INFORMATION THEORY & CODING

Week 4 : Asymptotic Equipartition Property (AEP)

Dr. Rui Wang

Department of Electrical and Electronic Engineering

Southern Univ. of Science and Technology (SUSTech)

Email: wang.r@sustech.edu.cn

September 29, 2020

Review Summary

Inequalities related to D and I

1. $D(p\|q) \geq 0$ with equality iff $p(x) = q(x)$, for all $x \in \mathcal{X}$ (*information inequality*).
2. $I(X;Y) = D(p(x,y)\|p(x)p(y)) \geq 0$, with equality iff $p(x,y) = p(x)p(y)$ (i.e., X and Y are independent).
3. If $|\mathcal{X}| = m$, and u is the uniform distribution over $\mathcal{X}$, then $D(p\|u) = \log m - H(p)$.

Jensen's Inequality

If f is a convex function, then $E[f(X)] \geq f(E[X])$.

Data-processing inequality

If $X \to Y \to Z$ forms a Markov chain, then $I(X;Y) \geq I(X;Z)$.

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Fano's inequality

Problem 2.5 (*Zero conditional entropy*)

Show that if $H(X|Y) = 0$, then X is a function of Y, i.e., for all y with $p(y) > 0$, there is only one possible value of x with $p(x, y) > 0$.

Proof.

Assume that there exists an y, say y_0 and two different values of x, say x_1 and x_2 such that $p(y_0, x_1) > 0$ and $p(y_0, x_2) > 0$. Then $p(y_0) \geq p(y_0, x_1) + p(y_0, x_2) > 0$, and $p(x_1|y_0)$ and $p(x_2|y_0)$ are not equal to 0 or 1. Thus,

$$\begin{aligned} H(X|Y) &= -\sum_y p(y) \sum_x p(x|y) \log p(x|y) \\ &\geq p(y_0)\left(-p(x_1|y_0) \log p(x_1|y_0) - p(x_2|y_0) \log p(x_2|y_0)\right) \\ &> 0 \end{aligned}$$

since $-t \log t \geq 0$ for $0 \leq t \leq 1$, and is strictly positive for $t \neq 0, 1$, which is a contradiction to $H(X|Y) = 0$. □

Fano's inequality

- If $H(X|Y) = 0$, X is a function of Y. we can estimate X from Y with zero probability of error.

- When $H(X|Y)$ is not zero, our estimate $\hat{X}$ may be wrong. Define
$$P_e = \Pr[\hat{X} \neq X],$$
as the detection error probability, we want to connect P_e with $H(X|Y)$.

Fano's inequality

Theorem 2.10.1

For any estimator $\hat{X}$ such that $X \to Y \to \hat{X}$, with $P_e = \Pr\{X \neq \hat{X}\}$, we have

$$H(P_e) + P_e \log(|\mathcal{X}| - 1) \geq H(X|\hat{X}) \geq H(X|Y).$$

This inequality can be weakened to

$$1 + P_e \log(|\mathcal{X}| - 1) \geq H(X|Y)$$

or

$$P_e \geq \frac{H(X|Y) - 1}{\log |\mathcal{X}| - 1}.$$

Fano's inequality

Theorem 2.10.1

For any estimator $\hat{X}$ *such that* $X \to Y \to \hat{X}$ *, with* $P_e = \Pr\{X \neq \hat{X}\}$*, we have*

$$H\left(P_e\right) + P_e \log(|\mathcal{X}| - 1) \geq H(X|\hat{X}) \geq H(X|Y).$$

Proof.

Define an error random variable as

$$E = \begin{cases} 1 & \text{if } \hat{X} \neq X, \\ 0 & \text{if } \hat{X} = X. \end{cases}$$

Using the chain rule for entropies to expand $H(E, X|\hat{X})$ in two different ways, we have

$$H(E, X|\hat{X}) = H(X|\hat{X}) + \underbrace{H(E|X, \hat{X})}_{=0} = \underbrace{H(E|\hat{X})}_{\leq H(P_e)} + \underbrace{H(X|E, \hat{X})}_{\leq P_e \log(|X|-1)} .$$

Since conditioning reduces entropy, $H(E|\hat{X}) \leq H(E) = H(P_e)$. Since E is a function of X and $\hat{X}$, the conditional entropy $H(E|X, \hat{X})$ is equal to 0. We now look at $H(X|E, \hat{X})$. By the equation $H(X|Y) = \sum_y p(y)H(X|Y = y)$, we have

$$\begin{aligned} H(X|E, \hat{X}) = \sum_{\hat{x} \in \mathcal{X}} \{&\Pr[\hat{X} = \hat{x}, E = 0]H(X|\hat{X} = \hat{x}, E = 0) \\ &+ \Pr[\hat{X} = \hat{x}, E = 1]H(X|\hat{X} = \hat{x}, E = 1)\}. \end{aligned}$$

Fano's inequality

Theorem 2.10.1

For any estimator $\hat{X}$ *such that* $X \to Y \to \hat{X}$ *, with* $P_e = \Pr\{X \neq \hat{X}\}$*, we have*

$$H(P_e) + P_e \log(|\mathcal{X}| - 1) \geq H(X|\hat{X}) \geq H(X|Y).$$

Proof.

$$H(E, X|\hat{X}) = H(X|\hat{X}) + \underbrace{H(E|X, \hat{X})}_{=0} = \underbrace{H(E|\hat{X})}_{\leq H(P_e)} + \underbrace{H(X|E, \hat{X})}_{\leq P_e \log(|X|-1)} .$$

$$H(X|E, \hat{X}) = \sum_{\hat{x} \in \mathcal{X}} \{\Pr[\hat{X} = \hat{x}, E = 0] H(X|\hat{X} = \hat{x}, E = 0)$$
$$+ \Pr[\hat{X} = \hat{x}, E = 1] H(X|\hat{X} = \hat{x}, E = 1)\}.$$

By definition of E, X is conditionally deterministic given $\hat{X} = \hat{x}$ and $E = 0$, then $H(X|\hat{X} = \hat{x}; E = 0) = 0$. If $\hat{X} = \hat{x}$ and $E = 1$, then X must take a value in the set $\{x \in \mathcal{X} : x \neq x\hat{x}\}$ which contains $|\mathcal{X}| - 1$ elements. Then $H(X|\hat{X} = \hat{x}, E = 1) \leq \log(|\mathcal{X}| - 1)$.

$$\begin{aligned} H(X|E, \hat{X}) &\leq \sum_{\hat{x} \in \mathcal{X}} \Pr[\hat{X} = \hat{x}, E = 1] \log(|\mathcal{X}| - 1) \\ &= \Pr[E = 1] \log(|\mathcal{X}| - 1) \\ &= P_e \log(|\mathcal{X}| - 1) \end{aligned}$$

□

Fano's inequality

Theorem 2.10.1

For any estimator $\hat{X}$ *such that* $X \to Y \to \hat{X}$ *, with* $P_e = \Pr\{X \neq \hat{X}\}$*, we have*

$$H(P_e) + P_e \log(|\mathcal{X}| - 1) \geq H(X|\hat{X}) \geq H(X|Y).$$

Proof.

$$H(E, X|\hat{X}) = H(X|\hat{X}) + \underbrace{H(E|X, \hat{X})}_{=0} = \underbrace{H(E|\hat{X})}_{\leq H(P_e)} + \underbrace{H(X|E, \hat{X})}_{\leq P_e \log(|X|-1)} \quad .$$

$$\begin{aligned} H(X|E, \hat{X}) = \sum_{\hat{x} \in \mathcal{X}} \{&\Pr[\hat{X} = \hat{x}, E = 0] H(X|\hat{X} = \hat{x}, E = 0) \\ &+ \Pr[\hat{X} = \hat{x}, E = 1] H(X|\hat{X} = \hat{x}, E = 1)\}. \end{aligned}$$

$$\begin{aligned} H(X|E, \hat{X}) &\leq \sum_{\hat{x} \in \mathcal{X}} \Pr[\hat{X} = \hat{x}, E = 1] \log(|\mathcal{X}| - 1) \\ &= \Pr[E = 1] \log(|\mathcal{X}| - 1) \\ &= P_e \log(|\mathcal{X}| - 1) \end{aligned}$$

By the data-processing inequality, we have $I(X; \hat{X}) \leq I(X; Y)$ and therefore $H(X|\hat{X}) \geq H(X|Y)$. □

Corollary

For any two random variables X and Y, let $p = Pr(X \neq Y)$.

$$H(p) + p\log(|\mathcal{X}| - 1) \geq H(X|Y).$$

Proof.

Let $\hat{X} = Y$ in Fano's inequality. $\square$

Fano's inquality

Remark

Suppose that there is no knowledge of Y. Thus, X must be guessed without any information. Let $X \in \{1, 2, \ldots, m\}$ and $p_1 \geq p_2 \geq \cdots \geq p_m$. Then the best guess of X is $\hat{X} = 1$ and the resulting probability of error is $P_e = 1 - p_1$. Fano's inquality becomes

$$H(P_e) + P_e \log(m-1) \geq H(X).$$

The probability mass function

$$(p_1, p_2, \cdots, p_m) = \left(1 - P_e, \frac{P_e}{m-1}, \cdots, \frac{P_e}{m-1}\right)$$

achieves this bound with equality.

Applications of Fano's inequality

- Prove converse in many theorems (including channel capacity)

- Compressed sensing signal model

$$y = Ax + w$$

where $A \in \mathcal{R}^{M \times d}$: projection matrix for dimension reduction. Signal x is sparse. Want to estimate x from y.

Fano's inequality

Lemma 2.10.1

If X and X' are i.i.d. with entropy $H(X)$,

$$\Pr[X = X'] \geq 2^{-H(X)},$$

with equality iff X has a uniform distribution.

Corollary

Let X, X' be independent with $X \sim p(x)$, $X' \sim r(x)$, $x, x' \in \mathcal{X}$. Then

$$\Pr\left[X = X'\right] \geq 2^{-H(p)-D(p\|r)}$$

$$\Pr\left[X = X'\right] \geq 2^{-H(r)-D(r\|p)}$$

Please refer to P40 of the textbook for the proof.

Stock Market

- Initial investment Y_0, daily return ratio r_i, in t-th day, your money is

$$Y_t = Y_0 r_1 \cdot \ldots \cdot r_t.$$

- Now if returns ratio r_i are i.i.d., with

$$r_i = \begin{cases} 4, & \text{w.p. } 1/2 \\ 0, & \text{w.p. } 1/2 \end{cases}$$

- So you think the expected return ratio is $E[r_i] = 2$.
- And then

$$E[Y_t] = E[Y_0 r_1 \cdot \ldots \cdot r_t] = Y_0 (E[r_i])^t = Y_0 2^t ???$$

Stock Market

- With $Y_0 = 1$, actual return Y_t goes like

$$1 \qquad 4 \qquad 16 \qquad 0 \qquad 0 \qquad \cdots$$

- Why?
    - The 'typical' sequences will end up with 0 return.
    - Occasionally, we got high return.
    - The expected return is increasing.
    - Expectation does not show the typical feature of this random sequence. We can turn to typical set.

Weak Law of Large Numbers

Theorem (Weak Law of Large Numbers)

Suppose that $X_1, X_2, \ldots, X_n$ are n independent, identically distributed (i.i.d.) random variables, then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \rightarrow E[X] \qquad \text{in probability},$$

i.e. for every number $\epsilon > 0$,

$$\lim_{n\rightarrow\infty} \Pr\left[|\frac{1}{n}\sum_{i=1}^{n} X_i - E[X]| \leq \epsilon \right] = 1.$$

Asymptotic Equipartition Property (AEP)

Definition (Convergence of random variables)

Given a sequence of random variables, $X_1, X_2, \ldots$, we say that the sequence $X_1, X_2, \ldots$ converges to a random variable X:

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

Asymptotic Equipartition Property (AEP)

Theorem 3.1.1 (AEP)

If $X_1, X_2, \ldots$ are i.i.d. $\sim p(x)$, then

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

Proof.

Since X_i are i.i.d., so are $\log p(X_i)$. Hence, by the weak law of large numbers,

$$\begin{aligned} -\frac{1}{n}\log p\left(X_1, X_2, \ldots, X_n\right) &= -\frac{1}{n}\sum_i \log p\left(X_i\right) \\ &\rightarrow -E[\log p(X)] \qquad \text{in probability} \\ &= H(X) \end{aligned}$$

$\square$

Typical Set

Definition

A *typical set* $A_\epsilon^{(n)}$ contains all sequence realizations $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ with

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

Consequences of AEP

Theorem 3.1.2

- *If* $(x_1, x_2, \ldots, x_n) \in A_\epsilon^{(n)}$*, then* $H(X) - \epsilon \le -\frac{1}{n} \log p(x_1, x_2, \ldots, x_n) \le H(X) + \epsilon$.
- $\Pr[A_\epsilon^{(n)}] > 1 - \epsilon$ *for n sufficiently large.*
- $|A_\epsilon^{(n)}| \le 2^{n(H(X)+\epsilon)}$*, where* $|A|$ *denotes the cardinality of the set* A*.*
- $|A_\epsilon^{(n)}| \ge (1-\epsilon)2^{n(H(X)-\epsilon)}$ *for n sufficiently large.*

Proof.

1. Immediate from the definition of $A_\epsilon^{(n)}$. □

The number of bits used to describe sequences in typical set is approximately $nH(X)$.

Consequences of AEP

Theorem 3.1.2

- *If* $(x_1, x_2, \ldots, x_n) \in A_\epsilon^{(n)}$*, then* $H(X) - \epsilon \le -\frac{1}{n} \log p(x_1, x_2, \ldots, x_n) \le H(X) + \epsilon.$
- $\Pr[A_\epsilon^{(n)}] > 1 - \epsilon$ *for n sufficiently large.*
- $|A_\epsilon^{(n)}| \le 2^{n(H(X)+\epsilon)}$*, where* $|A|$ *denotes the cardinality of the set A.*
- $|A_\epsilon^{(n)}| \ge (1-\epsilon)2^{n(H(X)-\epsilon)}$ *for n sufficiently large.*

Proof.

2. By Theorem 3.1.1, 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 $n \ge n_0$, 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 $\delta = \epsilon$, the conclusion follows. □

Consequences of AEP

Theorem 3.1.2

- *If* $(x_1, x_2, \ldots, x_n) \in A_\epsilon^{(n)}$, *then* $H(X) - \epsilon \le -\frac{1}{n}\log p(x_1, x_2, \ldots, x_n) \le H(X) + \epsilon$.
- $\Pr[A_\epsilon^{(n)}] > 1 - \epsilon$ *for* n *sufficiently large.*
- $|A_\epsilon^{(n)}| \le 2^{n(H(X)+\epsilon)}$, *where* $|A|$ *denotes the cardinality of the set* A.
- $|A_\epsilon^{(n)}| \ge (1-\epsilon)2^{n(H(X)-\epsilon)}$ *for* n *sufficiently large.*

Proof.

3.

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

□

Consequences of AEP

Theorem 3.1.2

- *If* $(x_1, x_2, \ldots, x_n) \in A_\epsilon^{(n)}$, *then* $H(X) - \epsilon \le -\frac{1}{n} \log p(x_1, x_2, \ldots, x_n) \le H(X) + \epsilon.$
- $\Pr[A_\epsilon^{(n)}] > 1 - \epsilon$ *for n sufficiently large.*
- $|A_\epsilon^{(n)}| \le 2^{n(H(X)+\epsilon)}$, *where* $|A|$ *denotes the cardinality of the set* A.
- $|A_\epsilon^{(n)}| \ge (1-\epsilon)2^{n(H(X)-\epsilon)}$ *for n sufficiently large.*

Proof.

4. For sufficiently large n, $\Pr[A_\epsilon^{(n)}] > 1 - \epsilon$, so that

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

□

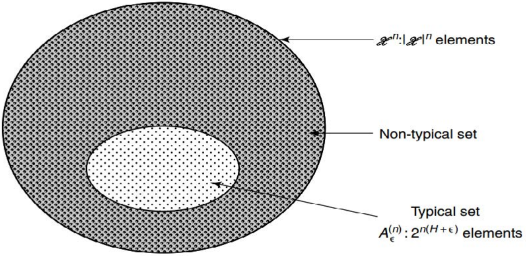
Typical set diagram

This enables us to divide all sequences into two sets

- Typical set: high probability to occur, sample entropy is close to true entropy

  so we will focus on analyzing sequences in typical set

- Non-typical set: small probability, can ignore in general



Asymptotic Equipartition Property (AEP)

Theorem 3.2.1

Let $X_1, X_2, ..., X_n$ be i.i.d. random variables with distribution $p(x)$, and $X^n = X_1 X_2 ... X_n$. For arbitrarily small $\epsilon > 0$, there exists a code that maps every realization $x^n = x_1 x_2 ... x_n$ of X^n into one binary string, such that the mapping is one-to-one (and therefore invertible) and

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

for a sufficiently large n.

Asymptotic Equipartition Property (AEP)

Theorem 3.2.1

$$E\left[\frac{1}{n}\ell(X^n)\right] \leq H(X) + \epsilon.$$

for n sufficiently large.

Proof.

Description in typical set requires no more than $n(H(X)+\epsilon)+1$ bits (correction of 1 bit because of integrality).

Description in atypical set ${A_\epsilon^{(n)}}^C$ requires no more than $n\log|\mathcal{X}|+1$ bits.

Add another bit to indicate whether in $A_\epsilon^{(n)}$ or not to get whole description. $\square$

Asymptotic Equipartition Property (AEP)

Theorem 3.2.1

$$E[\frac{1}{n}\ell(X^n)] \leq H(X) + \epsilon.$$

for n sufficiently large.

Proof.

Let $\ell(x^n)$ be the length of the binary description of x^n. Then, $\forall \epsilon > 0$, there exists n_0 s.t. $\forall n > n_0$,

$$\begin{aligned}
E\left(\ell\left(X^n\right)\right) &= \sum_{x^n} p\left(x^n\right)\ell\left(x^n\right) \\
&= \sum_{x^n \in A_\epsilon^{(n)}} p\left(x^n\right)\ell\left(x^n\right) + \sum_{x^n \in A_\epsilon^{(n)C}} p\left(x^n\right)\ell\left(x^n\right) \\
&\leq \sum_{x^n \leq A_\epsilon^{(n)}} p\left(x^n\right)\left(n(H+\epsilon)+2\right) + \sum_{x^n \in A_\epsilon^{(n)C}} p\left(x^n\right)\left(n\log|\mathcal{X}|+2\right) \\
&= \Pr[A_\epsilon^{(n)}](n(H+\epsilon)+2) + \Pr[A_\epsilon^{(n)^C}](n\log|\mathcal{X}|+2) \\
&\leq n(H+\epsilon) + \epsilon n(\log|\mathcal{X}|) + 2 \\
&= n\left(H+\epsilon'\right)
\end{aligned}$$

where $\epsilon' = \epsilon + \epsilon\log|\mathcal{X}| + \frac{2}{n}$ can be made arbitrarily small by choosing n properly. $\square$

Reading & Homework

Reading : 2.10 and whole Chapter 3

Homework : Problems 2.32, 3.8, 3.10