Information Theory (5XSE0)
Ch.3: Asymptotic Equipartition Property

Hamdi Joudeh

TU/e (Q3 2020-2021)

Reading

- Cover & Thomas, Ch. 3 (excluding 3.3) and 7.6.
- Supplementary: Gallager, 3.1; MacKay, Ch. 4.

Before we start

- We use X^n to briefly denote the random sequence $(X_1, X_2, \ldots, X_n)$. Similarly, we use x^n to denote the sequence $(x_1, x_2, \ldots, x_n)$. Therefore, $p(x^n)$ is the pmf of X^n.
- All logs in this chapter have base 2. Also, all codes are binary (i.e. $\mathcal{D} = \{0, 1\}$).

Consider a bent coin represented by a binary random variable X with $p(0) = 0.1$ and $p(1) = 0.9$, where 0 and 1 represent tails and heads, respectively. An outcome of n independent tosses of this coin is given by the sequence $x^n \equiv (x_1, x_2, \ldots, x_n)$, representing one of 2^n possible outcomes for this experiment. How many ones (heads) and zeros (tails) do we expect to see in x^n? From the law of large numbers, we know that $\frac{1}{n}\sum_{i=1}^n x_i$ will be close to $\mathbb{E}X = 0.9$, especially if n is large enough. It is therefore safe to claim that for large enough n (many tosses), a typical x^n will have roughly $0.9n$ ones and $0.1n$ zeros. How many typical sequences do we have? What is the total probability of these typical sequences? How does this generalize to other discrete random variables? These questions will be answered in this chapter.

1 Asymptotic Equipartition Property Theorem

The Asymptotic Equipartition Property AEP is the information-theoretic analog of the law of large numbers (LLN). Suppose that we have a sequence X^n of n i.i.d. copies of a random variable X. The LLN states that the empirical expectation $\frac{1}{n}\sum_{i=1}^n X_i$ is close to the true expectation $\mathbb{E}X$ for large n. Similarly, the AEP states that the empirical entropy $\frac{1}{n}\sum_{i=1}^n \log\frac{1}{p(X_i)} = \frac{1}{n}\log\frac{1}{p(X^n)}$ is close to the true entropy $H(X)$ for large n. As a result, almost all sequences that we expect see (i.e. typical sequences) have a probability of $p(X^n) \approx 2^{-nH(X)}$; the set of typical sequence has roughly $2^{nH(X)}$ elements; and all remaining sequences have negligible probability. Recall that the total number of sequences is $|\mathcal{X}|^n = 2^{n\log|\mathcal{X}|}$.

In cases where $H(X)$ is much smaller than $\log|\mathcal{X}|$, the set of typical sequences can be considerably smaller than the set of all sequences. Data compression follows as a consequence: we need roughly $nH(X)$ bits to represent the set of typical sequences, while remaining can be assigned longer codewords (with little consequence). Before formalizing all the above statements, we discuss the weak law of large numbers.

1.1 Weak Law of Large Numbers

Consider a random variable $X \sim p(x)$ with a finite expected value (i.e. $-\infty < \mathbb{E}X < \infty$). Now let $X^n = (X_1, X_2, \ldots, X_n)$ a sequence of i.i.d. random variables, where $X_i \sim p(x)$ for all i. The empirical

mean of X^n is defined as

$$\overline{X}_n \equiv \frac{1}{n}\sum_{i=1}^{n} X_i.$$

The weak law of large numbers states that the empirical mean $\overline{X}_n$ converges in probability to $\mathbb{E}X$.

Theorem 1. *Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim p(x)$. For every $\epsilon > 0$, we have*

$$\lim_{n\to\infty} \mathbb{P}\Big\{ \big|\overline{X}_n - \mathbb{E}X\big| > \epsilon \Big\} = 0. \tag{1}$$

Note that the sample mean $\overline{X}_n$ is a random variable, as it depends on the actual realizations of its constituent random variables. The true mean $\mathbb{E}X$ is fixed. Theorem 1 states that the probability that the sample mean will deviate from the true mean by more than ϵ diminishes for a large enough sample. The convergence in the above theorem is called convergence in probability, and will be denote by:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\text{p}} \mathbb{E}X.$$

Corollary 1. *For $g(X)$ such that $\mathbb{E}g(X) < \infty$, we have*

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) \xrightarrow{\text{p}} \mathbb{E}g(X).$$

This holds as $g(X_1), g(X_2), \ldots$ are also i.i.d. random variables.

1.2 The Asymptotic Equipartition Property Theorem

The AEP is given through the following theorem.

Theorem 2. (AEP). *If $X_1, X_2, \ldots, X_n$ are i.i.d. $\sim p(x)$, then*

$$-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) \xrightarrow{\text{p}} H(X).$$

Proof. The proof relies on taking $g(X_i) = -\log p(X_i)$ and invoking Corollary 1. In particular, we write:

$$\begin{aligned}
-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) &= -\frac{1}{n}\log \prod_{i=1}^{n} p(X_i) \\
&= -\frac{1}{n}\sum_{i=1}^{n} \log p(X_i) \\
&\xrightarrow{\text{p}} -\mathbb{E}\log p(X) \\
&= H(X).
\end{aligned}$$

This completes the proof of the AEP. □

Using the definition of convergence in probability, Theorem 2 can be stated more elaborately as follows. For any $\delta > 0$, there exists an integer n_δ such that for all $n \geq n_\delta$, we have

$$1 - \mathbb{P}\left\{\left|-\frac{1}{n}\log p(X^n) - H(X)\right| > \epsilon\right\} = \mathbb{P}\left\{\left|-\frac{1}{n}\log p(X^n) - H(X)\right| \leq \epsilon\right\} \geq 1 - \delta. \tag{2}$$

Equivalently, we may say that for large enough n, we have:

$$\mathbb{P}\left\{H(X) - \epsilon \leq -\frac{1}{n}\log p(X^n) \leq H(X) + \epsilon\right\} = \mathbb{P}\left\{2^{-n(H(X)+\epsilon)} \leq p(X^n) \leq 2^{-n(H(X)-\epsilon)}\right\} \approx 1.$$

The above implies that as n grows large, it is likely to observe a sequence $X^n = x^n$ which has a probability given by $p(x^n) \approx 2^{-n(H(X)\pm\epsilon)}$. It follows that *typical* outcomes are uniformly distributed over a set of roughly $2^{nH(X)}$ sequences. This is known as the *typical set*. In the words of Cover & Thomas, "*Almost all events are almost equally surprising*", which is especially the case for large n.

2 Typical Sets

Definition 1. For a pmf $p(x)$, integer n, and any $\epsilon > 0$, the typical set $\mathcal{A}_\epsilon^{(n)}$ is a set of sequences in $\mathcal{X}^n$ defined as

$$\mathcal{A}_\epsilon^{(n)} \equiv \left\{ x^n \in \mathcal{X}^n : 2^{-n(H(X)+\epsilon)} \leq p(x^n) \leq 2^{-n(H(X)-\epsilon)} \right\}.$$

It follows from the above definition that the typical set $\mathcal{A}_\epsilon^{(n)}$ is equivalently defined as

$$\mathcal{A}_\epsilon^{(n)} = \left\{ x^n \in \mathcal{X}^n : \left| -\frac{1}{n} \log p(x^n) - H(X) \right| \leq \epsilon \right\}. \tag{3}$$

Some key properties of the typical set are stated in the following theorem.

Theorem 3. *The typical set $\mathcal{A}_\epsilon^{(n)}$ satisfies the following properties:*

1. *If $x^n \in \mathcal{A}_\epsilon^{(n)}$, then $H(X) - \epsilon \leq -\frac{1}{n} \log p(x^n) \leq H(X) + \epsilon$.*
2. *$\mathbb{P}\left\{ \mathcal{A}_\epsilon^{(n)} \right\} \geq 1 - \epsilon$ for large enough n.*
3. *$\left| \mathcal{A}_\epsilon^{(n)} \right| \leq 2^{n(H(X)+\epsilon)}$, where $\left| \mathcal{A}_\epsilon^{(n)} \right|$ is the size of $\mathcal{A}_\epsilon^{(n)}$.*
4. *$\left| \mathcal{A}_\epsilon^{(n)} \right| \geq (1-\epsilon) 2^{n(H(X)-\epsilon)}$, for large enough n.*

Intuition. Here are some intuitions related to the above properties of the typical set:

1. Say we observed a sequence x^n by sampling X independently n times. If x^n is typical, then the "empirical" entropy given by $-\frac{1}{n} \log p(x^n) = \frac{1}{n} \sum_{i=1}^n \log \frac{1}{p(x_i)}$ will be close to the entropy $H(X)$.
2. As n grows large, a sequence x^n observed by i.i.d. sampling of X will very likely be typical. An implication is that the above entropy estimate will be close to the true entropy with high probability.
3. The third and fourth properties imply that the size of the typical set is $\approx 2^{nH(X)}$ for large n.

Proof. The proofs for the above properties are given as follows:

1. The first property follows immediately from (3).
2. Note that the probability of the typical set $\mathbb{P}\{\mathcal{A}_\epsilon^{(n)}\}$ is equal to $\mathbb{P}\{X^n \in \mathcal{A}_\epsilon^{(n)}\}$, i.e. the probability that a random sequence X^n is typical. With this in mind, we proceed as follows:

$$\begin{aligned} \mathbb{P}\left\{ X^n \in \mathcal{A}_\epsilon^{(n)} \right\} &= \mathbb{P}\left\{ \left| -\frac{1}{n} \log p(X^n) - H(X) \right| \leq \epsilon \right\} \\ &\geq 1 - \epsilon, \text{ for all } n \geq n_\epsilon \end{aligned} \tag{4}$$

   where (4) follows from (2) by setting $\delta = \epsilon$.

3. This property is shown as follows:

$$\begin{aligned} 1 &= \sum_{x^n \in \mathcal{X}^n} p(x^n) \\ &\geq \sum_{x^n \in \mathcal{A}_\epsilon^{(n)}} p(x^n) \\ &\geq \sum_{x^n \in \mathcal{A}_\epsilon^{(n)}} 2^{-n(H(X)+\epsilon)} \\ &= \left| \mathcal{A}_\epsilon^{(n)} \right| 2^{-n(H(X)+\epsilon)}. \end{aligned}$$

   This immediately implies that $\left| \mathcal{A}_\epsilon^{(n)} \right| \leq 2^{n(H(X)+\epsilon)}$.

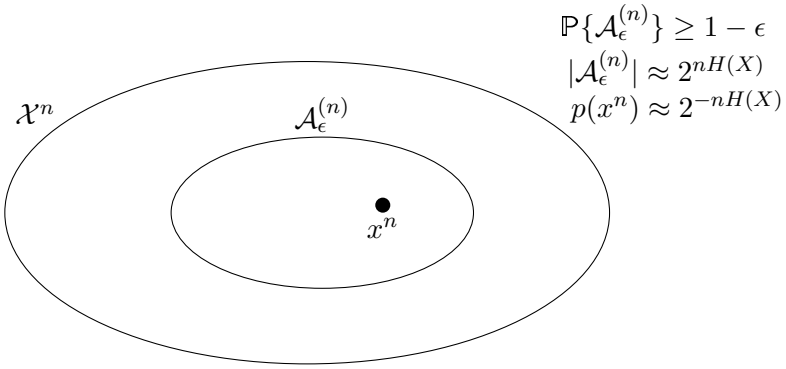
4. From the above, we know that for all $n \geq n_\epsilon$, we have $\mathbb{P}\left\{\mathcal{A}_\epsilon^{(n)}\right\} \geq 1-\epsilon$. It follows that:

$$\begin{aligned} 1-\epsilon &\leq \mathbb{P}\left\{\mathcal{A}_\epsilon^{(n)}\right\} \\ &= \sum_{x^n \in \mathcal{A}_\epsilon^{(n)}} p(x^n) \\ &\leq \sum_{x^n \in \mathcal{A}_\epsilon^{(n)}} 2^{-n(H(X)-\epsilon)} \\ &= \left|\mathcal{A}_\epsilon^{(n)}\right| 2^{-n(H(X)-\epsilon)}. \end{aligned} \tag{5}$$

This immediately implies that $\left|\mathcal{A}_\epsilon^{(n)}\right| \geq (1-\epsilon)2^{n(H(X)-\epsilon)}$.

□

The typical set is illustrated below. As n grows large, most of the probability concentrates in $\mathcal{A}_\epsilon^{(n)}$. This observation is exploited next to carry out compression.



3 Application to Data Compression

Now let us consider the problem of compressing an i.i.d. random sequence X^n, which we call a *source*. By compression we mean finding a description of X^n which is short on average. We will focus on binary descriptions: each possible source sequence x^n is encoded into a binary codeword $C(x^n)$ of length $l(x^n)$. The goal is to make the expected codeword length $\mathbb{E}\left[l(X^n)\right] \equiv \sum_{x^n \in \mathcal{X}^n} p(x^n) l(x^n)$ as small as possible.

An important requirement is non-singularity to avoid possible confusion when decoding (i.e. unique decodability). For this, we must have $x^n \neq \bar{x}^n \implies C(x^n) \neq C(\bar{x}^n)$. This is known as *lossless* compression (or lossless sources coding), as opposed to *lossy* compression where some confusion may be allowed.

Intuition. Think of X^n as text file, which is modeled by a random sequence to represent our lack of knowledge of its content (we may know its distribution from many previously observed files of the same category, e.g. English text). Encoding X^n into a binary codeword allows us to store X^n on a hard drive. Compressing X^n using a short binary description allows us to store X^n with a reduced hard-drive space requirement. This is what a compression software does (e.g. Zip or RAR), and such compression must be lossless as we do not wish to loose any text after decompression. Compression of images, however, could be lossy: we may reduce the resolution to store more photos. Here we focus on lossless compression.

We saw in the previous chapter that we can compress X^n using a source code (e.g. Huffman code) with expected length satisfying $nH(X) \leq \mathbb{E}\left[l(X^n)\right] < nH(X)+1$. Here we focus on an alternative coding scheme based on typical sequences and typical sets.

3.1 Raw bit content

Since X^n takes values in $\mathcal{X}^n$, which in turn has a size of $m \equiv |\mathcal{X}^n| = |\mathcal{X}|^n$, it is sufficient to use no more than $\lceil \log m \rceil = \lceil n \log |\mathcal{X}| \rceil < 1 + n \log |\mathcal{X}|$ bits to describe X^n. Note that the ceiling function $\lceil \cdot \rceil$ is used since the quantity $n \log |\mathcal{X}|$ may not be integer. This raw bit coding is done as follows. We assign a unique index from 0 to $m-1$ to each possible source sequence, from which $\mathcal{X}^n$ may be written as:

$$\mathcal{X}^n = \big\{ x^n(0), x^n(1), \ldots, x^n(m-1) \big\} \tag{6}$$

where $x^n(i)$ is the i-th source sequence. Each sequence $x^n(i)$ is then mapped into a corresponding codeword $C(x^n(i))$ of length $\lceil n \log |\mathcal{X}| \rceil$, given by the binary representation of the index m, that is:

i	$C(x^n(i))$
0	$00\cdots000$
1	$00\cdots001$
2	$00\cdots010$
3	$00\cdots011$
$\vdots$	
$m-1$	$11\cdots111$

All codewords have the same length $\lceil n \log |\mathcal{X}| \rceil$, and hence the average length of this description is less than $n \log |\mathcal{X}| + 1$ bits. For large n, this becomes close to $\log |\mathcal{X}|$ bits per source symbol.

Now let us consider a case where the joint pmf $p(x^n)$ is such that sequences in some subset $\mathcal{A} \subset \mathcal{X}^n$ are uniformly distributed, while sequences in the complement set $\mathcal{A}^{\mathrm{c}}$ (i.e. not in $\mathcal{A}$) have zero probability:

$$p(x^n) = \begin{cases} \frac{1}{|\mathcal{A}|}, \text{ for all } x^n \in \mathcal{A} \\ 0, \text{ for all } x^n \in \mathcal{A}^{\mathrm{c}}. \end{cases}$$

In this case, we only need to assign codewords to sequences in $\mathcal{A}$, as all remaining sequences never occur. For this, we need no more than $n \log |\mathcal{A}| + 1$ bits, which is roughly $\log |\mathcal{A}|$ bits per symbol when n is large.

We now return to our i.i.d. sequence X^n with an arbitrary joint pmf of the form $p(x^n) = \prod_{i=1}^n p(x_i)$. The AEP tells us that as n grows large, $p(x^n)$ becomes such that sequences in the typical set $\mathcal{A}_\epsilon^{(n)}$ are almost uniformly distributed, while remaining sequences in the complement ${\mathcal{A}_\epsilon^{(n)}}^{\mathrm{c}}$ have a probability of almost zero. Since $|\mathcal{A}_\epsilon^{(n)}| \approx 2^{nH(X)}$, we need roughly $nH(X)$ bits on average to described sequences in the typical set. Here, however, we cannot simply ignore sequences in ${\mathcal{A}_\epsilon^{(n)}}^{\mathrm{c}}$, as the event of encountering one such sequence remains possible, despite being highly improbable. Sequences in ${\mathcal{A}_\epsilon^{(n)}}^{\mathrm{c}}$ are assigned longer codewords, guaranteeing successful decoding, while only marginally affecting the average description length. The average description length is predominantly determined by the shorter codewords assigned to $\mathcal{A}_\epsilon^{(n)}$, and will remain around $H(X)$ bits per symbol.

3.2 Typicality compression

Let us partition the set of all possible source sequences $\mathcal{X}^n$ into two sets: the typical set $\mathcal{A}_\epsilon^{(n)}$ and its complement ${\mathcal{A}_\epsilon^{(n)}}^{\mathrm{c}} \equiv \mathcal{X}^n \setminus \mathcal{A}_\epsilon^{(n)}$. We can assume, without loss of generality, that the indexing in (6) is done such that $\mathcal{A}_\epsilon^{(n)}$ is given by the first $|\mathcal{A}_\epsilon^{(n)}|$ sequences, which implies that the next $|\mathcal{X}|^n - |\mathcal{A}_\epsilon^{(n)}|$ sequences form the complement ${\mathcal{A}_\epsilon^{(n)}}^{\mathrm{c}}$. A sequence x^n is then encoded into binary codeword of a length that depends on whether x^n is in $\mathcal{A}_\epsilon^{(n)}$ or in ${\mathcal{A}_\epsilon^{(n)}}^{\mathrm{c}}$. This is carried out as follows:

- Sequences in $\mathcal{A}_\epsilon^{(n)}$ are indexed and then each is assigned a binary representation. Since we have at most $2^{n(H(X)+\epsilon)}$ sequences in $\mathcal{A}_\epsilon^{(n)}$, it is sufficient to use $\lceil n(H(X)+\epsilon) \rceil < n(H(X)+\epsilon)+1$ bits. For this purpose, we may use the $\lceil n(H(X)+\epsilon) \rceil$ least significant bits in the raw bit code shown in the

above table. We then prefix each of the resulting binary sequences by a 0, from which we obtain codewords of length less than $n(H(X)+\epsilon)+2$ as follows:

$$x^n \to C'(x^n) \to \big(0, C'(x^n)\big)$$

where x^n is in $\mathcal{A}_\epsilon^{(n)}$, $C'(x^n)$ is the corresponding binary representation of length no more than $n(H(X)+\epsilon)+1$, and $\big(0, C'(x^n)\big)$ is the resulting binary codeword of length $< n(H(X)+\epsilon)+2$.

- Sequences in $\mathcal{A}_\epsilon^{(n)^{\mathrm{c}}}$ are indexed and assigned binary representations as well. Here it is sufficient to use $\lceil n \log|\mathcal{X}| \rceil \le n\log|\mathcal{X}|+1$ bits, since $\mathcal{A}_\epsilon^{(n)^{\mathrm{c}}}$ is a subset of $\mathcal{X}^n$ and $|\mathcal{X}|^n$ is the total number of sequences. For this purpose, we may directly use the corresponding raw bit code in the table. We then prefix each of these binary sequences by a 1, from which we obtain binary codewords of length less than $n(\log|\mathcal{X}|)+2$ as follows:

$$x^n \to C''(x^n) \to \big(1, C''(x^n)\big)$$

where x^n is in $\mathcal{A}_\epsilon^{(n)^{\mathrm{c}}}$, $C''(x^n)$ is the corresponding binary representation of length no more than $n\log|\mathcal{X}|+1$, and $\big(0, C''(x^n)\big)$ is the resulting codeword of length $< n\log|\mathcal{X}|+2$.

- The above code is one-to-one and is easily decodable: the first bit acts as a flag, indicating the length of the binary sequence that follows, and hence whether it is of type $C'(x^n)$ or type $C''(x^n)$. This allows us to decode the codeword into the corresponding sequence x^n with no confusion.

Next, we analyze the average codeword length for the above code. We have:

$$\begin{aligned}
\mathbb{E}\left[l(X^n)\right] &= \sum_{x^n\in\mathcal{X}^n} p(x^n)l(x^n) \\
&= \sum_{x^n\in\mathcal{A}_\epsilon^{(n)}} p(x^n)l(x^n) + \sum_{x^n\in\mathcal{A}_\epsilon^{(n)^{\mathrm{c}}}} p(x^n)l(x^n) \\
&\le \sum_{x^n\in\mathcal{A}_\epsilon^{(n)}} p(x^n)(n(H(X)+\epsilon)+2) + \sum_{x^n\in\mathcal{A}_\epsilon^{(n)^{\mathrm{c}}}} p(x^n)(n\log|\mathcal{X}|+2) \\
&= \mathbb{P}\left\{\mathcal{A}_\epsilon^{(n)}\right\}(n(H(X)+\epsilon)+2) + \mathbb{P}\left\{\mathcal{A}_\epsilon^{(n)^{\mathrm{c}}}\right\}(n\log|\mathcal{X}|+2) \\
&= \mathbb{P}\left\{\mathcal{A}_\epsilon^{(n)}\right\} n(H(X)+\epsilon) + \left(1-\mathbb{P}\left\{\mathcal{A}_\epsilon^{(n)}\right\}\right)(n\log|\mathcal{X}|)+2) \\
&\le n(H(X)+\epsilon) + \epsilon(n\log|\mathcal{X}|)+2, \quad \text{for all } n\ge n_\epsilon \\
&= n(H(X)+\epsilon')
\end{aligned}$$

where $\epsilon' = \epsilon + \epsilon\log|\mathcal{X}| + \frac{2}{n}$. Note that ϵ' can be made as small as desired by an appropriate choice of ϵ followed by an appropriate choice of n. Therefore, we have proved the following theorem.

Theorem 4. *Let X^n be i.i.d. $\sim p(x)$, and let $\epsilon > 0$ be a small real number. There exists a lossless binary source code for X^n such that*

$$\frac{1}{n}\mathbb{E}\left[l(X^n)\right] \le H(X)+\epsilon. \tag{7}$$

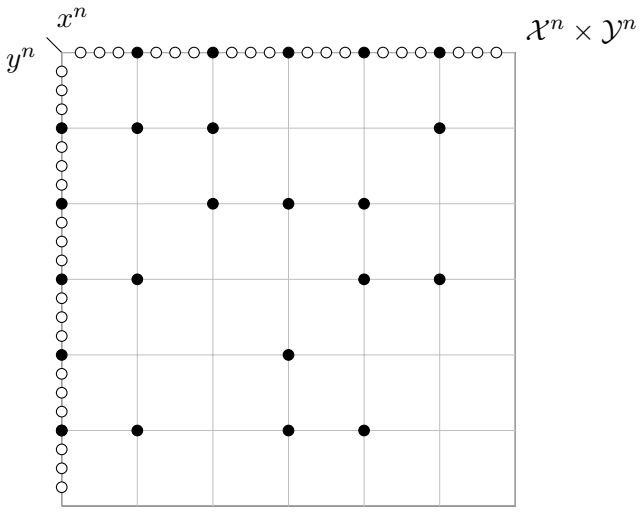
4 Jointly Typical Sets

We now extend the notion of typicality to pairs of jointly distributed sequences. Recall that Cartesian product of $\mathcal{X}^n$ and $\mathcal{Y}^n$ is defined as $\mathcal{X}^n\times\mathcal{Y}^n \equiv \big\{(x^n,y^n) : x^n\in\mathcal{X}^n, y^n\in\mathcal{Y}^n\big\}$.

Definition 2. For a joint pmf $p(x, y)$, integer n, and any $\epsilon > 0$, the jointly typical set $\mathcal{A}_\epsilon^{(n)}$ is a set of sequences in $\mathcal{X}^n \times \mathcal{Y}^n$ defined as

$$\mathcal{A}_\epsilon^{(n)} \equiv \Big\{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : 2^{-n(H(X)+\epsilon)} \leq p(x^n) \leq 2^{-n(H(X)-\epsilon)} \\ 2^{-n(H(Y)+\epsilon)} \leq p(y^n) \leq 2^{-n(H(Y)-\epsilon)} \\ 2^{-n(H(X,Y)+\epsilon)} \leq p(x^n, y^n) \leq 2^{-n(H(X,Y)-\epsilon)}\Big\}.$$

According to the above definition, if x^n is a typical sequence, i.e. $\left|-\frac{1}{n}\log p(x^n) - H(X)\right| \leq \epsilon$, and y^n is a typical sequence, i.e. $\left|-\frac{1}{n}\log p(y^n) - H(Y)\right| \leq \epsilon$, then the pair (x^n, y^n) is not necessarily jointly typical. An additional condition is required, which is $\left|-\frac{1}{n}\log p(x^n, y^n) - H(X,Y)\right| \leq \epsilon$.



Joint typicality is illustrated above. Sequences marked in black on the upper edge are typical x^ns, while sequences marked in black on the left edge are typical y^ns. A pair (x^n, y^n) for which both x^n and y^n are typical is not necessarily jointly typical. Jointly typical (x^n, y^n)s are marked in black in the interior.

Note that the size of the jointly typical set $\mathcal{A}_\epsilon^{(n)}$ is bounded as

$$(1-\epsilon)2^{n(H(X,Y)-\epsilon)} \leq \left|\mathcal{A}_\epsilon^{(n)}\right| \leq 2^{n(H(X,Y)+\epsilon)}. \tag{8}$$

where the lower bound holds for large enough n. This can be shown by following the same approach used to prove properties 3 and 4 in Theorem 3 (try this!). We are now ready to present the joint AEP theorem.

Theorem 5. (Joint AEP). *Let (X^n, Y^n) be i.i.d. with distribution $p(x^n, y^n) = \prod_{i=1}^n p(x_i, y_i)$, and $(\tilde{X}^n, \tilde{Y}^n)$ be i.i.d with distribution $p(x^n)p(y^n)$. Note that $\tilde{X}^n$ and $\tilde{Y}^n$ are independent, but have the same distributions as X^n and Y^n, respectively. The following points hold:*

1. $\mathbb{P}\left\{(X^n, Y^n) \in \mathcal{A}_\epsilon^{(n)}\right\} \geq 1 - \epsilon$ *for large enough n.*
2. $\mathbb{P}\left\{(\tilde{X}^n, \tilde{Y}^n) \in \mathcal{A}_\epsilon^{(n)}\right\} \leq 2^{-n(I(X,Y)-3\epsilon)}$.
3. $\mathbb{P}\left\{(\tilde{X}^n, \tilde{Y}^n) \in \mathcal{A}_\epsilon^{(n)}\right\} \geq (1-\epsilon)2^{-n(I(X,Y)+3\epsilon)}$ *for large enough n.*

Proof. We have the following.

1. The proof of the first point is left as an exercise.

2. For the second point, we have

$$\begin{aligned}\mathbb{P}\left\{(\tilde{X}^n, \tilde{Y}^n) \in \mathcal{A}_\epsilon^{(n)}\right\} &= \sum_{(x^n,y^n)\in\mathcal{A}_\epsilon^{(n)}} p(x^n)p(y^n) \\ &\leq 2^{n(H(X,Y)+\epsilon)}2^{-n(H(X)-\epsilon)}2^{-n(H(Y)-\epsilon)} \qquad (9) \\ &= 2^{-n(I(X;Y)-3\epsilon)}.\end{aligned}$$

   The inequality in (9) follows from the upper bound in (8), and Definition 2.

3. The proof of the third point is also left as an exercise.

□

Intuition. Since the pair (X^n, Y^n) is jointly generated, we expect it to be jointly typical as n grows large. For $(\tilde{X}^n, \tilde{Y}^n)$, we have $\tilde{X}^n \sim X^n$ and $\tilde{Y}^n \sim Y^n$. However, $\tilde{X}^n$ and $\tilde{Y}^n$ are generated independently, and therefore they have a much lower probability of being jointly typical, especially if $I(X;Y)$ is large. For large n, they may only be jointly typical with high probability if X^n and Y^n are independent, i.e. $p(x^n, y^n) = p(x^n)p(y^n)$. In this case, we clearly have $I(X;Y) = 0$.

Exercise 1. Prove points 1 and 3 in the joint AEP theorem.