CS258: Information Theory

Fan Cheng
Shanghai Jiao Tong University

http://www.cs.sjtu.edu.cn/~chengfan/
chengfan@sjtu.edu.cn
Spring, 2020

Outline

- Law of Large Numbers
- Asymptotic Equipartition Property
- Typical Set
- Data Compression

Grand Picture

1%的人掌握了99%的财富，1%的事件占据了99%的概率
20%的人完成了80%的工作，20%的任务耗费了80%的资源
一种信息论的观点

- 人多势众 → 势众人多?

$\boldsymbol{X = x_1}$	$X = x_2$	…	…	$X = x_n$
p_1	p_2	…	…	p_n

- We say the occurrence of some events is 99% in probability.
  - The number of such events may be very small.
- Two different points of view
  - Utility maximization
  - Fairness

Terminology of Probability Theory

- $\mathcal{X}$: sample space or alphabet. X: random variable. x: an event in $\mathcal{X}$
- **(i.i.d.): independent, identically distributed**
- $\Pr(X = x)$: the probability of event $x \in \mathcal{X}$
- For a set A,

$$\mathbf{Pr}(\boldsymbol{A}) := \sum_{\boldsymbol{x}\in\boldsymbol{A}} \mathbf{Pr}(\boldsymbol{X} = \boldsymbol{x})$$

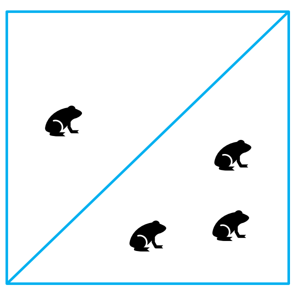
We say that events occurred in probability $\Pr(A)$ or the probability of set A is $\Pr(A)$

If X and X' are i.i.d. random variables, then

$$\Pr(X = X') = \sum_{x} \Pr(X = x)\Pr(X' = x) = \sum_{x} p^2(x)$$

- For two independent random variables X and Y, the probability mass function of $Z = X + Y$ is the **convolution** of the p.m.fs of X and Y

$$\mathbf{Pr}(\boldsymbol{Z} = \boldsymbol{z}) = \sum_{\boldsymbol{x}\in\boldsymbol{\mathcal{X}}} \mathbf{Pr}(\boldsymbol{X} = \boldsymbol{x})\,\mathbf{Pr}(\boldsymbol{Y} = \boldsymbol{z} - \boldsymbol{x})$$



- By counting the number of frogs,

$$\Pr(\text{frogs stay in the lower triangle}) = \frac{3}{4}$$

- If the probability of the frog in the upper triangle is $\frac{2}{3}$, then

$$\Pr(\text{frogs stay in the lower triangle}) = \frac{1}{3}$$

Convergence of random variables

Definition (Convergence of random variables). Given a sequence of random variables, $X_1, X_2, \dots,$ we say that the sequence $X_1, X_2, \dots,$ converges to a random variable X:

1. **In probability** if for every $\epsilon > 0, \Pr\{|X_n - X| > \epsilon\} \to 0$
2. **In mean square** if $E(X_n - X)^2 \to 0$
3. **With probability 1** (also called almost surely) if $\Pr\{\lim_{n\to\infty} X_n = X\} = 1$

The corresponding $\epsilon - \delta$ form

1. In probability
   - The set of events $A: |X_n - X| > \epsilon$
   - For any $\epsilon' > 0$, there exists $n > N(\epsilon')$,
     $\Pr(A) < \epsilon'$
   - **Equivalently, $\Pr(|X_n - X| \le \epsilon) \to 1$ or $\Pr(A^c) \to 1$**
2. In mean square
   - For any $\epsilon' > 0$, there exists $n > N(\epsilon')$,
     $$E(X_n - X)^2 < \epsilon'$$
3. With probability 1
   - Let $Y = \lim_{n\to\infty} X_n$. $Y = X$: For any $\epsilon' > 0$, there exists $n > N(\epsilon')$,
     $$|X_n - Y| < \epsilon'$$
   - $$\Pr(Y = X) = 1$$

- $(2) \to (1)$, $(3) \to (1)$

Law of Large Numbers

For i.i.d. random variables $X_1, X_2, \dots, X_n \sim p(x)$

$$\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i\,,$$

- Strong law of large number

$$\Pr\{\lim_{n\to\infty} \overline{X_n} = E(X_1)\} = 1.$$

- **Weak law of large number**

$$\boldsymbol{\overline{X_n} \to E(X_1)}$$

in probability

- $E(X)$ may not exist

The $\epsilon - \delta$ form of weak law of large numbers

- By the definition of "convergence in probability"

$$\Pr\left(\left|\overline{X_n} - E(X_1)\right| > \epsilon\right) \to 0$$

- For any $\epsilon' > 0$, there exists $N(\epsilon')$, when $n > N(\epsilon')$

$$\Pr\left(\left|\overline{X_n} - E(X_1)\right| > \epsilon\right) < \epsilon'$$

When n is sufficiently large, $\Pr\left(\left|\overline{X_n} - E(X_1)\right| \le \epsilon\right) > 1 - \epsilon'$; i.e.,

$$\boldsymbol{\Pr\left(\left|\overline{X_n} - E(X_1)\right| \le \epsilon\right) \to 1}$$

Asymptotic Equipartition Property

Theorem (AEP 渐近均分性) If $X_1, X_2, \ldots$ are i.i.d. $\sim p(x)$, then

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

Proof.

$$-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) = -\frac{1}{n}\sum_i \log p(X_i)$$
$$\rightarrow -E\log p(X) \text{ in probability}$$
$$= H(X)$$

- $-\frac{1}{n}\log p(X_1, \ldots, X_n) \rightarrow H(X)$
- 总概率$\rightarrow 1$

The counterpart of L.L.N in information theory

$$H(X) - \epsilon \leq -\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) \leq H(X) + \epsilon \text{ in prob.}$$

$$\mathbf{2^{-n(H(X)+\epsilon)} \leq p(X_1, X_2, \ldots, X_n) \leq 2^{-n(H(X)-\epsilon)} \Rightarrow A_\epsilon^{(n)}}$$

Typical Set

The **typical set (典型集)** $A_\epsilon^{(n)}$ with respect to $p(x)$ is the set of sequences $(x_1, x_2, \dots, x_n) \in \mathcal{X}^n$ with the property

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

1. If $(x_1, x_2, \dots, x_n) \in A_\epsilon^{(n)}$, then $H(X) - \epsilon \le -\frac{1}{n}\log p(x_1, x_2, \dots, x_n) \le H(X) + \epsilon$
2. $\Pr\left\{A_\epsilon^{(n)}\right\} \ge 1 - \epsilon$ for n sufficiently large.
3. $|A_\epsilon^{(n)}| \le 2^{n(H(X)+\epsilon)}$, where $|A|$ denotes the number of elements in the set A.
4. $\left|A_\epsilon^{(n)}\right| \ge (1-\epsilon)2^{n(H(X)-\epsilon)}$ for n sufficiently large.

Intuition

- 2. The typical set has probability nearly 1
- 3. All elements of the typical set are nearly equiprobable (等概率)
- 4. The number of elements in the typical set is nearly 2^{nH}

Typical Set (cont'd)

The **typical set (典型集) $A_\epsilon^{(n)}$** with respect to $p(x)$ is the set of sequences $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ with the property

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

By definition and $\epsilon - \delta$ form

1. If $(x_1, x_2, \ldots, x_n) \in A_\epsilon^{(n)}$, then $H(X) - \epsilon \leq -\frac{1}{n}\log p(x_1, x_2, \ldots, x_n) \leq H(X) + \epsilon$

Proof. By the definition of typical set.

2. $\Pr\left\{A_\epsilon^{(n)}\right\} \geq 1 - \epsilon$ for n sufficiently large.

Proof. By AEP Theorem, the probability of the event $(X_1, X_2, \ldots, X_n) \in A_\epsilon^{(n)}$ tends to 1 as $n \to \infty$. Thus, for any $\delta > 0$, there exists an n_0 such that for all $n \geq n_0$, we have

$$\Pr\left\{\left|-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) - H(X)\right| < \epsilon\right\} > 1 - \delta$$

Setting $\delta = \epsilon$.

Typical Set (cont'd)

The **typical set (典型集)** $A_\epsilon^{(n)}$ with respect to $p(x)$ is the set of sequences $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ with the property

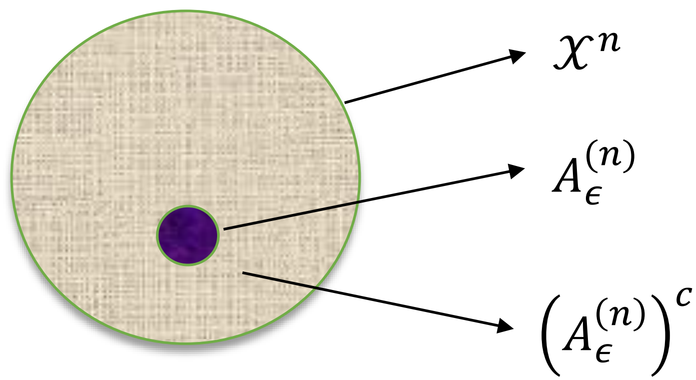
$$2^{-n(H(X)+\epsilon)} \leq p(x_1, x_2, \ldots, x_n) \leq 2^{-n(H(X)-\epsilon)}$$

3. $|A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$, where $|A|$ denotes the number of elements in the set A.

Proof.

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

Thus, $|A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$



$$\frac{|A_\epsilon^{(n)}|}{|\mathcal{X}^n|} \leq 2^{n(H(X)-\log|\mathcal{X}|)} \to 0$$

$$\Pr(\mathcal{X}^n) \approx \Pr(A_\epsilon^{(n)})$$

Typical Set (cont'd)

The **typical set (典型集) $A_\epsilon^{(n)}$** with respect to $p(x)$ is the set of sequences $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ with the property

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

4. $\left|A_\epsilon^{(n)}\right| \geq (1-\epsilon)2^{n(H(X)-\epsilon)}$ for n sufficiently large.

Proof. For sufficiently large n, $\Pr\left\{A_\epsilon^{(n)}\right\} > 1-\epsilon$, so that

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

Thus $\left|A_\epsilon^{(n)}\right| \geq (1-\epsilon)2^{n(H(X)-\epsilon)}$

High Probability Set

- $A_\epsilon^{(n)}$ is a very tiny set that contains most of the probability; i.e., high probability set

Definition. For each $n = 1, 2, \dots$, let $B_\delta^{(n)} \subseteq \mathcal{X}^n$ be the smallest set with

$$\Pr\left\{B_\delta^{(n)}\right\} \geq 1 - \delta$$

Theorem. Let $X_1, X_2, \dots, X_n$ be i.i.d $\sim p(x)$. For $\delta < \frac{1}{2}$ and any $\delta' > 0$, if $\Pr\left\{B_\delta^{(n)}\right\} \geq 1 - \delta$, then

$$\frac{1}{n}\log|B_\delta^{(n)}| > H - \delta' \text{ for } n \text{ sufficiently large.}$$

- Intuition: As $A_\epsilon^{(n)}$ has $2^{n(H\pm\epsilon)}$ elements, $|B_\delta^{(n)}|$ and $|A_\epsilon^{(n)}|$ are equal to the first order in the exponent
- Proof: (exercise 3.11)
  - For any two sets A, B, if $\Pr(A) \geq 1 - \epsilon_1$ $\Pr(B) \geq 1 - \epsilon_2$, then
    $$\Pr(A \cap B) > 1 - \epsilon_1 - \epsilon_2$$
  - $1 - \epsilon - \delta \leq \Pr\left(A_\epsilon^{(n)} \cap B_\delta^{(n)}\right) = \sum_{A_\epsilon^{(n)} \cap B_\delta^{(n)}} p(x^n) \leq \sum_{A_\epsilon^{(n)} \cap B_\delta^{(n)}} 2^{-n(H-\epsilon)}$
    $= \left|A_\epsilon^{(n)} \cap B_\delta^{(n)}\right| 2^{-n(H-\epsilon)} \leq \left|B_\delta^{(n)}\right| 2^{-n(H-\epsilon)}$
  - $\left|B_\delta^{(n)}\right| \geq \left|A_\epsilon^{(n)} \cap B_\delta^{(n)}\right| \geq 2^{n(H-\epsilon)}(1 - \epsilon - \delta)$

Data Compression: Problem Formulation

$$X^n = (X_1, \dots, X_n) \longrightarrow \boxed{\textbf{Encoder}} \xrightarrow{m \text{ bits}} \boxed{\textbf{Decoder}} \longrightarrow \hat{X}^n$$

(Data compression/Source coding) For a source sequence, we seek to find a **shorter encoding** for them:

“苟利国家生死以” $\rightarrow \{00, 01, 1, 110, 111, 010, 1010\}$

“government of the people, by the people, for the people” $\rightarrow \{\dots\dots\}$

Problem definition:

- Source: $X_1, X_2, \dots,$ are i.i.d. $\sim p(X)$. Source sequences: $X^n = (X_1, \dots, X_n)$ denotes the n-tuple that represents a sequence of n source symbols
- Alphabet: $\mathcal{X} = \{1, 2, \dots, |\mathcal{X}|\}$ – the possible values that each X_i can take on
- Encoder and decoder are a pair of functions f, g such that

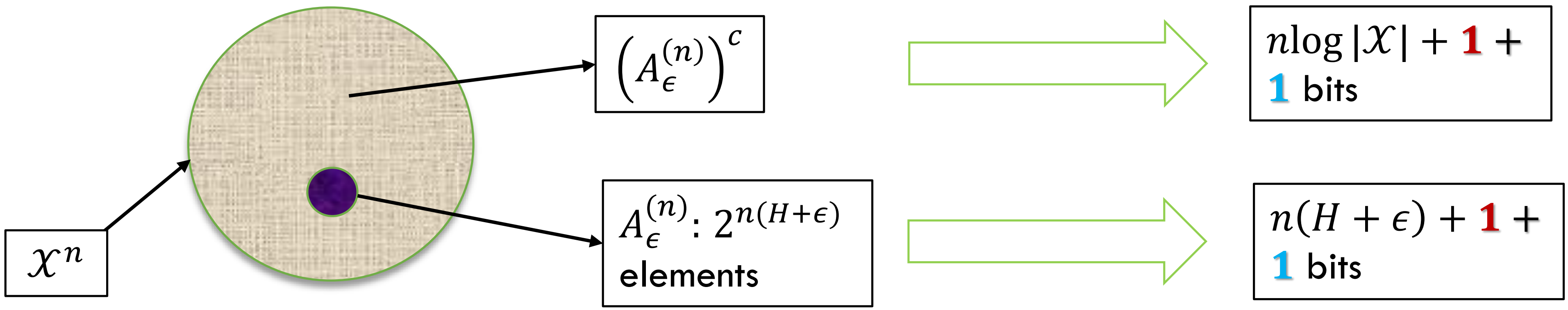
$$f: \mathcal{X} \rightarrow \{0,1\}^* \text{ and } g: \{0,1\}^* \rightarrow \mathcal{X}$$

- Probability of error $\boldsymbol{P_e = P(X^n \neq \hat{X}^n)}$

  If $P_e = 0$, “lossless”, otherwise “lossy”
- The rate of a scheme: $R = \dfrac{m}{n}$ ($R = \log|\mathcal{X}|$ is trivial!)

ToDo: Find an encoder and decoder pair such that $P_e \rightarrow 0$, as $n \rightarrow \infty$

Data Compression: Procedure



Divide and conquer: $x^n \in A_\epsilon^{(n)}$ and $x^n \notin A_\epsilon^{(n)}$

- $x^n \in A_\epsilon^{(n)}$:
  - Since there are $\leq\ 2^{n(H+\epsilon)}$ sequences in $A_\epsilon^{(n)}$, the indexing requires no more than $n(H\ +\epsilon)\ +\ 1$ bits. [The extra bit may be necessary because $n(H\ +\epsilon)$ may not be an integer.]
- $x^n \notin A_\epsilon^{(n)}$:
  - Similarly, we can index each sequence not in $A_\epsilon^{(n)}$ by using not more than $n\log|X|\ +\ 1$ bits.
- To deal with overlap in the $\{0{,}1\}$ sequences
  - We prefix all these sequences by a 0, giving a total length of $\leq\ n(H\ +\epsilon) +\ 2$ bits to represent each sequence in $A_\epsilon^{(n)}$
  - Prefixing these indices by 1, we have a code for all the sequences in $\mathcal{X}^n$.

Data Compression: Analysis

Homework
Cover: 3.1, 3.3, 3.4

$$
\begin{aligned}
E\big(l(X^n)\big) &= \sum_{x^n} p(x^n) l(x^n) \\
&= \sum_{x^n \in A_\epsilon^{(n)}} p(x^n) l(x^n) + \sum_{x^n \notin A_\epsilon^{(n)}} p(x^n) l(x^n) \\
&\le \sum_{x^n \in A_\epsilon^{(n)}} p(x^n)(n(H+\epsilon)+2) + \sum_{x^n \notin A_\epsilon^{(n)}} p(x^n)(n \log|\mathcal{X}| + 2\,) \\
&= \Pr\left\{A_\epsilon^{(n)}\right\}(n(H+\epsilon)+2) + \Pr\left\{\left(A_\epsilon^{(n)}\right)^c\right\}(n \log|\mathcal{X}| + 2) \\
&\le n(H+\epsilon) + \epsilon n(\log|\mathcal{X}|) + 2 \\
&= n(H+\epsilon')
\end{aligned}
$$

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

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

(Converse). For any scheme with rate $r < H(X)$, $P_e \to 1$

Let $r = H(X) - \epsilon$. For any scheme with rate r, it can encode at most 2^{nr} different symbols in $\mathcal{X}^n$. The correct decoding probability is $\approx 2^{nr} 2^{-nH} = 2^{-n(H-r)} \to 0$

$$P_e \to 1$$

Summary

All the materials can be found at:

- T. Cover : Ch. 3