HYKE Weekly Seminar

Chapter 11. Information Theory and Statistics

Youngjae Lee

June 25, 2021

Introduction

Today's goal is to describe the asymptotic behavior of i.i.d. random variables. There are three ways to describe this behavior.

1. Law of Large Numbers

$$\frac{1}{n}(X_1 + \cdots X_n) \to \mu$$

2. Central Limit Theorem

$$\frac{1}{\sqrt{n}}(X_1 + \cdots + X_n) \rightharpoonup N(\mu, \sigma^2)$$

3. Large Deviation Principle

$$P\left(\left|\frac{1}{n}(X_1 + \cdots X_n) - \mu\right| > \epsilon\right) \to e^{-nr(\epsilon)}$$

Sections

Method of Types

Notations

- $X_1, X_2, \cdots, X_n$: sequence of n symbols from an alphabet $\chi = \{a_1, a_2, \cdots, a_{|\chi|}\}$
- x^n or $\mathbf{x}$: sequence $x_1, x_2, \cdots, x_n$
  but I will only use $\mathbf{x}$ whenever possible.
- X^n: sequence $X_1, X_2, \cdots, X_n$

Type

Definition (Type)

The **type** $P_{\mathbf{x}}$ of a sequence $\mathbf{x} = x_1, x_2, \cdots, x_n$ is the relative proportion of occurences of each symbol of χ. That is,

$$P_{\mathbf{x}}(a) = \frac{\text{the number of } a \text{ in } \mathbf{x}}{n}$$

Example

Let $\chi = \{a, b, \cdots, z\}$, $\mathbf{x} = ajxaxpekbgjsazz$(15 letters). Then

$$P_{\mathbf{x}}(a) = \frac{3}{15}, P_{\mathbf{x}}(b) = \frac{1}{15}, \cdots P_{\mathbf{x}}(z) = \frac{2}{15}.$$

Types with Denominator n

Definition

$\mathcal{P}_n$ denote the set of **types with denominator** n.

Example

If $\chi = \{0, 1\}$. Then

$$\mathcal{P}_n = \left\{ (P(0), P(1)) : \left(\frac{0}{n}, \frac{n}{n}\right), \left(\frac{1}{n}, \frac{n-1}{n}\right), \cdots, \left(\frac{n}{n}, \frac{0}{n}\right) \right\}.$$

Example

$\chi = \{a, b, \cdots, z\}$. Then $\mathcal{P}_n$ is the set of functions

$$P_{\mathbf{x}} : \chi \to \left\{ \frac{0}{n}, \frac{1}{n}, \cdots, \frac{n}{n} \right\} \quad \text{with} \quad \sum_{a \in \chi} P_{\mathbf{x}}(a) = 1.$$

Type class of P

Definition

If $P \in \mathcal{P}_n$, the set of sequences of length n and type P is called the **type class** of P, denoted $T(P)$:

$$T(P) = \{\mathbf{x} \in \chi^n : P_{\mathbf{x}} = P\}$$

Type class of P Cont.

Example

Let $\chi = \{a, b, c\}$ and $\mathbf{x} = aacba$. Then the type $P_{\mathbf{x}}$ is

$$P_{\mathbf{x}}(a) = \frac{3}{5}, \quad P_{\mathbf{x}}(b) = \frac{1}{5}, \quad P_{\mathbf{x}}(c) = \frac{1}{5}.$$

$T(P_{\mathbf{x}})$ is the set of $\mathbf{y} \in \mathcal{P}_5$ such that $P_{\mathbf{x}} = P_{\mathbf{y}}$. That is,

$$T(P_{\mathbf{x}}) = \{aaabc, aaacb, aabac, \cdots, cbaaa\}.$$

The number of elements in $T(P_{\mathbf{x}})$ is

$$|T(P)| = \binom{5}{3,1,1} = \frac{5!}{3!1!1!} = 20.$$

Main Theorem: 11.1.1

In fact, we can bound the number of types is by polynomial in n.

Theorem

$$|\mathcal{P}_n| \leq (n+1)^{|\chi|}$$

Proof.

Recall that $\mathcal{P}_n$ is the set of functions

$$P_{\mathbf{x}} : \chi \to \left\{\frac{0}{n}, \frac{1}{n}, \cdots, \frac{n}{n}\right\} \quad \text{with} \quad \sum_{a \in \chi} P_{\mathbf{x}}(a) = 1.$$

Therefore, $\mathcal{P}_n \subset \left\{\frac{0}{n}, \frac{1}{n}, \cdots, \frac{n}{n}\right\}^{\chi}$ and hence

$$|\mathcal{P}_n| \leq (n+1)^{|\chi|}. \qquad \square$$

Remark

There are only polynomial number($\leq (n+1)^{\chi}$) of types of length n, while there are exponential number($|\chi|^n$) of sequence in n.

From now on, we assume that the sequence $\mathbf{x} = X^n = X_1, X_2, \cdots, X_n$ is drawn i.i.d. $\sim Q(x)$. Let

$$Q^n(\mathbf{x}) = \prod_{i=1}^{n} Q(x_i)$$

denote the product distribution associated with Q. The next theorem shows that the Q^n can be accurately described in terms of entropy.

Theorem 11.1.2

Theorem

If $X_1, X_2, \cdots, X_n$ are drawn i.i.d. according to $Q(x)$, the probability of $\mathbf{x}$ *depends only on its type and is given by*

$$Q^n(\mathbf{x}) = 2^{-n(H(P_\mathbf{x})+D(P_\mathbf{x}||Q))}$$

Proof

$$
\begin{aligned}
Q^n(\mathbf{x}) &:= \prod_{i=1}^{n} Q(x_i) \\
&= \prod_{a \in \chi} Q(a)^{N(a|\mathbf{x})} \\
&= \prod_{a \in \chi} Q(a)^{nP_{\mathbf{x}}(a)} \\
&= \prod_{a \in \chi} 2^{nP_{\mathbf{x}}(a) \log Q(a)} \\
&= \prod_{a \in \chi} 2^{nP_{\mathbf{x}}(a) \log Q(a) - P_{\mathbf{x}}(a) \log P_{\mathbf{x}}(a) + P_{\mathbf{x}}(a) \log P_{\mathbf{x}}(a)} \\
&= 2^{n \sum_{a \in \chi} (-P_{\mathbf{x}}(a) \log \frac{P_{\mathbf{x}}(a)}{Q(a)} + P_{\mathbf{x}}(a) \log P_{\mathbf{x}}(a))} \\
&= 2^{n(-D(P_{\mathbf{x}}||Q) - H(P_{\mathbf{x}}))}
\end{aligned}
$$

Corollary

Corollary

If $\mathbf{x}$ *is the type class of* Q*, then*

$$Q^n(\mathbf{x}) = 2^{-nH(Q)}.$$

Proof.

If $\mathbf{x} \in T(Q)$, then $P_{\mathbf{x}} = Q$. Therefore,

$$Q^n(\mathbf{x}) = 2^{-n(H(P_{\mathbf{x}}))+D(P_{\mathbf{x}}||Q)} = 2^{-n(H(P_{\mathbf{x}}))} = 2^{-nH(Q)} \qquad \square$$

Theorem

Theorem (Size of a type class $T(P)$)

For any type $P \in \mathcal{P}_n$,

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{nH(P)} \leq |T(P)| \leq 2^{nH(P)}.$$

Theorem (Probability of type class)

For any $P \in \mathcal{P}_n$ and any disribution Q, the probability of the type class $T(P)$ under Q^n is $2^{-nD(P||Q)}$ to first order in the exponent. More precisely,

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-nD(P||Q)} \leq Q^n(T(P)) \leq 2^{-nD(P||Q)}.$$

Summary

- $|\mathcal{P}_n| \leq (n+1)^{|\chi|}$
- $Q^n(\mathbf{x}) = 2^{-n(D(P_\mathbf{x}||Q+H(P_\mathbf{x})))}$
- $|T(P)| \approx 2^{nH(P)}$
- $Q^n(T(P)) \approx 2^{-nD(P||Q)}$

Law of Large Numbers

Review of Chapter 3

Theorem (AEP)

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

$$-\frac{1}{n}\log p(X_1, X_2, \cdots, X_n) \to H(X) \quad \textit{in probability.}$$

Definition (Typical Set $A_\epsilon^{(n)}$)

The **typical set** $A_\epsilon^{(n)}$ w.r.t. $p(x)$ is the set of sequences $(x_1, x_2, \cdots, x_n) \in \chi^n$ with the property

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

Typical Set T_Q^ϵ

Definition

Given an $\epsilon > 0$ we can define a **typical set** T_Q^ϵ of sequences for the distribution Q^n as

$$T_Q^\epsilon = \{\mathbf{x} : D(P_\mathbf{x}||Q) \leq \epsilon.\}$$

Proposition

The probability that $\mathbf{x}$ is not typical is

$$1 - Q^n(T_Q^\epsilon) \leq 2^{-n\left(\epsilon - |\chi| \frac{\log(n+1)}{n}\right)},$$

which goes to 0 as $n \to \infty$. Hence $\Pr\{\mathbf{x} \in T_Q^\epsilon\} \to 1$ as $n \to \infty$.

Proof of the Proposition

$$
\begin{aligned}
1 - Q^n(T_Q^\epsilon) &= \sum_{P: D(P||Q) > \epsilon} Q^n(T(P)) \\
&\leq \sum_{P: D(P||Q) > \epsilon} 2^{-nD(P||Q)} \\
&\leq \sum_{P: D(P||Q) > \epsilon} 2^{-n\epsilon} \\
&\leq (n+1)^{|\chi|} 2^{-n\epsilon} \\
&= 2^{-n(\epsilon - |\chi| \frac{\log(n+1)}{n})}
\end{aligned}
$$

Theorem 11.2.1

Theorem

Let $X_1, X_2, \cdots, X_n$ be i.i.d. $\sim P(x)$. Then

$$Pr\{D(P_{\mathbf{x}}||P) > \epsilon\} \leq 2^{-n\left(\epsilon - |\chi| \frac{\log(n+1)}{n}\right)},$$

and consequentely, $D(P_{\mathbf{x}}||P) \to 0$ with probability 1.

Remark

In chapter 3, we proved

$$Pr\left\{A_{\epsilon}^{(n)}\right\} > 1 - \epsilon \quad \text{as } n \to \infty.$$

Strong Typical Set

Definition (Strong Typical Set $A_\epsilon^{*(n)}$)

$$A_\epsilon^{*(n)} = \left\{ \mathbf{x} \in \chi^n : \begin{array}{ll} \left| \frac{1}{n} N(a|\mathbf{x}) - P(a) \right| < \frac{\epsilon}{|\chi|}, & \text{if } P(a) > 0 \\ N(a|\mathbf{x}) = 0 & \text{if } P(a) = 0 \end{array} \right\}.$$

By the strong law of large numbers, $\Pr\left\{A_\epsilon^{*(n)}\right\} \to 1$ as $n \to \infty$. The strong typical set is useful in proving stronger results such as universal coding, rate distortion theory, and large deviation theory.

Universal Source Coding

Review of Chapter 5

In chapter 5 we studied Huffman coding. It compresses an i.i.d. D-ary source with a **known** distribution $p(x)$ with entropy

$$H_D(X) \leq L^* := \min_{\sum D^{-l_i} \leq 1} \sum p_i l_i < H_D(X) + 1.$$

If the code is designed for incorrect distribution $q(x)$, a penalty of $D(p||q)$ is incurred:

$$H(p) + D(p||q) \leq L = \sum p(x) l(x) < H(p) + D(p||q) + 1.$$

Thus, Huffman coding is sensitive to the assumed distribution.

Motivation

Motivation

What compression can be achieved if the true distribution $p(x)$ is unknown? Is there a universal code such that $H(X) < R$?

Answer

Yes! Details are provided in chapter 13.

Fixed-rate block code of rate R

Definition

A **fixed-rate block code** of rate R for a source $X_1, X_2, \cdots, X_n$ which has an unknown distribution Q consists of two mappings:

- the encoder $f_n : \chi^n \to \{1, 2, \cdots, 2^{nR}\}$
- the decoder $\phi_n : \{1, 2, \cdots, 2^{nR}\} \to \chi^n$

Here R is called the **rate** of the code. The probability of error for the code w.r.t. Q is

$$P_e^{(n)} = Q^n(\mathbf{x} : \phi_n(f_n(\mathbf{x})) \neq \mathbf{x}).$$

Universal Code

Definition

A rate R block code for a source will be called **universal** if

1. the functions f_n and ϕ_n do not depend on the distribution Q
2. $P_e^{(n)} \to 0$ as $n \to \infty$ if $H(Q) < R$.

Theorem 11.3.1

Theorem

There exists a sequence of $(2^{nR}, n)$ universal source codes such that $P_e^{(n)} \to 0$ for every source Q such that $H(Q) < R$.

Remark

Smaller R reduces the number of Qs that satisfy $H(Q) < R$. That is, we have to set R large enough for meaningful results.

Remark

Compared to Huffman code, universal coding requires more code block(R). Therefore, if you are aware of the distribution $p(x)$, it is desirable to use Huffman codes.

Large Deviation Theory

Motivation

- Law of large numbers(LLN)
  describes the result of performing the same experiment a large number of times
- Centrl limit theorem(CLT)
  establishes that their properly normalized sum of i.i.d. random variable tends toward a normal distribution.
- Large Deviation Principle(LDP)
  concerns the asymptotic behaviour of remote tails of sequences of probability distributions

What about rate of convergence?

Large Deviation Principle

For i.i.d. random variables $X_1, \cdots, X_n$ the rate of convergence is exponential:

$$P\left(\left|\frac{1}{n}(X_1 + \cdots X_n) - \mu\right| > \epsilon\right) \to e^{-nr(\epsilon)}.$$

Surprisingly, in addition to the i.i.d. random variables, many probabilistic models have been proved to follow **exponential deacy**. In other words, it has been revealed that LDP is universal property.

Miscellaneous about Srinivasa Varadhan

- A unified formalization of large deviation theory was developed in 1966, in a paper by Varadhan.
- Varadhan won the Abel prize for
  "his fundamental contributions to probability theory and in particular for creating a unified theory of large deviation"
- Varadhan is professor. Insuk Seo's advisor.

Theorem 11.4.1

Theorem (Sanov's Theorem)

Let $X_1, X_2, \cdots, X_n$ be i.i.d. $\sim Q(x)$. Let $E \subset \mathcal{P}$ be a set of probability distributions. Then

$$Q^n(E) := Q^n(E \cap \mathcal{P}_n) \leq (n+1)^{|\mathcal{X}|} 2^{-nD(P^*||Q)},$$

where

$$P^* = \arg\min_{P \in E} D(P||Q)$$

is the distribution in E that is closest to Q in relative entropy.

If, in addition, the set E is the closure of its interior, then

$$\frac{1}{n} \log Q^n(E) \to -D(P^*||Q).$$

Proof (upper bound)

$$
\begin{aligned}
Q^n(E) &= \sum_{P \in E \cap \mathcal{P}_n} Q^n(T(P)) \\
&\leq \sum_{P \in \mathcal{P}_n} 2^{-nD(P||Q)} \\
&\leq \sum_{P \in \mathcal{P}_n} \max_{P \in E \cap \mathcal{P}_n} 2^{-nD(P||Q)} \\
&= \sum_{P \in \mathcal{P}_n} 2^{-n \min_{P \in E \cap \mathcal{P}_n} D(P||Q)} \\
&\leq \sum_{P \in \mathcal{P}_n} 2^{-n \min_{P \in E} D(P||Q)} \\
&= \sum_{P \in \mathcal{P}_n} 2^{-nD(P^*||Q)} \\
&\leq \sum_{P \in \mathcal{P}_n} (n+1)^{|\mathcal{X}|} 2^{-nD(P^*||Q)}
\end{aligned}
$$

Proof (lower bound)

Assume E is the closure of its interior. Since $\bigcup_n \mathcal{P}_n$ is dense in the set of all distributions, we can find a sequence of distributions $P_n \in E \cap \mathcal{P}_n$ with the property $D(P_n||Q) \to D(P^*||Q)$.

$$\begin{aligned} Q^n(E) &= \sum_{P \in E \cap \mathcal{P}_n} Q^n(T(P)) \\ &\geq Q^n(T(P_n)) \\ &\geq \frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-nD(P_n||Q)}. \end{aligned}$$

Remarks

General form of Sanov's theorem

$$-\inf_{x\in\mathring{\Gamma}} I(x) \le \liminf_{\epsilon\to 0} \epsilon \log \mu_\epsilon(\Gamma) \le \limsup_{\epsilon\to 0} \epsilon \log \mu_\epsilon(\Gamma) \le -\inf_{x\in\bar{\Gamma}} I(x),$$

where $\{\mu_\epsilon\}$ is a family of probability measures satisfying LDP, Γ is some measurable set, and I is a rate function.

Meaning of Sanov's theorem

In the language of large deviations theory, Sanov's theorem identifies the rate function for large deviations of the empirical measure of a sequence of i.i.d. random variables.

Examples of Sanov's Theorem

Example 1: Coin

Question

What is the approximate probability that the front will come out more than 700 times when the coin is thrown 1000 times?

Answer

From Sanov's theorem, the probability is

$$P(\overline{X}_n \geq 0.7) \approx 2^{-nD(P^*||Q)},$$

where P^* is the $(0.7, 0.3)$ distribution and Q is the $(0.5, 0.5)$ distribution. In this case,

$$\begin{aligned} D(P^*||Q) &= 1 - H(P^*) = 1 - H(0.7) \\ &= 1 + (0.7 \log 0.7 + 0.3 \log 0.3) = 0.119. \end{aligned}$$

Therefore, $P(\overline{X}_n \geq 0.7) \approx 2^{-119}$.

Useful Formula

Suppose that we wish to find

$$\Pr\left\{\frac{1}{n}\sum_{i=1}^{n} g_j(X_i) \geq \alpha_j,\ j = 1, 2, \cdots, k\right\}.$$

Then the set E is defined as

$$E = \left\{P : \sum_{a} P(a) g_j(a) \geq \alpha_j,\ j = 1, 2, \cdots, k\right\}.$$

To find the closest distribution in E to Q, use Lagrange multipliers:

$$J(P) = \sum_{x} P(x) \log \frac{P(x)}{Q(x)} + \sum_{i} \lambda_i \sum_{x} P(x) g_i(x) + \nu \sum_{x} P(x).$$

It follows that the closest distribution to Q is of the form

$$P^*(x) = \frac{Q(x) e^{\sum_i \lambda_i g_i(x)}}{\sum_{a \in \chi} Q(a) e^{\sum_i \lambda_i g_i(a)}}.$$

Example 2. Dice

Question

Suppose that we toss a fair die n times. What is the probability that the average of the thorws is greater than or equal to 4?

Observation

We wish to find

$$\Pr\left\{\frac{1}{n}\sum_{i=1}^{n} iP(i) \geq 4\right\}.$$

In this case, $Q(x) = \frac{1}{6}$, $k = 1$ and $g(a) = a$.

Dice Cont.

$$P^*(x) = \frac{Q(x)e^{\sum_i \lambda_i g_i(x)}}{\sum_{a\in\chi} Q(a)e^{\sum_i \lambda_i g_i(a)}}$$
$$= \frac{2^{\lambda x}}{\sum_{i=1}^{6} 2^{\lambda i}}$$

Solving numerically, we obtain $\lambda = 0.2519$,

$$P^* = (0.1031, 0.1227, 0.1461, 0.1740, 0.2072, 0.2468),$$

and therefore $D(P^*||Q) = 0.0624$. Thus, from Sanov's theorem, it follows that

$$\Pr\left\{\frac{1}{n}\sum_{i=1}^{n} iP(i) \geq 4\right\} \approx 2^{-0.0624n}.$$

Uncovered Sections

6. Conditional Limit Theorem
7. Hypothesis Teseting
8. Chernoff-Stein Lemma
9. Chernoff Information
10. Fisher Information and the Cramér-Rao Inequality

The End