

# ECE253/CSE208 Introduction to Information Theory

## Lecture 5: Asymptotic Equipartition Property (AEP)

Dr. Yu Zhang

ECE Department

University of California, Santa Cruz

- Chap 3 of *Elements of Information Theory (2nd Edition)* by Thomas Cover & Joy Thomas
- Lecture 1 Typical Sequences of *Information Theory for Wireless Comms.* by Dr. Saif Mohammed

## Typical Sequences

- LLN in probability — Sample mean converges to the true mean:

$$\bar{X} = \frac{1}{n} \sum_i X_i \xrightarrow[n \rightarrow \infty]{\text{i.p.}} \mathbb{E}(X).$$

- LLN in info theory — Sample entropy converges to the true entropy:

$$\bar{H}(X) = \frac{1}{n} \log \frac{1}{p(X_1, \dots, X_n)} \xrightarrow[n \rightarrow \infty]{\text{i.p.}} H(X).$$

### Example

Consider i.i.d.  $\{X_i\}_{i=1}^n \sim \text{Bern}(p)$ , then  $p(x_1, x_2, \dots, x_n) = \prod_{i=1}^n p(x_i)$ . For example,  $p(1, 0, 1, 1, 0, 1) = p^{\sum x_i} \times (1-p)^{n-\sum x_i} = p^4(1-p)^2$ . Clearly, not all sequences are generated equally.

We will see that  $p(X_1, X_2, \dots, X_n)$  is close to  $2^{-nH(X)}$  with high probability. That is, the probability  $p(X_1, X_2, \dots, X_n)$  assigned to an observed sequence is close to  $2^{-nH(X)}$ .

Almost all events are almost equally surprising.

$$\Pr\left\{(X_1, \dots, X_n) : p(X_1, \dots, X_n) = 2^{-n(H \pm \epsilon)}\right\} \approx 1, \text{ if } \{X_i\}_{i=1}^n \sim p(x) \text{ are i.i.d.}$$

# Typical Sets

We can thus divide the set of all sequences into two classes:

1. **Typical set**, where the probability of each typical sequence is close to  $2^{-nH(X)}$ .
2. **Atypical set** that contains all the other sequences.
3. Typical set is primarily a theoretical tool that is defined to help prove some theorems, even though its concept is somehow counter-intuitive, as we will see later.

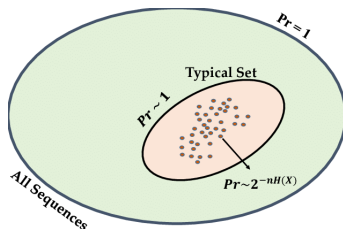


Figure: Typical sequences and typical set.

# Asymptotic Equipartition Property (AEP)

**Theorem (AEP: Empirical entropy converges to the true entropy.)**

*If  $\{X_i\}_{i=1}^n \sim p(x)$  are i.i.d., then  $-\frac{1}{n} \log p(X_1, X_2, \dots, X_n) \xrightarrow{i.p.} H(X)$ .*

**Proof:** by the weak law of large numbers (WLLN), we have

$$-\frac{1}{n} \log p(X_1, X_2, \dots, X_n) = -\frac{1}{n} \sum_i \log p(X_i) \xrightarrow{i.p.} -\mathbb{E}[\log p(X)] = H(X)$$

## Example (Sanity check of AEP)

Consider i.i.d.  $\{X_i\}_{i=1}^n \sim \text{Bern}(p)$ , let  $q = 1 - p$ . We have

$$p(x_1, \dots, x_n) = p^{\sum_{i=1}^n x_i} \times q^{n - \sum_{i=1}^n x_i} \xrightarrow{i.p.} p^{np} q^{nq}.$$

$$H(X) = -p \log p - q \log q \implies -nH(X) = \log(p^{np} q^{nq}).$$

This matches the AEP:  $p(X^n) \xrightarrow{i.p.} 2^{-nH(X)}$ .

**Q:** AEP is based on the assumption that  $X^n$  are i.i.d. How about for non-iid case?

**A:** Entropy rate of stochastic processes; see the next lecture (Chap 4).

## Weakly Typical Sequences

Some sequences are “typical” in the sense that their information is about the same as the self-information expected. We define those typical sequences as follows:

### Definition ( $\epsilon$ -typical sequence and $\epsilon$ -typical set)

A sequence  $(x_1, x_2, \dots, x_n) \in \mathcal{X}^n$  is an  $\epsilon$ -typical sequence with respect to  $p(x)$  if

$$2^{-n(H(X)+\epsilon)} \leq p(x_1, x_2, \dots, x_n) \leq 2^{-n(H(X)-\epsilon)}.$$

Further, a typical set  $A_\epsilon^{(n)}$  is the set containing all  $\epsilon$ -typical sequences  $(x_1, \dots, x_n) \in \mathcal{X}^n$ .

- Intuitively, we would like to assign shorter bit strings to the 'typical' sequences to reduce the expected length of the code. It turns out that as  $n$  grows large, almost all the probability concentrates on the typical set.
- Typical set is made by all the sequences that are giving us an amount of information close to the average information of the source distribution.
- The least and most probable sequences give us less information than the average.
- Hence, AEP filters out a lot of highly unlikely sequences as well as a small number of highly likely sequences.

## Typical Sequences (cont'd)

### Example (Most likely sequence is often not in the typical set)

- For i.i.d.  $X_i \sim \text{Bern}(0.9)$ ,  $H(X) = 0.469$ . The most likely sequence of outcome is the sequence of all 1's,  $(1, 1, \dots, 1)$ .

$$-\frac{1}{n} \log_2 p((x_1, x_2, \dots, x_n) = (1, 1, \dots, 1)) = -\frac{1}{n} \log_2 (0.9^n) = 0.152.$$

Hence, for small enough  $\epsilon$ , all-one sequence is not in the typical set.

- For Bernoulli RVs, the typical set consists of sequences with average numbers of 0's and 1's in  $n$  independent trials. Because if a sequence has  $np$  1's and  $nq$  0's for  $n$  trials, then  $p(x_1, \dots, x_n) = p^{np} q^{nq} \implies$

$$-\frac{1}{n} \log_2 p(x_1, x_2, \dots, x_n) = -p \log p - q \log q = H(X).$$

If  $p = 0.9, n = 10$ , then the typical set consist of all sequences that have a single 0 in the entire sequence. If  $p = 0.5$ , then every possible binary sequences belong to the typical set.

## Typical Sequences and Set (cont'd)

Example (The “typicality” is in the sense of *sample entropy close to the true entropy*, rather than “most likely”)

A computer program is used to generate a binary sequence of length 10 digits (i.i.d.  $X_i \sim \text{Bern}(\frac{1}{3})$ ). One of the following four sequences is generated from the program. Which one is it?

(a) 0 0 0 0 0 0 0 0 0 0 0 0,  $\Pr(a) = (2/3)^{12} = 7.7 \times 10^{-3}$

(b) 1 0 1 1 0 1 0 1 0 1 0 0,  $\Pr(b) = (2/3)^6 \times (1/3)^6 = 1.2 \times 10^{-4}$

(c) 0 0 0 1 0 0 0 1 0 0 1 0,  $\Pr(c) = (2/3)^9 \times (1/3)^3 = 9.6 \times 10^{-4}$

(d) 1 1 1 1 1 1 1 1 1 1 1 1,  $\Pr(d) = (1/3)^{12} = 1.9 \times 10^{-6}$ .

The answer is sequence (c), although sequence (a) has a higher probability of occurrence. An intuition based reasoning is that, since the source outputs are i.i.d., roughly  $1/3$  of the 12 digits should be zero and  $2/3$  should be one. This is in fact true as the length of the sequence is increasing. Those sequences are called “typical sequences”.

## Typical Sequences and Set (cont'd)

- Consider a random source i.i.d.  $X_i \sim \text{Bern}(p)$  generating a sequence of length  $n$ .
- There are  $\binom{n}{np}$  independent sequences that have exactly  $np$  ones and the probability of each such sequence is  $p^{np}(1-p)^{n(1-p)}$ .
- Approximate  $\binom{n}{np}$  by using the Stirling's formula:  $n! \approx \sqrt{2\pi n} \left(\frac{n}{e}\right)^n$ .

$$\begin{aligned}\log \binom{n}{np} &= \log \left( \frac{n!}{(np)!(n-np)!} \right) \\ &\approx \log \left( \frac{\sqrt{2\pi n} \left(\frac{n}{e}\right)^n}{\sqrt{2\pi np} \left(\frac{np}{e}\right)^{np} \sqrt{2\pi n(1-p)} \left(\frac{n(1-p)}{e}\right)^{n(1-p)}} \right) \\ &= -\log \left( \sqrt{2\pi np(1-p)} \right) - n \times [p \log p + (1-p) \log(1-p)]\end{aligned}$$

$$\Rightarrow \boxed{\binom{n}{np} \approx \frac{2^{nH(p)}}{\sqrt{2\pi np(1-p)}}}$$

Hence, the number of such sequences increases as  $2^{nH(p)}$ , but it's a much smaller subset of all possible sequences.



# Properties of AEP

## Theorem

Let  $x^n := (x_1, x_2, \dots, x_n)$ , we have the following properties of the AEP:

1. If  $x^n \in A_\epsilon^{(n)}$  then  $H(X) - \epsilon \leq -\frac{1}{n} \log p(x^n) \leq H(X) + \epsilon$ .
2.  $\Pr\{A_\epsilon^{(n)}\} > 1 - \epsilon$  for  $n$  sufficiently large.
3.  $|A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$ .
4.  $|A_\epsilon^{(n)}| \geq (1 - \epsilon)2^{n(H(X)-\epsilon)}$  for  $n$  sufficiently large.

The above Theorem asserts that i) the typical set has probability nearly 1; ii) all elements in it are nearly equiprobable; and iii) the size of the typical set is about  $2^{nH(X)}$ .

## Proof of AEP Properties

### Proof.

1. From the definition of  $\epsilon$ -typical sequences, if  $(x_1, \dots, x_n) \in A_\epsilon^{(n)}$ , then we have

$$2^{-n(H(X)+\epsilon)} \leq p(x^n) \leq 2^{-n(H(X)-\epsilon)}$$

Taking the log and dividing by  $-n$  yields Property 1.

2. Since  $-\frac{1}{n} \log p(X^n) \xrightarrow{\text{i.p.}} H(X)$ , for any  $\delta > 0$ ,  $\exists n_0$  such that for all  $n \geq n_0$ ,

$$\Pr \left( \left| -\frac{1}{n} \log p(X^n) - H(X) \right| < \epsilon \right) > 1 - \delta.$$

Finally, we know  $\left| -\frac{1}{n} \log p(X^n) - H(X) \right| < \epsilon$  holds iff  $X^n$  is  $\epsilon$ -typical, so we can set  $\delta := \epsilon$  to obtain Property 2.

## Proof of AEP Properties (cont.)

### Proof.

3. By Property 1, we have  $2^{-n(H(X)+\epsilon)} \leq p(x^n)$ ,  $\forall x^n \in A_\epsilon^{(n)} \implies$

$$2^{-n(H(X)+\epsilon)} \left| A_\epsilon^{(n)} \right| \leq \sum_{x^n \in A_\epsilon^{(n)}} p(x^n) \leq \sum_{x^n \in \mathcal{X}^n} p(x^n) = 1,$$

which proves Property 3.

4. By Property 2, for  $n$  sufficiently large, we have  $\Pr\{A_\epsilon^{(n)}\} > 1 - \epsilon$ . Hence,

$$1 - \epsilon < \Pr\{A_\epsilon^{(n)}\} = \sum_{x^n \in A_\epsilon^{(n)}} p(x^n) \leq \sum_{x^n \in A_\epsilon^{(n)}} 2^{-n(H(X)-\epsilon)} = 2^{-n(H(X)-\epsilon)} \left| A_\epsilon^{(n)} \right|.$$

The “ $\leq$ ” follows from the upper bound of  $p(x^n)$ .

## Properties of AEP (cont'd)

For small  $\epsilon$ , we have  $|A_\epsilon^{(n)}| \approx 2^{nH(X)}$ . Thus, the fraction of sequences that are typical is

$$\rho_n := \frac{|A_\epsilon^{(n)}|}{|\mathcal{X}^n|} \approx \frac{2^{nH(X)}}{|\mathcal{X}|^n} = \frac{2^{nH(X)}}{2^{n \log |\mathcal{X}|}} = 2^{-n(\log |\mathcal{X}| - H(X))}.$$

- For non-uniform distribution:  $H(X) < \log |\mathcal{X}|$ ,  $\rho_n \rightarrow 0$  as  $n \rightarrow \infty$ .
- For uniform distribution:  $H(X) = \log |\mathcal{X}| \rightarrow \rho_n = 1$ , every sequence is typical.

Everything outside the typical set has a negligible probability.

- $|A_\epsilon^{(n)}|$  is **exponentially small fraction in  $n$** . However, the typical sequences make up most of the probability because  $\Pr\{A_\epsilon^{(n)}\} > 1 - \epsilon$ .
- In other words, the probability of a generated sequence being in the typical set is high, even though the number of elements in the typical set is much smaller than the total number of possible sequences.
- For  $n$  sufficiently large, we can almost think of the sequence  $X^n$  as being obtained by choosing a sequence from the weakly typical set according to the uniform distribution  $\rightarrow$  “asymptotic equipartition”.

# Strongly Typical Sequences

## Definition (Strongly Typical Sets)

The strongly typical set  $T_\delta^{(n)}$  with respect to a distribution function  $p(x)$  is the set of sequences  $x^n \in \mathcal{X}^n$  such that

$$\sum_{x \in \mathcal{X}} \left| \frac{1}{n} N(x; x^n) - p(x) \right| \leq \delta,$$

where  $N(x; x^n)$  is the number of occurrences of  $x$  in the sequence  $x^n$ , and  $\delta > 0$  is an arbitrarily small number. The sequences in  $T_\delta^{(n)}$  are called strongly  $\delta$ -typical sequences.

## Theorem (Strong AEP)

*There exists  $\eta > 0$  such that  $\eta \rightarrow 0$  as  $\delta \rightarrow 0$ , and the following hold*

1. *If  $x^n \in T_\delta^{(n)}$ , then  $H(X) - \eta \leq -\frac{1}{n} \log p(x^n) \leq H(X) + \eta$ .*
2.  *$\Pr\{T_\delta^{(n)}\} > 1 - \delta$  for  $n$  sufficiently large.*
3.  *$|T_\delta^{(n)}| \leq 2^{n(H(X) + \eta)}$ .*
4.  *$|T_\delta^{(n)}| \geq (1 - \delta) 2^{n(H(X) - \eta)}$  for  $n$  sufficiently large.*

## Strong Typicality vs Weak Typicality

- Weak typicality (entropy typicality): empirical entropy  $\approx$  true entropy.
- Strong typicality (letter typicality): empirical distribution  $\approx$  true distribution.
- Strong typicality  $\implies$  Weak typicality, but not vice versa.
- Strong typicality works only for finite alphabet, i.e.,  $|\mathcal{X}| < \infty$ .

## High-probability Set

To this end, we know that the  $A_\epsilon^{(n)}$  is a fairly small set that has most of the probability.

**Q:** Is it the smallest set with such a property?

### Definition

For  $\delta > 0$ , let  $B_\delta^{(n)} \subset \mathcal{X}^n$  be the smallest set such that  $\Pr(X^n \in B_\delta^{(n)}) \geq 1 - \delta$ .

### Theorem

Let  $\delta < \frac{1}{2}$ . For any  $\delta' > 0$ ,

$$\frac{1}{n} \log |B_\delta^{(n)}| > H - \delta'$$

for  $n$  sufficiently large.

### Typical set vs High-probability set.

For sufficiently large  $n$  (depending on  $\delta$  and  $\delta'$ ),  $B_\delta^{(n)}$  has at least  $2^{n(H-\delta')}$  elements. The  $\epsilon$ -typical set  $A_\epsilon^{(n)}$  has about  $2^{n(H \pm \epsilon)}$  elements. Thus,  $A_\epsilon^{(n)}$  and  $B_\delta^{(n)}$  have roughly the same number of elements to first order in the exponent.

## Encoding for the Typical Set

The fact that the typical set has probability approaching 1 as  $n$  grows large means that we “only need” to care about encoding the sequences in the typical set.

The number of bits required to encode a set of size  $\mathcal{S}$  is  $\lceil \log |\mathcal{S}| \rceil$ , where the ceiling operator  $\lceil a \rceil$  outputs the smallest integer number no less than  $a$ .

Let i.i.d.  $\{X_i\}_{i=1}^n \sim p(x)$ . Consider the following scheme for coding  $x^n \in \mathcal{X}^n$ .

- First, consider a complete order of all the sequences in  $\mathcal{A}_\epsilon^{(n)}$  and its complement, according to a certain criterion (e.g., lexicographic order, “ABC, ACB, BAC, BCA, CAB, CBA”).
- We use the first bit as an indicator to show if  $x^n$  is typical, say, start with 0 if the sequence is typical, otherwise start with 1.



## Encoding for the Typical Set (cont'd)

- If  $x^n \in A_\epsilon^{(n)}$ , since  $|A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$ , use  $n(H(X) + \epsilon) + 1$  bits for encoding (the additional 1 bit is due to integrality),
- If  $x^n \notin A_\epsilon^{(n)}$ , use no more than  $n \log |\mathcal{X}| + 1$  bits to encode it.

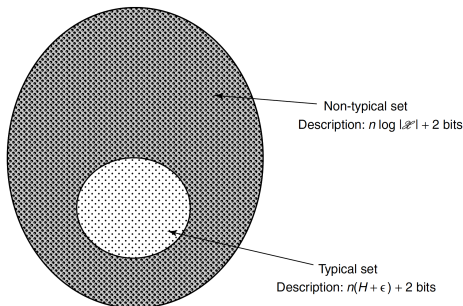


Figure: Encoding for the typical set.

## Consequences of AEP

Let  $\ell(x^n)$  denote the length of the codeword (a binary string) corresponding to  $x^n \in \mathcal{X}^n$ . For a sufficiently large  $n$ , we have

$$\begin{aligned} \mathbb{E}[\ell(x^n)] &\leq P\left(x^n \in A_\epsilon^{(n)}\right) \times (n(H + \epsilon) + 2) + P\left(x^n \notin A_\epsilon^{(n)}\right) \times (n \log |\mathcal{X}| + 2) \\ &= 2 + P\left(x^n \in A_\epsilon^{(n)}\right) \times (n(H + \epsilon)) + P\left(x^n \notin A_\epsilon^{(n)}\right) \times (n \log |\mathcal{X}|) \\ &\leq 2 + n(H + \epsilon) + \epsilon n \log |\mathcal{X}| =: n(H + \tilde{\epsilon}) \end{aligned}$$

where  $\tilde{\epsilon} = \epsilon(1 + \log |\mathcal{X}|) + \frac{2}{n}$  can be arbitrarily small by appropriate choices of  $\epsilon$  and  $n$ .

### Theorem ( $H(X)$ bits are needed to encode $X^n$ per symbol on average)

Consider i.i.d.  $\{X_i\}_{i=1}^n \sim p(x)$ . Let  $\epsilon > 0$ , then there exists a code that maps sequences  $x^n$  into binary strings, such that the mapping is one-to-one and  $E\left[\frac{1}{n}\ell(x^n)\right] \leq H(x) + \epsilon$  for  $n$  sufficiently large.

The above theorem explains the achievability part of the *Source Coding Theorem*: A sequence of symbols can be compressed to a binary string with an average of  $H(X)$  bits per symbol. This further reinforces the interpretation of the entropy as the average information content of a random source.

## Jointly Typical Sequences

Two sequences  $x^n$  and  $y^n$  are jointly  $\epsilon$ -typical if

1. the pair  $(x^n, y^n)$  is  $\epsilon$ -typical with respect to the joint distribution  $p(x^n, y^n) = \prod_{i=1}^n p(x_i, y_i)$  (i.e., pairwise independence).
2. both  $x^n$  and  $y^n$  are  $\epsilon$ -typical w.r.t. their marginal distributions  $p(x^n)$  and  $p(y^n)$ .

The set of all such pairs of sequences  $(x^n, y^n)$  is denoted by

$$A_{\epsilon}^{(n)}(X, Y) = \left\{ (x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : \begin{aligned} &\left| -\frac{1}{n} \log p(X^n) - H(X) \right| < \epsilon, \\ &\left| -\frac{1}{n} \log p(Y^n) - H(Y) \right| < \epsilon, \\ &\left| -\frac{1}{n} \log p(X^n, Y^n) - H(X, Y) \right| < \epsilon \end{aligned} \right\}$$

## Joint AEP

### Theorem

Let  $(X^n, Y^n)$  be sequences of length  $n$  drawn i.i.d.  $\sim p(x^n, y^n) = \prod_{i=1}^n p(x_i, y_i)$ . Then,

1.  $\Pr \left( (X^n, Y^n) \in A_\epsilon^{(n)} \right) \rightarrow 1$  as  $n \rightarrow \infty$ .
2.  $(1 - \epsilon)2^{n(H(X,Y) - \epsilon)} \leq |A_\epsilon^{(n)}| \leq 2^{n(H(X,Y) + \epsilon)}$ .
3. If  $(\tilde{X}^n, \tilde{Y}^n) \sim p(x^n)p(y^n)$ , then

$$(1 - \epsilon)2^{-n(I(X;Y) + 3\epsilon)} \leq \Pr \left( (\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)} \right) \leq 2^{-n(I(X;Y) - 3\epsilon)},$$

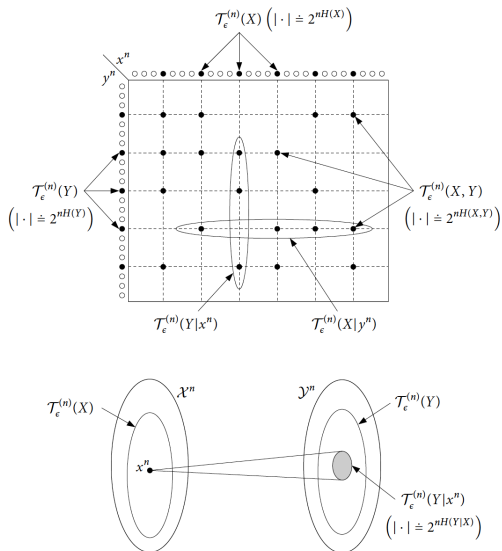
where both lower bounds in parts 2 and 3 hold for  $n$  sufficiently large.

### Implication:

- Typical sets  $|X^n| \approx 2^{nH(X)}$  and  $|Y^n| \approx 2^{nH(Y)}$ .
- Not all pair of typical  $X^n$  and typical  $Y^n$  are jointly typical: only about  $2^{nH(X,Y)}$ .
- Intuitive argument for joint typicality lemma: the probability of any randomly chosen pair is jointly typical is about  $\frac{2^{nH(X,Y)}}{2^{n(H(X)+H(Y))}} = 2^{-nI(X;Y)}$

We use the joint AEP and random coding to prove the channel coding theorem (Chap 7).

# Illustration of Joint AEP



**Figure:** Source: Chapter 2 of Network Information Theory by El Gamal and Kim.

## Proof of Joint AEP (Part 1)

By LLN,  $-\frac{1}{n} \log p(X^n) \xrightarrow{\text{i.p.}} -E[\log p(X)] = H(X)$  ().

Hence, given  $\epsilon > 0$ ,

- $\exists n_1$ , such that for all  $n > n_1$ ,  $\Pr \left( \underbrace{\left| -\frac{1}{n} \log p(X^n) - H(X) \right|}_{A} \geq \epsilon \right) < \frac{\epsilon}{3}$
- $\exists n_2$ , such that for all  $n > n_2$ ,  $\Pr \left( \underbrace{\left| -\frac{1}{n} \log p(Y^n) - H(Y) \right|}_{B} \geq \epsilon \right) < \frac{\epsilon}{3}$
- $\exists n_3$ , such that for all  $n > n_3$ ,  $\Pr \left( \underbrace{\left| -\frac{1}{n} \log p(X^n, Y^n) - H(X, Y) \right|}_{C} \geq \epsilon \right) < \frac{\epsilon}{3}$

Hence, choosing  $n > \max\{n_1, n_2, n_3\}$ , we have

$\Pr(A \cup B \cup C) \leq \Pr(A) + \Pr(B) + \Pr(C) < \epsilon$ , which implies  $\Pr(\bar{A}\bar{B}\bar{C}) \geq 1 - \epsilon$ .

## Proof of Joint AEP (Part 2)

Note that if  $(x^n, y^n) \in A_\epsilon^{(n)}$ , then  $|\frac{1}{n} \log p(X^n, Y^n) - H(X, Y)| < \epsilon \implies$

$$2^{-n(H(X,Y)+\epsilon)} < p(x^n, y^n) < 2^{-n(H(X,Y)-\epsilon)}.$$

The upper bound of  $|A_\epsilon^{(n)}|$  is due to

$$1 = \sum p(x^n, y^n) \geq \sum_{A_\epsilon^{(n)}} p(x^n, y^n) \geq |A_\epsilon^{(n)}| 2^{-n(H(X,Y)+\epsilon)}.$$

To derive the lower bound of  $|A_\epsilon^{(n)}|$ : For sufficiently large  $n$ ,  $\Pr(A_\epsilon^{(n)}) \geq 1 - \epsilon$ , and thus

$$1 - \epsilon \leq \sum_{(x^n, y^n) \in A_\epsilon^{(n)}} p(x^n, y^n) \leq |A_\epsilon^{(n)}| 2^{-n(H(X,Y)-\epsilon)}.$$

## Proof of Joint AEP (Part 3)

If  $\tilde{X}^n$  and  $\tilde{Y}^n$  are independent but have the same marginals as  $X^n$  and  $Y^n$ , then

$$\begin{aligned}\Pr\left(\left(\tilde{X}^n, \tilde{Y}^n\right) \in A_{\epsilon}^{(n)}\right) &= \sum_{(x^n, y^n) \in A_{\epsilon}^{(n)}} p(x^n) p(y^n) \\ &\leq 2^{n(H(X, Y) + \epsilon)} \times 2^{-n(H(X) - \epsilon)} \times 2^{-n(H(Y) - \epsilon)} \\ &= 2^{-n(I(X; Y) - 3\epsilon)}.\end{aligned}$$

By similar arguments, we can show that for  $n$  sufficiently large,

$$\begin{aligned}\Pr\left(\left(\tilde{X}^n, \tilde{Y}^n\right) \in A_{\epsilon}^{(n)}\right) &= \sum_{(x^n, y^n) \in A_{\epsilon}^{(n)}} p(x^n) p(y^n) \\ &\geq (1 - \epsilon) 2^{n(H(X, Y) - \epsilon)} \times 2^{-n(H(X) + \epsilon)} \times 2^{-n(H(Y) + \epsilon)} \\ &= (1 - \epsilon) 2^{-n(I(X; Y) + 3\epsilon)}.\end{aligned}$$



*Thank You!*

Email: <zhangy@ucsc.edu>

Homepage: <https://people.ucsc.edu/~yzhan419/>