \delta\right\}. \tag{3}$$

Figure 1 provides an intuitive picture of the notion of conditional typicality.

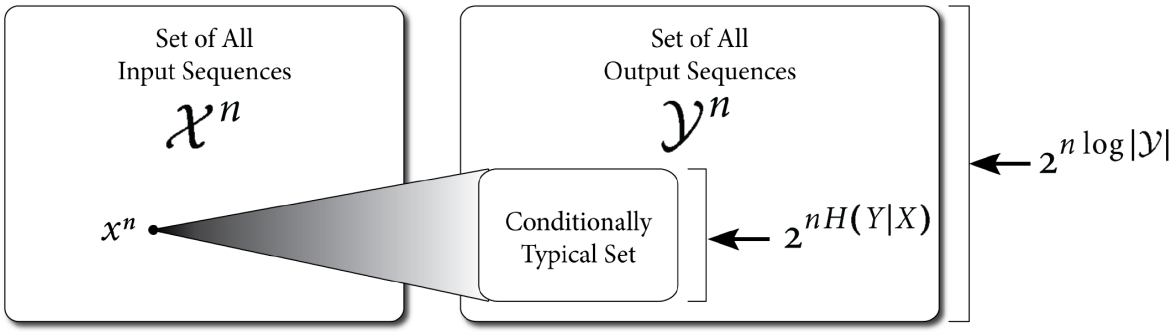

Figure 1: This figure depicts the notion of a conditionally typical set. Associated to every input sequence x^n is a conditionally typical set consisting of the likely output sequences. The size of this conditionally typical set is $\approx 2^{nH(Y|X)}$. It is exponentially smaller than the set of all output sequences whenever the conditional random variable is not uniform.

2.2 Properties of the Conditionally Typical Set

The set $T_\delta^{Y^n|x^n}$ of conditionally typical sequences enjoys the following three properties:

2.2.1 Unit Probability

Property 2.1 (Unit Probability)**.** *The set $T_\delta^{Y^n|x^n}$ asymptotically has probability one when the sequence x^n is random. So as n becomes large, it is highly likely that random sequences Y^n and X^n are such that Y^n is a conditionally typical sequence. We formally state this property as follows:*

$$\forall \varepsilon > 0 \qquad \mathbb{E}_{X^n} \left\{ \Pr_{Y^n|X^n} \left\{ Y^n \in T_\delta^{Y^n|X^n} \right\} \right\} \geq 1 - \varepsilon \qquad \textit{for sufficiently large } n. \tag{4}$$

We now prove the first property. This is just again another application of the law of large numbers. Consider that

$$\begin{aligned}
&\mathbb{E}_{X^n} \left\{ \Pr_{Y^n|X^n} \left\{ Y^n \in T_\delta^{Y^n|X^n} \right\} \right\} \\
&= \mathbb{E}_{X^n} \left\{ \mathbb{E}_{Y^n|X^n} \left\{ I_{T_\delta^{Y^n|X^n}}(Y^n) \right\} \right\} &(5)\\
&= \mathbb{E}_{X^n,Y^n} \left\{ I_{T_\delta^{Y^n|X^n}}(Y^n) \right\} &(6)\\
&= \sum_{x^n \in \mathcal{X}^n, y^n \in \mathcal{Y}^n} p_{X^n,Y^n}(x^n, y^n) I_{T_\delta^{Y^n|x^n}}(y^n), &(7)
\end{aligned}$$

where I denotes an indicator function. Given random variables X and Y, let us define the random variable $g(X,Y) = -\log p_{Y|X}(Y|X)$. Consider that the sample conditional entropy $\overline{H}(Y^n|X^n)$ for

the random sequences X^n and Y^n factors as follows:

$$\overline{H}(Y^n|X^n) = -\frac{1}{n}\log p_{Y^n|X^n}(Y^n|X^n) \tag{8}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left[-\log p_{Y|X}(Y_i|X_i)\right] \tag{9}$$

$$= \frac{1}{n}\sum_{i=1}^{n} g\left(X_i, Y_i\right). \tag{10}$$

This is the sample average of the random variable $g(X,Y)$ and the expectation of this random variable is

$$\mathbb{E}_{X,Y}\{g(X,Y)\} = \mathbb{E}_{X,Y}\{\left[-\log p_{Y|X}(Y|X)\right]\} \tag{11}$$

$$= \sum_{x,y} p_{X,Y}(x,y)\left[-\log p_{Y|X}(y|x)\right] \tag{12}$$

$$= H(Y|X). \tag{13}$$

Given all of the above, we can rewrite (7) as follows:

$$\Pr_{X^nY^n}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} g\left(X_i, Y_i\right) - \mathbb{E}_{X,Y}\{g(X,Y)\}\right| \leq \delta\right\}. \tag{14}$$

By applying the law of large numbers, this is larger than $1-\varepsilon$ for all $\varepsilon \in (0,1)$ and sufficiently large n.

2.2.2 Exponentially Smaller Cardinality

Property 2.2 (Exponentially Smaller Cardinality)**.** *The number $\left|T_\delta^{Y^n|x^n}\right|$ of δ-conditionally typical sequences is exponentially smaller than the total number $|\mathcal{Y}|^n$ of sequences for any conditional random variable $Y|X$ that is not uniform. We formally state this property as follows:*

$$\left|T_\delta^{Y^n|x^n}\right| \leq 2^{n(H(Y|X)+\delta)}. \tag{15}$$

We can also lower bound the expected size of the δ-conditionally typical set when n is sufficiently large and x^n is a random sequence:

$$\forall \varepsilon > 0 \qquad \mathbb{E}_{X^n}\left\{\left|T_\delta^{Y^n|X^n}\right|\right\} \geq (1-\varepsilon)\, 2^{n(H(Y|X)-\delta)} \qquad \text{for sufficiently large } n. \tag{16}$$

2.2.3 Equipartition

Property 2.3 (Equipartition)**.** *The probability of a given δ-conditionally typical sequence y^n (corresponding to the sequence x^n) is approximately uniform:*

$$2^{-n(H(Y|X)+\delta)} \leq p_{Y^n|X^n}\left(y^n|x^n\right) \leq 2^{-n(H(Y|X)-\delta)}. \tag{17}$$

3 Decoding Algorithm and Error Analysis

3.1 Steps for Decoding

After receiving the sequence y^n from the channel outputs, Bob performs the following decoding algorithm:

1. Test whether y^n is in the typical set $T_\delta^{Y^n}$ corresponding to the distribution

$$p_Y(y) \equiv \sum_x p_{Y|X}(y|x)p_X(x).$$

   If it is not, then he reports an error.

2. He then tests if there is some message m such that the sequence y^n is in the conditionally typical set $T_\delta^{Y^n|x^n(m)}$. If m is the unique message such that $y^n \in T_\delta^{Y^n|x^n(m)}$, then he declares m to be the transmitted message. If there is no message m such that $y^n \in T_\delta^{Y^n|x^n(m)}$ or multiple messages m' such that $y^n \in T_\delta^{Y^n|x^n(m')}$, then he reports an error.

Observe that the decoder is a function of the channel, so that we might say that we construct channel codes "from the channel."

3.2 Error Analysis

As discussed in the above decoding algorithm, there are three kinds of errors that can occur in this communication scheme when Alice sends the codeword $x^n(m)$ over the channels:

$\mathcal{E}_0(m)$**:** The event that the channel output y^n is not in the typical set $T_\delta^{Y^n}$.

$\mathcal{E}_1(m)$**:** The event that the channel output y^n is in $T_\delta^{Y^n}$ but not in the conditionally typical set $T_\delta^{Y^n|x^n(m)}$.

$\mathcal{E}_2(m)$**:** The event that the channel output y^n is in $T_\delta^{Y^n}$ but it is in the conditionally typical set for some other message:

$$\left\{y^n \in T_\delta^{Y^n}\right\} \text{ and } \left\{\exists m' \neq m : y^n \in T_\delta^{Y^n|x^n(m')}\right\}. \tag{18}$$

It is helpful to analyze the expectation of the average error probability, where the expectation is with respect to the random selection of the code and the average is with respect to a uniformly random choice of the message m. Let $\mathcal{C} \equiv \{X^n(1), X^n(2), \ldots, X^n(|\mathcal{M}|)\}$ denote the random variable corresponding to the random selection of a code. The expectation of the average error probability of a randomly selected code is as follows:

$$\mathbb{E}_{\mathcal{C}} \left\{ \frac{1}{|\mathcal{M}|} \sum_m \Pr\left\{\mathcal{E}_0(m) \cup \mathcal{E}_1(m) \cup \mathcal{E}_2(m)\right\} \right\}. \tag{19}$$

Our first "move" is to exchange the expectation and the sum, following from linearity of the expectation:

$$\frac{1}{|\mathcal{M}|}\sum_{m}\mathbb{E}_{\mathcal{C}}\left\{\Pr\left\{\mathcal{E}_0(m)\cup\mathcal{E}_1(m)\cup\mathcal{E}_2(m)\right\}\right\}. \tag{20}$$

Since all codewords are selected in the same way (randomly and independently of the message m and according to the same distribution $p_{X^n}(x^n)$), the following equality holds for all $m, m' \in \mathcal{M}$:

$$\mathbb{E}_{\mathcal{C}}\left\{\Pr\left\{\mathcal{E}_0(m)\cup\mathcal{E}_1(m)\cup\mathcal{E}_2(m)\right\}\right\} = \mathbb{E}_{\mathcal{C}}\left\{\Pr\left\{\mathcal{E}_0(m')\cup\mathcal{E}_1(m')\cup\mathcal{E}_2(m')\right\}\right\}, \tag{21}$$

implying that it suffices to analyze $\mathbb{E}_{\mathcal{C}}\left\{\Pr\left\{\mathcal{E}_0(m)\cup\mathcal{E}_1(m)\cup\mathcal{E}_2(m)\right\}\right\}$ for just a single message m. Without loss of generality, we can pick $m = 1$ (the first message). Using the above, we find that the expectation of the average error probability simplifies as follows:

$$\frac{1}{|\mathcal{M}|}\sum_{m}\mathbb{E}_{\mathcal{C}}\left\{\Pr\left\{\mathcal{E}_0(m)\cup\mathcal{E}_1(m)\cup\mathcal{E}_2(m)\right\}\right\} = \mathbb{E}_{\mathcal{C}}\left\{\Pr\left\{\mathcal{E}_0\left(1\right)\cup\mathcal{E}_1\left(1\right)\cup\mathcal{E}_2\left(1\right)\right\}\right\}. \tag{22}$$

So we can then apply the union bound:

$$\begin{aligned}\mathbb{E}_{\mathcal{C}}\left\{\Pr\left\{\mathcal{E}_0(1)\cup\mathcal{E}_1(1)\cup\mathcal{E}_2(1)\right\}\right\}&\\ \leq \mathbb{E}_{\mathcal{C}}\left\{\Pr\left\{\mathcal{E}_0\left(1\right)\right\}\right\} + \mathbb{E}_{\mathcal{C}}\left\{\Pr\left\{\mathcal{E}_1(1)\right\}\right\} + \mathbb{E}_{\mathcal{C}}\left\{\Pr\left\{\mathcal{E}_2(1)\right\}\right\}.&\end{aligned} \tag{23}$$

We now analyze each error individually. For each of the above events, we can exploit indicator functions in order to simplify the error analysis (we are also doing this to help build a bridge between this classical proof and the packing lemma approach for the quantum case—projectors in some sense replace indicator functions later on). Recall that an indicator function $I_{\mathcal{A}}\left(x\right)$ is equal to one if $x \in \mathcal{A}$ and equal to zero otherwise. So the following three functions being equal to one or larger then corresponds to error events $\mathcal{E}_0(1)$, $\mathcal{E}_1(1)$, and $\mathcal{E}_2(1)$, respectively:

$$1 - I_{T_\delta^{Y^n}}\left(y^n\right), \tag{24}$$

$$I_{T_\delta^{Y^n}}\left(y^n\right)\left(1 - I_{T_\delta^{Y^n|x^n(1)}}\left(y^n\right)\right), \tag{25}$$

$$\sum_{m'\neq 1} I_{T_\delta^{Y^n}}\left(y^n\right) I_{T_\delta^{Y^n|x^n(m')}}\left(y^n\right). \tag{26}$$

(The last sum of indicators is a consequence of applying the union bound again to the error $\mathcal{E}_2(1)$, which itself is a union of events.)

By exploiting the indicator function from (24), we have that

$$\begin{aligned}&\mathbb{E}_{\mathcal{C}}\left\{\Pr\left\{\mathcal{E}_0\left(1\right)\right\}\right\}\\ &= \mathbb{E}_{X^n(1)}\left\{\mathbb{E}_{Y^n|X^n(1)}\left\{1 - I_{T_\delta^{Y^n}}\left(Y^n\right)\right\}\right\} &(27)\\ &= 1 - \mathbb{E}_{X^n(1),Y^n}\left\{I_{T_\delta^{Y^n}}\left(Y^n\right)\right\} &(28)\\ &= 1 - \mathbb{E}_{Y^n}\left\{I_{T_\delta^{Y^n}}\left(Y^n\right)\right\} &(29)\\ &= \Pr\left\{Y^n \notin T_\delta^{Y^n}\right\} \leq \varepsilon, &(30)\end{aligned}$$

where the first line follows because Y^n is generated according to the conditional distribution $p_{Y^n|X^n}$ and from $X^n(1)$ (since the first message was transmitted) and all other codewords have no role in

the test, so that we marginalize over them. In the last line we have exploited the high probability property of the typical set $T_\delta^{Y^n}$. In the above, we are also exploiting the fact that $\mathbb{E}\{I_{\mathcal{A}}\} = \Pr\{\mathcal{A}\}$. By exploiting the indicator function from (25), we have that

$$\begin{aligned}
&\mathbb{E}_{\mathcal{C}}\{\Pr\{\mathcal{E}_1(1)\}\} \\
&= \mathbb{E}_{X^n(1)}\left\{\mathbb{E}_{Y^n|X^n(1)}\left\{I_{T_\delta^{Y^n}}(Y^n)\left(1 - I_{T_\delta^{Y^n|X^n(1)}}(Y^n)\right)\right\}\right\} && (31)\\
&\leq \mathbb{E}_{X^n(1)}\left\{\mathbb{E}_{Y^n|X^n(1)}\left\{1 - I_{T_\delta^{Y^n|X^n(1)}}(Y^n)\right\}\right\} && (32)\\
&= 1 - \mathbb{E}_{X^n(1)}\left\{\mathbb{E}_{Y^n|X^n(1)}\left\{I_{T_\delta^{Y^n|X^n(1)}}(Y^n)\right\}\right\} && (33)\\
&= \mathbb{E}_{X^n(1)}\left\{\Pr_{Y^n|X^n(1)}\left\{Y^n \notin T_\delta^{Y^n|X^n(1)}\right\}\right\} \leq \varepsilon, && (34)
\end{aligned}$$

where in the last line we have exploited the high probability property of the conditionally typical set $T_\delta^{Y^n|X^n(1)}$. We finally consider the probability of the last kind of error by exploiting the indicator function in (26):

$$\begin{aligned}
&\mathbb{E}_{\mathcal{C}}\{\Pr\{\mathcal{E}_2(1)\}\} \\
&\leq \mathbb{E}_{\mathcal{C}}\left\{\sum_{m'\neq 1} I_{T_\delta^{Y^n}}(y^n) I_{T_\delta^{Y^n|X^n(m')}}(y^n)\right\} && (35)\\
&= \sum_{m'\neq 1} \mathbb{E}_{\mathcal{C}}\left\{I_{T_\delta^{Y^n}}(y^n) I_{T_\delta^{Y^n|X^n(m')}}(y^n)\right\} && (36)\\
&= \sum_{m'\neq 1} \mathbb{E}_{X^n(1),X^n(m'),Y^n}\left\{I_{T_\delta^{Y^n}}(y^n) I_{T_\delta^{Y^n|X^n(m')}}(y^n)\right\} && (37)\\
&= \sum_{m'\neq 1} \sum_{x^n(1),x^n(m'),y^n} p_{X^n}(x^n(1)) p_{X^n}(x^n(m'))\times \\
&\qquad p_{Y^n|X^n}(y^n|x^n(1)) I_{T_\delta^{Y^n}}(y^n) I_{T_\delta^{Y^n|x^n(m')}}(y^n) && (38)\\
&= \sum_{m'\neq 1} \sum_{x^n(m'),y^n} p_{X^n}(x^n(m')) p_{Y^n}(y^n) I_{T_\delta^{Y^n}}(y^n) I_{T_\delta^{Y^n|x^n(m')}}(y^n)\,. && (39)
\end{aligned}$$

The first inequality is from the union bound, and the first equality follows from the way that we select the random code: for every message m, the codewords are selected independently and randomly according to p_{X^n} so that the distribution for the joint random variable $X^n(1)X^n(m')Y^n$ is

$$p_{X^n}(x^n(1))\, p_{X^n}(x^n(m'))\, p_{Y^n|X^n}(y^n|x^n(1)). \tag{40}$$

The second equality follows from marginalizing over $X^n(1)$. Continuing, we have

$$\begin{aligned}
&\leq 2^{-n[H(Y)-\delta]} \sum_{m'\neq 1} \sum_{x^n(m'),y^n} p_{X^n}(x^n(m')) I_{T_\delta^{Y^n|x^n(m')}}(y^n) && (41)\\
&= 2^{-n[H(Y)-\delta]} \sum_{m'\neq 1} \sum_{x^n(m')} p_{X^n}(x^n(m')) \sum_{y^n} I_{T_\delta^{Y^n|x^n(m')}}(y^n) && (42)\\
&\leq 2^{-n[H(Y)-\delta]} 2^{n[H(Y|X)+\delta]} \sum_{m'\neq 1} \sum_{x^n(m')} p_{X^n}(x^n(m')) && (43)\\
&\leq |\mathcal{M}|\, 2^{-n[I(X;Y)-2\delta]}. && (44)
\end{aligned}$$

The first inequality follows from the bound $p_{Y^n}(y^n) I_{T_\delta^{Y^n}}(y^n) \leq 2^{-n[H(Y)-\delta]}$ that holds for typical sequences. The second inequality follows from the cardinality bound $\left|T_\delta^{Y^n|x^n(m')}\right| \leq 2^{n[H(Y|X)+\delta]}$ on the conditionally typical set. The last inequality follows because

$$\sum_{x^n(m')} p_{X^n}(x^n(m')) = 1, \tag{45}$$

$|\mathcal{M}|$ is an upper bound on $\sum_{m' \neq 1} 1 = |\mathcal{M}| - 1$, and by the identity $I(X;Y) = H(Y) - H(Y|X)$. Thus, we can make this error arbitrarily small by choosing the message set size $|\mathcal{M}| = 2^{n[I(X;Y)-3\delta]}$. Putting everything together, we have the following bound on (19):

$$\varepsilon' \equiv 2\varepsilon + 2^{-n\delta}, \tag{46}$$

as long as we choose the message set size as given above. It follows that there exists a particular code with the same error bound on its average error probability. We can then exploit an expurgation argument to convert an average error bound into a maximal one (the expurgation step throws away the worse half of the codewords, guaranteeing a bound of $2\varepsilon'$ on the maximum error probability). Thus, we have shown the achievability of an $(n, C(\mathcal{N}) - \delta', 2\varepsilon')$ channel code for all $\delta' > 0, \varepsilon' \in (0, 1/2)$ and sufficiently large n (where $\delta' = 3\delta$). Finally, as a simple observation, our proof above does not rely on whether the definition of conditional typicality employed is weak or strong.