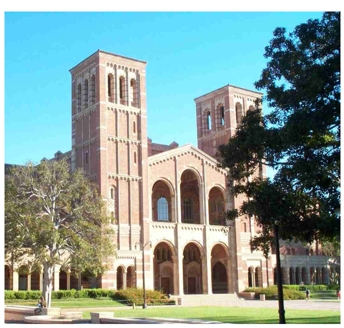
EE 231A: Information Theory
Lecture 3

A. Types of convergence, the weak law of large numbers, and the Asymptotic Equipartition Property
B. Properties of the typical set
C. AEP Data Compression
D. High Probability Sets vs. typical sets.

1

EE 231A: Information Theory
Lecture 3



A. Types of convergence, the weak law of large numbers, and the Asymptotic Equipartition Property
B. Properties of the typical set
C. AEP Data Compression
D. High Probability Sets vs. typical sets.

2

Convergence in probability

- $f_n \to f$ in probability iff for any $\varepsilon > 0$

$$P(|f_n - f| < \varepsilon) \to 1$$

  – i.e., a probability is converging.

Convergence with probability 1

$$f_n \to f \qquad \text{w.p.1}$$

– means that f_n is converging to f for every realization (except a set of measure zero).

Weak law of large numbers

- Let $X_1, X_2, \ldots, X_n$ be a sequence of i.i.d. random variables.

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to EX \qquad \text{in probability}$$

5

Asymptotic Equipartition Property

- Let $X_1, X_2, \ldots, X_n$ be a sequence of i.i.d. random variables.

$$-\frac{1}{n}\log p(X_1,\ldots,X_n) \to H(X) \qquad \text{in prob.}$$

- Proof:

$$\begin{aligned} -\frac{1}{n}\log p(X_1,\ldots,X_n) &= -\frac{1}{n}\sum_i \log p(X_i) \\ &\to -E\log p(X) \qquad \text{in probability} \\ &= H(X) \end{aligned}$$

6

EE 231A: Information Theory Lecture 3

A. Types of convergence, the weak law of large numbers, and the Asymptotic Equipartition Property
B. Properties of the typical set
C. AEP Data Compression
D. High Probability Sets vs. typical sets.

7

Asymptotic Equipartition Property

- Let $X_1, X_2, \ldots, X_n$ be a sequence of i.i.d. random variables.

$$-\frac{1}{n}\log p(X_1,\ldots,X_n) \to H(X) \qquad \text{in prob.}$$

- The Typical Set is the set of sequences $x_1,\ldots, x_n$ within epsilon of convergence behavior, i.e. the set where

$$H(X)-\varepsilon \le -\frac{1}{n}\log p(x_1,\ldots,x_n) \le H(X)+\varepsilon.$$

8

The Typical Set

- The typical set $A_\varepsilon^{(n)}$ for $p(x)$ is the set of sequences x^n for which

$$2^{-n(H(X)+\varepsilon)} \leq p(x_1, x_2, \ldots, x_n) \leq 2^{-n(H(X)-\varepsilon)}$$

9

Properties of $A_\varepsilon^{(n)}$

10

1. Restatement of $A_\varepsilon^{(n)}$

$$x^n \in A_\varepsilon^{(n)} \Leftrightarrow H(X) - \varepsilon \le -\frac{1}{n}\log p(x^n) \le H(X) + \varepsilon$$

Proof:

Take $-\frac{1}{n}\log$ of each term in the definition of $A_\varepsilon^{(n)}$.

11

2. Probability of the typical set converges to 1.

$P(A_\varepsilon^{(n)}) > 1 - \varepsilon$ for n sufficiently large.

– This is restatement of AEP. $P(A_\varepsilon^{(n)}) \to 1.$

12

3. Cardinality Upper Bound

$$\underbrace{\left|A_\varepsilon^{(n)}\right|}_{\#\text{ of elements in } A_\varepsilon^{(n)}} \le 2^{n(H(X)+\varepsilon)}$$

- Proof:

$$\begin{aligned} 1 &= \sum_{x^n} p(x^n) \\ &\ge \sum_{x^n \in A_\varepsilon^{(n)}} p(x^n) \\ &\ge \sum_{x^n \in A_\varepsilon^{(n)}} 2^{-n(H(X)+\varepsilon)} \\ &= \left|A_\varepsilon^{(n)}\right| 2^{-n(H(X)+\varepsilon)} \end{aligned}$$

$$\left|A_\varepsilon^{(n)}\right| \le 2^{n(H(X)+\varepsilon)}$$

13

4. Cardinality Lower Bound

$\left|A_\varepsilon^{(n)}\right| \ge (1-\varepsilon)2^{n(H(X)-\varepsilon)}$ for n sufficient large.

- Proof:

$$\begin{aligned} 1-\varepsilon &< P(A_\varepsilon^{(n)}) \quad \text{for } n \text{ sufficiently large.} \\ &\le \sum_{x^n \in A_\varepsilon^{(n)}} 2^{-n(H(X)-\varepsilon)} \\ &= 2^{-n(H(X)-\varepsilon)} \left|A_\varepsilon^{(n)}\right| \end{aligned}$$

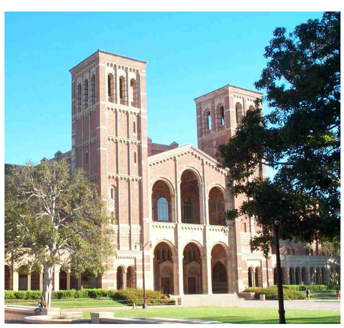
14

AEP Summary

- Consider drawing a sequence of n samples at random by drawing samples i.i.d. according to some distribution.
- With very high probability the sequence that occurs will be in the typical set and have probability about . 2^{-nH}
- There are about 2^{nH} events in the typical set and they have about the same probability.
- "Almost all events are equally surprising."

15

EE 231A: Information Theory Lecture 3



A. Types of convergence, the weak law of large numbers, and the Asymptotic Equipartition Property
B. Properties of the typical set
C. AEP Data Compression
D. High Probability Sets vs. typical sets.

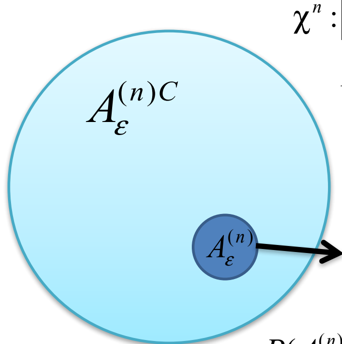
16

A small set with a lot of probability.

- Generally speaking, the typical set contains a small number of sequences, but almost all the probability.

$\chi^n : |\chi|^n$ elements

$P\left(x^n \in \chi^n\right) = 1$



$A_\varepsilon^{(n)} : 2^{n(H+\varepsilon)}$ elements

$P(A_\varepsilon^{(n)}) \to 1.$

$P(A_\varepsilon^{(n)}) > 1-\varepsilon$ for n suff. large.

AEP data compression concept

- Idea: Provide short description for elements of typical set. Don't worry too much about other sequences.

A short description for typical sequences

- Label each element in the typical set with a unique label using $\lceil \log_2 |A_\varepsilon^{(n)}| \rceil$ bits.

$$\lceil \log_2 |A_\varepsilon^{(n)}| \rceil \le \log_2 2^{n(H(X)+\varepsilon)} + 1$$
$$= n(H(X)+\varepsilon) + 1$$

- Add a leading zero to indicate membership in $A_\varepsilon^{(n)}$. $n(H(X)+\varepsilon)+2$ bits

19

Long description for (rare) atypical sequences

- Label each sequence not in $A_\varepsilon^{(n)}$ with $n \log|\chi| + 2$ bits (a leading 1 indicates $\notin A_\varepsilon^{(n)}$).
- Code is easily decodable.
- We used a brute force labeling of $A_\varepsilon^{(n)C}$.
- Typical sequences have $\approx nH$ bits.

20

Expected length of codeword

- $$E[l(x^n)] = \sum_{x^n} p(x^n) l(x^n)$$
$$= \sum_{x^n \in A_\varepsilon^{(n)}} p(x^n) l(x^n) + \sum_{x^n \notin A_\varepsilon^{(n)}} p(x^n) l(x^n)$$
$$\leq \sum_{x^n \in A_\varepsilon^{(n)}} p(x^n)[n(H+\varepsilon)+2] + \sum_{x^n \notin A_\varepsilon^{(n)}} p(x^n)[n \log|\mathcal{X}|+2]$$
$$= P(A_\varepsilon^{(n)})[n(H+\varepsilon)+2] + P(A_\varepsilon^{(n)C})[n \log|\mathcal{X}|+2]$$
$$\leq n(H+\varepsilon) + \varepsilon n \log|\mathcal{X}| + 2$$
$$= n(H+\varepsilon+\varepsilon \log|\mathcal{X}| + \frac{2}{n})$$
$$= n(H+\varepsilon')$$

21

Theorem 3.2.1

- For X^n i.i.d. $\sim p(x)$, we can map sequences x^n to binary strings such that the mapping is one-to-one (invertible) and

$$E[\frac{1}{n} l(X^n)] \leq H(X) + \varepsilon$$

for $\varepsilon > 0$ and n sufficiently large.

- Thus we can represent sequences X^n using $nH(X)$ bits on the average.

22

EE 231A: Information Theory
Lecture 3

A. Types of convergence, the weak law of large numbers, and the Asymptotic Equipartition Property
B. Properties of the typical set
C. AEP Data Compression
D. High Probability Sets vs. typical sets.

23

High probability sets and $A_\varepsilon^{(n)}$

- Consider a binary sequence with probability of ones p=0.9.

$$H(X) = 0.4690$$

- Sequences with about 90% ones are in $A_\varepsilon^{(n)}$.

$$\boxed{-\frac{1}{10}\log\left((0.9)^9\,0.1\right) = 0.4690}$$

- $P(A_\varepsilon^{(n)}) \to 1$ but for small ε, the most probable x^n, the all-ones sequence, is not in $A_\varepsilon^{(n)}$.

- Consider n=10: $\boxed{-\frac{1}{10}\log\left((0.9)^{10}\right) = 0.1520}$

24

High probability sets

- $B_\delta^{(n)} \subset \chi^n$ is any set with

$$P\{B_\delta^{(n)}\} \geq 1-\delta$$

- $B_\delta^{(n)} \cap A_\varepsilon^{(n)}$ must still have large probability.
- Hence…

25

Theorem 3.3.1

- $X_1, X_2, \ldots \text{ i.i.d. } \sim p(x)$, for $\delta < \frac{1}{2}, \delta' > 0$, if $P\{B_\delta^{(n)}\} > 1-\delta$ then

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

for n sufficiently large.

26

Implication of Theorem 3.3.1

- Thus $B_\delta^{(n)}$ must have at least 2^{nH} elements, to first order in the exponent.
- Thus, even though $A_\varepsilon^{(n)}$ may not contain the most probable sequence, it is about as small as the smallest set containing $1-\delta$ of the probability.

27

All the high-probability sets have about the same cardinality.

- Definition: $a_n \doteq b_n$ means

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{a_n}{b_n} = 0$$

- $a_n \doteq b_n$ means a_n and b_n are equal to first order in the exponent.
- $| B_\delta^{(n)} | \doteq | A_\varepsilon^{(n)} | \doteq 2^{nH}$.

28