EL-GY 6063: Information Theory
Lecture 5

1 The Asymptotic Equipartition Property (AEP)

The weak law of large number is stated as follows:

Theorem 1. *Let $Y_i, i \in \mathbb{N}$ be a sequence of independent and identically distributed random variables with $E(Y_i) = \bar{y}$. Then,*

$$P(|\frac{1}{n}\sum_{i=1}^{n} Y_i - \bar{y}| > \epsilon) \to 0 \text{ as } n \to \infty,$$

For any $\epsilon > 0$.

Examples of weak law of large numbers:
1) Assume a fair coin is thrown n times where n is large. Then, the number of heads is very close to $\frac{n}{2}$ with probability 1.
proof. Let Y_i be the indicator of the event that the ith throw was a head, then $\sum_{i=1}^{n} Y_i$ is the total number of heads seen after n throws. From the weak law we know that $\frac{1}{n}\sum_{i=1}^{n} Y_i \approx \mathbb{E}(Y_i) = P(Y_i = 1) \cdot 1 + P(Y_i = 0) \cdot 0 = \frac{1}{2}$ so $\sum_{i=1}^{n} Y_i \approx \frac{n}{2}$.
2) Assume that a fair die is thrown n times where n is large. Then, the number of times 1 is seen is approximately $\frac{n}{6}$. The number of times 1 or 3 are observed is almost $\frac{2n}{6} = \frac{n}{3}$.

Other than the weak law of large numbers, we need to remind ourselves of functions of random variables.

Let X be a random variable and let $f(\cdot)$ be a (measurable) function of X, then, Y=f(X) is a random variable as well.

Example. Let X be the outcome of a throw of a fair die. Then, X has six possible outcomes $\{1, 2, 3, \cdots, 6\}$. Let Y be the indicator of the event that the outcome is even. Then, Y is a function of X. Clearly $P(Y = 1) = P(X \in \{2, 4, 6\}) = P(X = 2) + P(X = 4) + P(X = 6) = \frac{1}{2}$.

Example. In the previous example define $Y = P_X(X)$. Then, Y is a random variable as well. Clearly, $P_X(\cdot)$ only takes the value $\frac{1}{6}$. In other words, the alphabet of Y is $\mathcal{Y} = \{\frac{1}{6}\}$. In this case, $P_Y(\frac{1}{6}) = 1$.

Example. Let X be a Bernoulli random variable with parameter $\frac{1}{3}$. Define $Y = -\log P_X(X)$. Then, the alphabet of Y is $\mathcal{Y} = \{-\log\frac{1}{3}, -\log\frac{2}{3}\}$. Also, $P_Y(-\log\frac{1}{3}) = P(P_X(X) = -\log\frac{1}{3}) = P(X = 1) = \frac{1}{3}$. Similarly, $P_Y(-\log\frac{2}{3}) = \frac{2}{3}$.

The reason that we are interested in the function of the random variable X defined by $Y = -\log P_X(X)$ is that $\mathbb{E}(Y) = \sum_{x\in\mathcal{X}} P_X(x)Y = -\sum_{x\in\mathcal{X}} P_X(x)\log P_X(x) = H(X)$.

Note: The random variable Y is sometimes interpreted as the amount of information revealed by knowing the value of X. In this interpretation $\mathbb{E}(Y)$ is the average amount of information or uncertainty in X which to no surprise is equal to $H(X)$.

Example 1. *i) Let X_i be independent and identically distributed random variables with finite alphabet. Let $Y_i = -\log P_X(X_i)$. Use the Chebychev inequality to bound the probability*

$$P(|\frac{1}{n}\sum_{i=1}^{n} Y_i - H(X)| \geq \epsilon).$$

Solution. Reminder for Chebychev Inequality: Let Z be a random variable and $\epsilon > 0$ a positive real. Then, $P(|Z - \mathbb{E}(Z)| > \epsilon) \leq \frac{Var(Z)}{\epsilon^2}$.

Define $Z = \frac{1}{n}\sum_{i=1}^{n} Y_i$. We want to apply the Chebychev inequality, so we need to find $\mathbb{E}(Z)$ and $Var(Z)$. Note that Y_i are independent and identically distributed since X_i are independent and identically distributed. We have:

$$\mathbb{E}(Z) = \mathbb{E}(\frac{1}{n}\sum_{i=1}^{n} Y_i) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(Y_i) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(-\log P_X(X_i)) = \frac{1}{n}\sum_{i=1}^{n} H(X_i) = H(X).$$

Also,

$$Var(Z) = \mathbb{E}(Z^2) - \mathbb{E}^2(Z) = \frac{1}{n}Var(X) = \frac{1}{n}(\sum_x (P_X(x)\log^2(P_X(x)) - H^2(X))).$$

So, from the Chebychev inequality, we have:

$$P(|\frac{1}{n}\sum_{i=1}^{n} Y_i - H(X)| \geq \epsilon) \leq \frac{\frac{1}{n}\sum_x (P_X(x)\log^2(P_X(x)) - H^2(X))}{\epsilon^2} = \frac{1}{n}\frac{\sum_x (P_X(x)\log^2(P_X(x)) - H^2(X))}{\epsilon^2}.$$

Note that $\frac{1}{n}\sum_{i=1}^{n} Y_i = -\frac{1}{n}\sum_{i=1}^{n}\log P_X(X_i) = -\frac{1}{n}\log\prod_{i=1}^{n} P_X(X_i) = -\frac{1}{n}\log P_{X^n}(X^n)$.

ii) As an example, let X be uniformly distributed among $\{0,1,2\}$. Then, $P_X(x) = \frac{1}{3}, \forall x$. Then, $\sum_x P_X(x)\log^2(P_X(x)) = \log^2\frac{1}{3} = \log^2 3 = H^2(X)$. So, from the above bound,

$$P(|\frac{1}{n}\sum_{i=1}^{n} Y_i - H(X)| \geq \epsilon) = 0, \forall \epsilon.$$

The equation above implies that $\frac{1}{n}\sum_{i=1}^{n} Y_i = -\frac{1}{n}\log P_{X^n}(X^n)$ is a constant (i.e. it is a variable that takes only one value) and the constant value is $H(X) = \log 3$. This is clear from our previous observations as well. Remember that $P_{X^n}(X^n) = (\frac{1}{3})^n$. In other words the random variable $P_{X^n}(X^n)$ only takes one value and that value is $(\frac{1}{3})^n$. Now we have $-\frac{1}{n}\log(\frac{1}{3})^n = \log 3$ with probability 1.

iii) Now let $P_X(0) = P_X(1) = \frac{1}{4}$ and $P_X(2) = \frac{1}{2}$. Then, $\sum_x P_X(x)\log^2(P_X(x)) = \frac{1}{4}(4+4) + \frac{1}{2} = \frac{5}{2}$. On the other hand, $H(X) = \frac{3}{2}$. So,

$$P(|\frac{1}{n}\sum_{i=1}^{n} Y_i - H(X)| \geq \epsilon) \leq \frac{1}{n}\frac{\sum_x (P_X(x)\log^2(P_X(x)) - H^2(X))}{\epsilon^2} = \frac{1}{n}\frac{1}{4\epsilon^2}.$$

Let us take $\epsilon = 0.1$. Then,

$$P(|\frac{1}{n}\sum_{i=1}^{n} Y_i - H(X)| \geq \epsilon) \leq \frac{1}{n}\frac{\sum_x (P_X(x)\log^2(P_X(x)) - H^2(X))}{\epsilon^2} = \frac{25}{n}.$$

We will see that in communication problems we often require this bound to be smaller than numbers such as 0.001. For this, we require $n \approx 25000$. As a result, typicality is a property of sequences with very very large lengths. It is not observed for small blocklengths such as 100 or 200.

Note: One could possibly derive better bounds on the probability using the Chernoff bound.

The Asymptotic Equipartition Property (AEP) is a consequence of the weak law which is widely used in information theory. We are ready to state the AEP theorem.

Theorem 2 (AEP)**.** *Let $X_i, i \in \mathbb{N}$ be a sequence of independent and identically distributed random variables. Then,*

$$P(|-\frac{1}{n}\log P_{X^n}(X^n) - H(X)| > \epsilon) \to 0 \text{ as } n \to \infty,$$

For any $\epsilon > 0$.

In other words, $-\frac{1}{n}\log P_{X^n}(x^n)$ converges to $H(X)$ in probability.

proof. Let $Y_i = -\log P_X(X_i)$. Then $\frac{1}{n}\sum_{i=1}^n Y_i = -\frac{1}{n}\sum_{i=1}^n \log P_X(X_i) = -\frac{1}{n}\log\prod_{i=1}^n P_X(X_i) = -\frac{1}{n}\log P_{X^n}(X^n)$. Also, Y_i are i.i.d. and $\bar{y} = E(Y_i) = E(-\log P_X(X_i)) = H(X)$. So, from the weak law of large numbers:

$$P(|\frac{1}{n}\sum_{i=1}^n Y_i - \bar{y}| > \epsilon) \to 0 \Rightarrow P(|-\frac{1}{n}\log P_{X^n}(X^n) - H(X)| > \epsilon) \to 0.$$

Interpretation of the AEP: The AEP states that for large n, the set of x^n for which $P_{X^n}(x^n) \approx 2^{-nH(X)}$ has probability close to 1. Alternatively, the set of x^n for which this property does not hold has total probability close to 0. So, the space of all x^n can be partitioned into two sets. The first set contains all sequences for which $P_{X^n}(x^n) \approx 2^{-nH(X)}$. This set looks almost uniformly distributed since x^n's all have almost the same probability $2^{-nH(X)}$. The other set is the complement of the first set and contains all other sequences. This set has probability 0. It is as if all of the probability is concentrated in the small set of sequences for which $P_{X^n}(x^n) \approx 2^{-nH(X)}$. For this reason, theorems such as the AEP theorem are sometimes called the concentration of measure theorems.

This small high probability set of sequences with almost uniform distributions is called the weak typical set and is defined below:

Definition 1. *Let $X_i, i \in \mathbb{N}$ be a sequence of independent and identically distributed random variables. The weakly typical set of sequences of length n with respect to P_X is defined as:*

$$\mathcal{A}_\epsilon^n(X) = \{x^n || -\frac{1}{n}\log P_{X^n}(x^n) - H(X)| \le \epsilon\}.$$

The sequence x^n is called an ϵ-(weakly) typical sequence with respect to P_X if $x^n \in \mathcal{A}_\epsilon^n(X)$.

Example. Assume that $X_i, i \in \{1, 2, \cdots, n\}$ be a sequence of independent Bernoulli random variables with parameter $\frac{1}{3}$. Show that the sequence $x^n = 11\cdots100\cdots0$ where the first $\frac{n}{3}$ elements are 1 and the last $\frac{2n}{3}$ are 0 is a typical sequence for any $\epsilon > 0$.

Solution.

$$P_{X^n}(x^n) = \frac{1}{3}^{\frac{n}{3}}\frac{2}{3}^{\frac{2n}{3}} = 2^{\log\frac{1}{3}^{\frac{n}{3}}}2^{\log\frac{2}{3}^{\frac{2n}{3}}} = 2^{\frac{n}{3}\log\frac{1}{3}+\frac{2n}{3}\log\frac{2}{3}} = 2^{n(\frac{1}{3}\log\frac{1}{3}+\frac{2}{3}\log\frac{2}{3})} = 2^{-nH(X)},$$

So, $x^n \in \mathcal{A}_\epsilon^n(X), \forall\epsilon > 0$.

Note that if the position of 0s and 1s is changed by a permutation, the sequence remains typical.

Question: What is the highest probability sequence x^n in the previous example? Is is a typical sequence?

Solution. Take an arbitrary sequence x^n, let the number of ones in the sequence be k and number of zeros $n-k$. Then,

$$P_{X^n}(x^n) = \frac{1}{3}^k\frac{2}{3}^{n-k},$$

Clearly, the probability is larger for smaller k since there power of $\frac{2}{3}$ increases compared to that of $\frac{1}{3}$. So, the most probable sequence is the all zeros sequence. The probability of the sequence is $\frac{2}{3}^n$, so $-\frac{1}{n}\log P_{X^n}(00\cdots0) = -\log P_X(0) = -\log\frac{2}{3} = \log 3 - 1 \approx 1.7$. Note that $H(X) \approx 0.91$. So, for $\epsilon \le 0.7$, the sequence is not typical although it is the most probable sequence. Furthermore, note that $\frac{2}{3}^n$ goes to 0 as $n \to \infty$, so even the most probable sequence has its probability approaching 0 as $n \to \infty$.

The weakly typical set is a high probability set of sequences not a set of high probability sequences!

Question: Assume that $\epsilon = 0.1$. In the previous example, take an arbitrary sequence x^n, let the number of ones in the sequence be k and number of zeros $n-k$. For what values of k, is x^n an ϵ-typical sequence?

Solution. From the previous part, we have:

$$P_{X^n}(x^n) = \frac{1}{3}^k\frac{2}{3}^{n-k} \Rightarrow \log P_{X^n(x^n)} = k\log\frac{1}{3} + (n-k)\log\frac{2}{3} =$$

let $q = \frac{k}{n}$. Then,

$$\log P_{X^n(x^n)} = n(q \log \frac{1}{3} + (1-q) \log \frac{2}{3}) = -n(D(Q_X||P_X) + H(Q_X)).$$

where Q_X is the distribution of a Bernoulli variable with parameter $\frac{k}{n}$. In order for x^n to be ϵ-typical we need $|-\frac{1}{n}\log P_{X^n(x^n)} - H(X)| \leq \epsilon$. This requires $D(Q_X||P_X) \leq \epsilon = 0.1$. From this equation, k can be calculated for any fixed n using a computer program. (final result is close to $0.2 \leq \frac{k}{n} \leq 0.5$).

Theorem 3. *The set $\mathcal{A}_\epsilon^n(X)$ satisfies the following properties:*
i) $P(\mathcal{A}_\epsilon^n(X)) \to 1$ as $n \to \infty$. Alternatively, for any $\epsilon > 0$ and for sufficiently large n, we have $P(\mathcal{A}_\epsilon^n(X)) > 1-\epsilon$,
ii) $|\mathcal{A}_\epsilon^n(X)| \leq 2^{n(H(X)+\epsilon)}$,
iii)$(1-\epsilon)2^{n(H(X)-\epsilon)} \leq |\mathcal{A}_\epsilon^n(X)|$.

proof. i) Follows from the weak law of large numbers as explained before:

$$P(\mathcal{A}_\epsilon^n(X)) = P(X^n \in \mathcal{A}_\epsilon^n(X)) = P(|-\frac{1}{n}\log P_{X^n}(X^n) - H(X)| \leq \epsilon) \to 1,$$

as $n \to \infty$.
ii) Remember that for any set $\mathcal{E}$, we defined $P_X(\mathcal{E}) = P(X \in \mathcal{E}) = \sum_{x\in\mathcal{E}} P_X(x)$. For example, in a throw of a fair die, the set of possible outcomes is $\{1,2,3,4,5,6\}$. Let $\mathcal{E} = \{2,4,6\}$ be the set of even outcomes. Then, $P_X(\mathcal{E}) = P_X(2) + P_X(4) + P_X(6)$.

Similarly, $P(\mathcal{A}_\epsilon^n(X)) = \sum_{x^n\in\mathcal{A}_\epsilon^n(X)} P_{X^n}(x^n)$. we know the following two facts:
1) From the definition of weak typicality, for any element $x^n \in \mathcal{A}_\epsilon^n(X)$, we have $2^{-n(H(X)+\epsilon)} \leq P_{X^n}(x^n) \leq 2^{-n(H(X)-\epsilon)}$.
2) $P(\mathcal{A}_\epsilon^n(X)) \leq 1$.
So,

$$1 \geq P(\mathcal{A}_\epsilon^n(X)) = \sum_{x^n\in\mathcal{A}_\epsilon^n(X)} P_{X^n}(x^n) \geq \sum_{x^n\in\mathcal{A}_\epsilon^n(X)} 2^{-n(H(X)+\epsilon)} = 2^{-n(H(X)+\epsilon)} \sum_{x^n\in\mathcal{A}_\epsilon^n(X)} 1 = 2^{-n(H(X)+\epsilon)}|\mathcal{A}_\epsilon^n(X)|$$
$$\Rightarrow 1 \geq 2^{-n(H(X)+\epsilon)}|\mathcal{A}_\epsilon^n(X)| \Rightarrow 2^{n(H(X)+\epsilon)} \geq |\mathcal{A}_\epsilon^n(X)|.$$

iii) From part 1, $P(\mathcal{A}_\epsilon^n(X)) > 1-\epsilon$ for sufficiently large n. The result in part iii) follows by the same argument as in part ii).

Interpretation of the theorem: Let X be a Bernoulli source with parameter 0.01. Then $H(X) \approx 0.08$. There are a total of 2^n binary sequences. In order to index each sequence by a binary sequence, we require n bits. However, according to the previous theorem approximately $2^{0.08n}$ happen with probability one. The rest of the outcomes have total probability close to 0. Instead of n bits, we only require $0.08n$ bits to index this small set of sequences. This result is used for efficient lossless and lossy source coding as we will see later.

Example. Let X_i be a sequence of i.i.d Bernoulli($\frac{1}{2}$) random variables. Let k be the number of ones in x^n. For what values of k is x^n an ϵ-typical sequence for an arbitrary $\epsilon > 0$?

Solution. $P_{X^n}(x^n) = \frac{1}{2}^n$ for any value of k. So, all sequences are weakly typical. In this case, for any ϵ, we have $|\mathcal{A}_\epsilon^n(X)| = 2^{nH(X)} = 2^n$.

Next, we define joint typicality for a pair of random variables.

Definition 2. *Let $(X_i, Y_i), i \in \mathbb{N}$ be a sequence of pairs independent and identically distributed random variables. The weakly typical set of sequences of length n with respect to $P_{X,Y}$ is defined as:*

$$\mathcal{A}_\epsilon^n(X,Y) = \{(x^n,y^n) \in \mathcal{A}_\epsilon^n(X) \times \mathcal{A}_\epsilon^n(Y) \big| |-\frac{1}{n}\log P_{X^n,Y^n}(x^n,y^n) - H(X,Y)| \leq \epsilon\}.$$

The sequence pair (x^n, y^n) is called an ϵ-(weakly) typical sequence with respect to $P_{X,Y}$ if $(x^n,y^n) \in \mathcal{A}_\epsilon^n(X,Y)$.

Example. Let $(X_i, Y_i), i \in \mathbb{N}$ be a sequence of pairs independent and identically distributed random variables distributed according to $P_X P_Y$ (i.e. X_i and Y_i are independent of each other.). Then,

$$\mathcal{A}_{\frac{\epsilon}{2}}(X) \times \mathcal{A}_{\frac{\epsilon}{2}}(Y) \subseteq \mathcal{A}_\epsilon^n(X,Y) \subseteq \mathcal{A}_\epsilon(X) \times \mathcal{A}_\epsilon(Y).$$

proof. The right hand side follows from the definition. For the left hand side we have:

$$\begin{aligned}
&(x^n, y^n) \in \mathcal{A}_{\frac{\epsilon}{2}}(X) \times \mathcal{A}_{\frac{\epsilon}{2}}(Y) \iff |-\frac{1}{n}\log P_{X^n}(x^n) - H(X)| \leq \frac{\epsilon}{2} \quad \& \quad |-\frac{1}{n}\log P_{Y^n}(y^n) - H(Y)| \leq \frac{\epsilon}{2} \\
&\Rightarrow |-\frac{1}{n}\log P_{X^n}(x^n) - H(X)| + |-\frac{1}{n}\log P_{Y^n}(y^n) - H(Y)| \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon \\
&\Rightarrow |-\frac{1}{n}\log P_{X^n}(x^n) - H(X) - \frac{1}{n}\log P_{Y^n}(y^n) - H(Y)| \leq \epsilon \\
&\Rightarrow |-\frac{1}{n}\log P_{X^n}(x^n) P_{Y^n}(y^n) - H(X,Y)| \leq \epsilon \\
&\Rightarrow |-\frac{1}{n}\log P_{X^n,Y^n}(x^n, y^n) - H(X,Y)| \leq \epsilon \\
&\Rightarrow (x^n, y^n) \in \mathcal{A}_\epsilon^n(X,Y).
\end{aligned}$$

Note that $\mathcal{A}_\epsilon(X) \times \mathcal{A}_\epsilon(Y) \not\subseteq \mathcal{A}_\epsilon^n(X,Y)$. Provide an example.

Example. Let $X = Y$ with probability 1. Then, $\mathcal{A}_\epsilon^n(X,Y) = \{(a^n, a^n) | a^n \in \mathcal{A}_\epsilon^n(X)\}$.

proof. First note that $H(X,Y) = H(X) + H(Y|X) = H(X)$. Let $a^n \in \mathcal{A}_\epsilon^n(X)$. Then, $P_{X^n,Y^n}(a^n, a^n) = P_{X^n}(a^n) P_{Y^n|X^n}(a^n|a^n) = P_{X^n}(a^n)$. On the other hand since $a^n \in \mathcal{A}_\epsilon^n(X)$ we know that $|-\frac{1}{n}\log P_{X^n}(a^n) - H(X)| \leq \epsilon$. So, $|-\frac{1}{n}\log P_{X^n,Y^n}(a^n, a^n) - H(X,Y)| \leq \epsilon$.

Now take (a^n, b^n) such that $a^n \neq b^n$. Then $P_{X^n,Y^n} = 0$. So, $(a^n, b^n) \notin \mathcal{A}_\epsilon^n(X,Y)$.

Similarly, if $a^n \notin \mathcal{A}_\epsilon^n(X)$, then $(a^n, a^n) \notin \mathcal{A}_\epsilon^n(X,Y)$.

Note that in this case $|\mathcal{A}_\epsilon^n(X,Y)| = |\mathcal{A}_\epsilon^n(X)|$.

Theorem 4. *The set $\mathcal{A}_\epsilon^n(X,Y)$ satisfies the following properties:*
i) $P(\mathcal{A}_\epsilon^n(X,Y)) \to 1$ as $n \to \infty$. Alternatively, for any $\epsilon > 0$ and for sufficiently large n, we have $P(\mathcal{A}_\epsilon^n(X,Y)) > 1 - \epsilon$,
ii) $|\mathcal{A}_\epsilon^n(X,Y)| \leq 2^{n(H(X,Y)+\epsilon)}$,
iii)$(1-\epsilon)2^{n(H(X,Y)-\epsilon)} \leq |\mathcal{A}_\epsilon^n(X,Y)|$.

Proof is similar to the single variable case.

Exercise: Prove that $|\mathcal{A}_\epsilon^n(X,Y)| = |\mathcal{A}_\epsilon^n(X)|$ if and only if Y is a function of X. Furthermore, $|\mathcal{A}_\epsilon^n(X,Y)| = |\mathcal{A}_\epsilon^n(X)| = |\mathcal{A}_\epsilon^n(Y)|$ if and only if X is a one-to-one function of Y.

Joint typicality for vectors of random variables:

Definition 3. *Let $(X_{1,i}, X_{2,i}, \cdots, X_{m,i}), i \in \mathbb{N}$ be a sequence of vectors of independent and identically distributed random variables. The weakly typical set of sequences of length n with respect to $P_{X_1,X_2,\cdots,X_m}$ is defined as:*

$$\mathcal{A}_\epsilon^n(X_1, X_2, \cdots, X_m) = \{(x_1^n, x_2^n, \cdots, x_m^n) || -\frac{1}{n}\log P_{X_1^n, X_2^n, \cdots, x_m^n}(x_1^n, x_2^n, \cdots, x_m^n) - H(X_1, X_2, \cdots, X_m)| \leq \epsilon\}.$$

Example. Let $(X_{1,i}, X_{2,i}, \cdots, X_{m,i}), i \in \mathbb{N}$ be a sequence of pairs independent and identically distributed random variables distributed according to $P_{X_1} P_{X_2} \cdots P_{X_m}$. Then,

$$\mathcal{A}_{\frac{\epsilon}{m}}(X_1) \times \mathcal{A}_{\frac{\epsilon}{m}}(X_2) \cdots \times \mathcal{A}_\epsilon^n(X_m) \subseteq \mathcal{A}_\epsilon^n(X_1, X_2, \cdots, X_m).$$

Proof is similar to the previous case.

The conditional typical set is defined as follows:

Definition 4. *Let $(X_i, Y_i), i \in \mathbb{N}$ be a sequence of pairs independent and identically distributed random variables. The set of conditionally weakly typcial sequences of length n with respect to the sequence y^n and distribution $P_{X|Y}$ is defined as:*

$$\mathcal{A}_\epsilon^n(X|Y^n = y^n) = \{x^n | (x^n, y^n) \in \mathcal{A}_\epsilon^n(X,Y)\}.$$

Lemma 1. *If $y^n \notin \mathcal{A}_\epsilon^n(Y)$, then $\mathcal{A}_\epsilon^n(X|Y^n = y^n) = \phi$. Otherwise If $y^n \in \mathcal{A}_\epsilon^n(Y)$, then,*
i) $P(\mathcal{A}_\epsilon^n(X|Y^n = y^n)|Y^n = y^n) \to 1$ as $n \to \infty$.
ii) $|\mathcal{A}_\epsilon^n(X|Y^n = y^n)| \leq 2^{n(H(X|Y)+2\epsilon)}$,
iii)$(1-\epsilon)2^{n(H(X|Y)-2\epsilon)} \leq |\mathcal{A}_\epsilon^n(X|Y^n = y^n)|$.

Example. Let $(X_i, Y_i), i \in \mathbb{N}$ be a sequence of pairs independent and identically distributed random variables distributed according to $P_X P_Y$ (i.e. X_i and Y_i are independent of each other.). Let $y^n \notin \mathcal{A}_{\frac{\epsilon}{2}}^n(Y)$. Then,

$$\mathcal{A}_{\frac{\epsilon}{2}}^n(X) \subseteq \mathcal{A}_\epsilon^n(X|Y^n = y^n).$$

proof.

$$\begin{aligned} &x^n, y^n \in \mathcal{A}_{\frac{\epsilon}{2}}^n(X) \times \mathcal{A}_{\frac{\epsilon}{2}}^n(Y) \Rightarrow (x^n, y^n) \in \mathcal{A}_\epsilon^n(X, Y) \\ &\Rightarrow x^n \in \mathcal{A}_\epsilon^n(X|Y^n = y^n). \end{aligned}$$

Alternatively, in some textbooks, the conditional typical set is defined as $\mathcal{A}_\epsilon^n(X|Y^n = y^n) = \{x^n \big|\big| -\frac{1}{n}\sum_{i=1}^n \log P_{X^n|Y^n}(x^n|y^n) - H(X|Y)| \leq \epsilon\}$. In this case, there are minor modification in the results provided above but most remain unchanged.