4. Method of Types, Strong AEP and Universal Source Coding

Vaughan Sohn

October 21, 2024

Contents

Overview

- 지금까지 우리는 source data의 true distribution을 알고 있다는 가정안에서 optimal 인 source coding 방법에 대해서 소개했었다.
- 이번에는 실제 상황과 유사한, source의 true distribution을 모를 때의 source coding 을 살펴보려고 한다. $\rightarrow$ *Universal source coding theorem*
- Universal source coding theorem을 다루기 위하여 large deviation theory에서 사용되는 type이라는 개념을 도입하고자한다.
- Type을 사용하면 weak typical set의 *subset*인 strong typical set을 정의할 수 있고 strong typical set의 성질을 활용하여 universal source coding theorem을 증명하고자 한다.

Types

Limitation of weak typical set

- Lecture 3에서 소개한 weak typical set은 다음과 같이 정의된다.

$$A_\delta^{(n)}(P) = \{x^n \in \mathcal{X}^n \;:\; e^{-n(H(P)+\delta)} \leq P^n(x^n) \leq e^{-n(H(P)-\delta)}\}$$

- $X_i \sim \mathrm{Bern}(1/2)$인 상황을 가정하자. ($p(0) = 1/2, p(1) = 1/2$)
- 그렇다면, 어떠한 sequence에 대해서도 다음이 성립하게 된다.

$$e^{-n(H(P)+\delta)} \leq P^n(x^n) = 2^{-n} \leq e^{-n(H(P)-\delta)}$$

$\Rightarrow$ sample space 전체가 weak typical set에 해당하게 된다! (something strange ...)

- 또한 weak typical set에 있는 sample이라고해서 반드시 Empirical distribution이 true distribution과 유사다하는 보장이 존재하지 않는다.

Idea of strong typical set

어떤 sample x^n에서 각 symbol이 나타난 횟수가 실제 true distribution과 유사한 sample 들만을 모아서 strong typical set을 정의하고자한다.

$$N(a|x^n) \sim nP(a), \quad \forall a$$

Empirical discrete probability distribution

Definition 1 (empirical probability distribution)

For $x^n \in X^n$, we denoted **empirical distribution** of x^n as $\hat{P}_{x^n}$,

$$\hat{P}_{x^n}(a) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(x_i = a) = \frac{1}{n} N(a|x^n), \quad \forall a \in X$$

where $N(a|x^n)$ is the number of a 's in the given string x^n.

✓meaning: 실제로 우리가 얻을 수 있는 하나의 sample x^n에서 각 symbol이 등장한 횟수($*$)를 세어 추정한 probability distribution.
$\rightarrow$ 즉, empirical distribution은 x^n이라는 realized value에 의존한다.

Definition 2 (set of valid empirical probability distribution)

Let $\mathcal{P}_n$ be the set of all valid empirical distributions over X for a sequence of length n.

Example: $X = \{0, 1\}$

$$\mathcal{P}_n = \Big\{ \qquad \qquad \qquad \qquad \Big\}$$

Types

Definition 3 (type)

For a distribution $P \in \mathcal{P}_n$, the type T_P is a set of length n sequences that have empirical distribution P.

$$T_P \triangleq \{x^n \in \mathcal{X}^n \ : \ \hat{P}_{x^n} = P\}, \quad T_P \subseteq \mathcal{X}^n$$

✓<u>meaning</u>: 특정 empirical distribution을 가지는 x^n들의 집합.

- sequence에서 각 symbol들이 등장하는 횟수만 동일하다면, symbol의 순서가 달라도 동일한 type에 속하게 될 것이다. (e.g., $T_P = \{001, 010, 100\}$)
- Example: $X = \{0, 1\}, P = [\frac{k}{n}, \frac{n-k}{n}]$일 때, type의 크기는?

$$|T_P| = \qquad\qquad .$$

Types : size

Theorem 4 (number of types)

The number of types (=size of valid empirical probability distribution set) have loose upper bound:

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

Theorem 5 (size of each type)

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

$$|T_P| \doteq e^{nH(P)}$$

where $|\mathcal{X}| = M$ and $\mathcal{X} = \{a_1, a_2, \ldots, a_M\}$.

✓meaning: n이 커질수록, type의 크기가 증가하는 속도는 $H(P)$에 의해 조절된다.
rare한 event일수록 type의 크기가 증가하는 속도도 느리다!

- if $H(P) \downarrow$ then $e^{nH(P)} \downarrow$
- if $H(P) \uparrow$ then $e^{nH(P)} \uparrow$

To prove Theorem 5, we will use *Stirling's approximation*.

Lemma 6 (Stirling's approximation)

$$n! \approx \sqrt{2\pi n} \cdot n^n e^{-n}$$

More precisely,

$$\lim_{n \to \infty} \frac{n!}{\sqrt{2\pi n} \cdot n^n e^{-n}} = 1$$

Or even more precisely,

$$\sqrt{2\pi n} \cdot n^n e^{-n} e^{\frac{1}{12n+1}} \leq n! \leq \sqrt{2\pi n} \cdot n^n e^{-n} e^{\frac{1}{12n}}$$

Types : size

∗ Proof (Theorem 5):

$$
\begin{aligned}
|T_P| &= \binom{n}{nP(a_1), nP(a_2), \ldots, nP(a_M)} \\
&= \frac{n!}{\prod_{i=1}^{M} (nP(a_i))!} \\
&\doteq \frac{n^n e^{-n} \sqrt{2\pi n}}{\prod_{i=1}^{M} (nP(a_i))^{nP(a_i)} e^{-nP(a_i)} \sqrt{2\pi n P(a_i)}} \\
&\doteq \frac{1}{\prod_{i=1}^{M} P(a_i)^{nP(a_i)}} = \frac{1}{\prod_{i=1}^{M} e^{nP(a_i)\log P(a_i)}} = e^{nH(P)}.
\end{aligned}
$$

Theorem 7 (probability of any sequence in the type)

If the probability of the sequence under the **true** *distribution Q^n $(X_i \sim Q)$, then $Q^n(x^n)$ is all the same for all $x^n \in T_P$. and the probability is:*

$$Q^n(x^n) = e^{-n[H(P)+D(P\|Q)]}$$

✓meaning: 만약 실제 X_i가 i.i.d Q distribution을 따른다면, 동일한 type set T_P안에 들어있는 sequence들의 probability $Q^n(x^n)$은 전부 동일하다.

∗ Proof:

$$\begin{aligned} Q^n(x^n) &= \prod_{i=1}^{n} Q(x_i) = \prod_{a\in\mathcal{X}} Q(a)^{N(a|x^n)} = \prod_{a\in\mathcal{X}} Q(a)^{nP(a)} \\ &= \exp\left(\log\left[\prod_{a\in\mathcal{X}} Q(a)^{nP(a)}\right]\right) = \exp\left(n\sum_{a\in\mathcal{X}} P(a)\log Q(a)\right) \\ &= \exp\left(n\sum_{a\in\mathcal{X}} P(a)\log P(a) - (P(a)\log P(a) - P(a)\log Q(a))\right) \\ &= \boxed{e^{-n[H(P)+D(P\|Q)]}} \end{aligned}$$

Types : probability

Corollary 8 (probability of the type)

$$Q^n(T_P) \doteq e^{-nD(P||Q)}$$

$*$ Proof: (hint. using Theorem 5, 7)
$\Rightarrow$

Lemma 9 (probability of any sequence in the empirical type and true type)

Probability of any sequence $x_n, (X_i \sim Q)$ in the type with empirical distribution P and true distribution Q is satisfies the following inequality:

$$Q^n(T_Q) \geq Q^n(T_P), \qquad \forall P, Q \in \mathcal{P}_n$$

$*$ Proof: (hint. using $\frac{n!}{m!} \leq n^{n-m}$)
$\Rightarrow$

Some remarks

- 각 type은 그 정의에 의하여 sample space $\mathcal{X}^n$에 대하여 서로 disjoint하며 collectively exhaustive하므로 $\mathcal{X}^n$의 *partitions*이다. $(\bigcup_{P \in \mathcal{P}_n} T_P)$
- $\mathcal{P}_n$의 크기에 대한 bound는 loose한 bound이지만, 이 theorem으로부터 types의 개수가 sequence의 길이 n에 대해 **polynomial**하게 증가한다는 것을 알 수 있다.
- 그러나 $\mathcal{X}^n$을 이루는 각각의 disjoint set T_P는 크기가 n에 대해 exponential하게 증가하며, 증가속도는 $H(P)$에 의해 결정된다.
- 또한, 동일한 type에 속하는 sequence들의 확률은 모두 동일하며, 그 값은 n에 대해 exponential하게 감소한다. 감소속도는 empirical distribution P가 true distribution Q와 얼마나 가까운지에 따라 결정된다.
- 만약 두 확률이 거의 유사하다면 $(D(P||Q) \downarrow)$, type T_P에 있는 sample들이 발생할 확률은 n에 따라 빠르게 증가하게 된다.
- $P = Q$일 때, Corollary 8에 의하면 $Q^n(T_Q) \doteq 1$을 얻을 수 있지만, 이는 T_Q의 확률이 exponential하게 증가하거나 감소하지 않는다는 사실만을 의미할 뿐, 아무런 정보도 우리에게 주지 않는다.

$$Q^n(T_Q) \doteq 0 \doteq 1 \doteq e^0 \cdots$$

Summary

Summary of Types

Type; a new way of classifying length n sequences.

- $T_P \triangleq \left\{x^n \in \mathcal{X}^n : \hat{P}_{x^n} = P\right\}$
- $|T_P| \doteq e^{nH(P)}$
- $Q^n\left(x^n\right) = e^{-n(H(P)+D(P\|Q))}$ for $x^n \in T_P$
- $Q^n\left(T_P\right) \doteq Q^n\big(\left\{x^n : \hat{P}_{x^n} = P\right\}\big) = e^{-nD(P\|Q)}$

Method of Types

Method of types : size

이제 하나의 type이 아니라 여러가지 type들의 union set에 대하여 분석해보자.

- Type은 서로 disjoint하기 때문에 다음을 만족한다. (*assume* $H(P_1) > H(P_2)$)

$$e^{nH(P_1)} \leq |T_{P_1} \cup T_{P_2}| = |T_{P_1}| + |T_{P_2}| \doteq e^{nH(P_1)} + e^{nH(P_2)} \leq 2e^{nH(P_1)} \doteq \boxed{e^{nH(P_1)}}$$

- M개의 type으로 일반화 하면, 다음과 같다. (*assume* $H(P_1) \geq H(P_i), \forall i$)

$$e^{nH(P_1)} \leq \left| \bigcup_{i=1}^{M} T_{P_i} \right| \leq M \, |T_{P_1}| \doteq M e^{nH(P_1)} \doteq \boxed{e^{nH(P_1)}}$$

$\Rightarrow$ Theorem 5에 의하면, type의 개수는 polynomial이기 때문에, 위와 같은 approximation이 가능하다. ($M \neq e^n$)

- 더 나아가, 이를 이용하면 어떤 특별한 constraint를 만족하는 type들의 union set $\mathcal{A}$에 대한 크기도 표현할 수 있다. (assume $P^* = \arg\max_{P \in \mathcal{A}} H(P)$)

$$\left| \bigcup_{P \in \mathcal{A}} T_P \right| = \sum_{P \in \mathcal{A}} |T_P| \doteq \sum_{P \in \mathcal{A}} e^{nH(P)} \doteq \boxed{e^{nH\left(P^*\right)}}$$

Method of types : probability

Theorem 10 (Sanov's theorem)

Let Q be a distribution over $\mathcal{X}$. Let $\mathcal{A}$ be a set of distributions over $\mathcal{X}$, and suppose that $Q \notin \mathcal{A}$:

$$Q^n\left(\left\{x^n : \hat{P}_{x^n} \in \mathcal{A}\right\}\right) \doteq e^{-n \min_{P \in \mathcal{A}} D(P\|Q)}$$

$*$ Proof: hint. union type set의 크기를 구하기 위해 적용했던 과정을 그대로 확률에도 적용하면, 쉽게 증명할 수 있다.

$$e^{-nD(P^{**}\|Q)} \doteq Q^n\left(T_{P^{**}}\right) \le Q^n\left(\bigcup_{P \in \mathcal{A}} T_P\right) \le (n+1)^{|\mathcal{X}|} Q^n\left(T_{P^{**}}\right) \doteq e^{-nD(P^{**}\|Q)}$$

where

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

Some remarks & Summary

Some remarks

- 특정한 조건을 만족하는 empirical probability들의 집합 $\mathcal{A}$를 다음과 같이 가정해보자.

$$\mathcal{A} \triangleq \{P \ : \ \mathbb{E}_P[F(X)] > \alpha\}$$

  $\Rightarrow X$의 함숫값 $F(X)$들의 expectation이 threshold α보다 크도록 하는 분포
- 집합 $\mathcal{A}$를 잘 정의하면, decode F를 했을 때 정확도가 α보다 크도록 만드는, 또는 특정 error rate보다 더 작은 에러를 만들어내는 분포들의 집합을 정의할 수도 있다.
- 따라서 이런 아이디어를 바탕으로 strong typical set을 정의하고자한다.

Summary of method of types

- $|\bigcup T_P| \doteq e^{nH(P^*)}$ where $P^* = \arg\max H(P)$
- $Q^n(\bigcup T_P) \doteq e^{-nD(P^{**}\|Q)}$ where $P^{**} = \arg\min D(P\|Q)$

Strong Typicality

Strong typical set

Definition 11 (strong typical set)

For a given distribution Q with $Q(x) > 0, \forall x$, and a constant $\delta > 0$, the **strong typical set** is defined as

$$\tilde{T}_Q = \bigcup_{P \in \mathcal{A}} T_P = \bigcup_{\{P : |P(x) - Q(x)| \leq \delta, \forall x\}} T_P$$

where $\mathcal{A}$ is defined

$$A = \{P : |P(x) - Q(x)| \leq \delta, \forall x \in \mathcal{X}\}$$

✓meaning: source에 속하는 모든 symbol x에 대하여, true distribution과의 차이가 δ 보다 작은 empirical distribution들의 집합 $\mathcal{A}$에 대한 union type set.

- $P(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(x_i = x)$
- $Q(x) = \mathbb{E}_Q[\mathbf{1}(x_i = x)]$

Strong AEP

Theorem 12 (Strong AEP)

For any ϵ, $\delta > 0$, *there exists a* N_0 *s.t.* $\forall n > N_0$,

$$Q^n(\tilde{T}_Q) > 1 - \epsilon$$

more precisely,

$$Q^n(\tilde{T}_Q) \to 1.$$

✓meaning: 특정 sample이 strong typical set에 속할 확률은 1로 수렴한다.

∗ Proof: (hint. using WLLN, union bound)

- $Q^n(\tilde{T}_Q) \to 1$임을 보이는 대신, $Q^N(\tilde{T}_Q^C) \to 0$임을 보이고자한다.

$$Q^n\left(\left\{x^n : \left|\hat{P}_{x^n}(a) - Q(a)\right| > \delta \text{ for some } a \in \mathcal{X}\right\}\right) \to 0.$$

- 어떤 *특정한* $a \in \mathcal{X}$에 대하여 WLLN을 적용하면 다음을 보일 수 있다.

$$Q^n\left(\left\{x^n : \left|\hat{P}_{x^n}(a) - Q(a)\right| > \delta\right\}\right) < \frac{\epsilon}{|\mathcal{X}|}$$

Strong AEP

* Proof: (contd.)
    - Union bound를 적용하면, for some a에 대한 확률은 for any $a \in \mathcal{X}$에 대한 확률들의 합으로 bound된다.

$$
\begin{aligned}
&Q^n\left(\left\{x^n : \left|\hat{P}_{x^n}(a) - Q(a)\right| > \delta \text{ for some } a \in \mathcal{X}\right\}\right) \\
&\leq \sum_{a \in \mathcal{X}} Q^n\left(\left\{x^n : \left|\hat{P}_{x^n}(a) - Q(a)\right| > \delta\right\}\right) \\
&\leq \epsilon.
\end{aligned}
$$

    - 따라서 $Q^n(\tilde{T}_Q)$의 complement가 ϵ으로 bound되므로 다음을 얻는다.□

$$Q^n(\tilde{T}_Q) > 1 - \epsilon.$$

Comparison between weak typical set

weak typical set $A_\delta^{(n)}(Q)$

$$\left\{x_1^n \in \mathcal{X}^n : \left| -\frac{1}{n} \log Q^n(x_1^n) - H(Q) \right| \leq \delta \right\}$$

- $\forall x_1^n \in A_{\delta_n}^{(n)}(Q),\ Q^n(x^n) \doteq e^{-nH(Q)}$.
- $|A_{\delta_n}^{(n)}(Q)| \doteq e^{nH(Q)}$.
- (weak AEP) $Q^n(A_{\delta_n}^{(n)}) \to 1$ *with sufficiently large* n.

strong typical set $\tilde{T}_Q$

$$\left\{x_1^n \in \mathcal{X}^n : \left|\hat{P}_{x^n}(a) - Q(a)\right| \leq \delta, \forall a \in \mathcal{X}\right\}$$

- $\forall x_1^n \in \tilde{T}_Q,\ Q^n(x^n) \doteq e^{-nH(Q)}$.
- $|\tilde{T}_Q| \doteq e^{nH(Q)}$.
- (strong AEP) $Q^n(\tilde{T}_Q) \to 1$ *with sufficiently large* n.
- $|\hat{P}_{x^n}(a) - Q(a)| < \delta, \forall a \in \mathcal{X}$.

Summary

Summary of strong typical set

- empirical distribution과 true distribution의 값이 모든 symbol에 대해서 δ보다 멀리 떨어지지 않도록 하는 constraint를 만족하는 probability에 대한 union type set으로 strong typical set을 정의할 수 있다.
- method of types를 이용하면, strong typical set의 크기와 strong typical set에 속하는 모든 원소들의 확률이 weak typical set의 특징을 만족함을 보일 수 있다.
- 즉, strong typical set에 속하는 data만 encoding하는 방식을 차용하면, strong typical set을 이용하여 source coding theorem을 증명할 수 있다.

Universal Source Coding

Unknown distribution source

다음과 같이 unknown distribution Q를 따르는 source data ($DMS(Q)$)에 대한 optimal block source coding 방법을 찾는 것이 목적이다.

$$f_n : \mathcal{X}^n \to \{0,1\}^{nR},$$
$$g_n : \{0,1\}^{nR} \to \mathcal{X}^n$$

Definition 13 (probability of error)

The **probability of error** for the code with respect to the distribution Q is

$$P_e^{(n)} \triangleq Q^n\left(\{x^n : g_n\left(f_n\left(x^n\right)\right) \neq x^n\}\right).$$

Unknown distribution source

Definition 14 (error exponent)

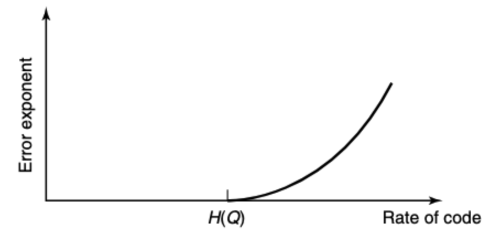
We say that an **error exponent** E is achievable at rate R if there exists a sequence of (n, R) codes with

$$P_e^{(n)} \dot{\leq} e^{-nE},$$

which means

$$\limsup_{n\to\infty} \frac{1}{n} \log P_e^{(n)} \leq -E.$$

- Unknown distribution이기 때문에 $H(P) > R$이더라도 R이 길어질수록 error probability 점차 0에 수렴할 뿐, 여전히 존재한다.
- 따라서 R이 커질 수록 error가 얼마나 빠르게 감소하는지를 표현하기 위해서 error exponent를 도입한다. (If $E \uparrow$, then error decreasing fast.)



Universal source codes

Definition 15 (universal source codes)

A rate R block code for a source will be called **universal** if the functions f_n and g_n do not depend on the distribution Q and if $P_\epsilon^{(n)} \to 0$ as $n \to \infty$ if $R > H(Q)$.

Idea for universal source codes

- source coding theorem에 의하면, $H(Q) \leq R$으로 설정하면 error가 0으로 수렴하도록 만들 수 있다.
- 우리는 true distribution을 모르기 때문에, 대신 다음을 만족하는 *empirical* probability set을 가정하자.

$$\mathcal{A} = \{P : H(P) \leq R\}$$

- $\mathcal{A}$에 대한 union type sets에 속하는 sample data들만 encoding하자!

$$T_{\mathcal{A}} \triangleq \bigcup_{P \in \mathcal{A}} T_P = \{x^n : H(\hat{P}_{x^n}) \leq R\}$$

Theorem 16 (universal source coding theorem)

There exists a sequence of (n, R) universal source codes such that $P_c^{(n)} \to 0$ for every source Q such that $H(Q) < R$. With this universal source coding, the achievable error exponent is

$$E = \min_{P:H(P)>R} D(P||Q),$$

i.e.,

$$P_e^{(n)} \dot{\leq} e^{-n \min_{P:H(P)>R} D(P||Q)}.$$

∗ Proof: (hint. Sanov's theorem)

- 아이디어에서 소개했던 union type set $T_{\mathcal{A}}$에 대해 methods of types를 적용하면, 집합의 크기를 구할 수 있다.

$$\begin{aligned}|T_{\mathcal{A}}| &= \sum_{P\in\mathcal{A}_n:H(P)\leq R} |T_P| \doteq \sum_{P\in\mathcal{A}_n:H(P)\leq R} e^{nH(P)} \\ &\leq (n+1)^{|X|} e^{nR}\end{aligned}$$

Universal source coding theorem

* <u>Proof</u>: (contd.)
  - 우리가 고안한 방법은 $T_{\mathcal{A}}$에 속하는 sample에 대해서만 encoding을 수행하기 때문에, error probability는 다음과 같이 정의된다.

$$P_e^{(n)} = 1 - Q^n(A)$$

  - Method of types를 이용하면 $Q^n(T_P)$를 다음과 같이 전개할 수 있다.$\square$

$$\begin{aligned} P_e^{(n)} &= 1 - Q^n(A) \\ &= \sum_{P:{\color{red}H(P)>R}} Q^n\left(T_P\right) \\ &\le (n+1)^{|X|} \max_{P:H(P)>R} Q^n\left(T_P\right) \\ &\le (n+1)^{|X|} e^{-n \min_{P:H(P)>R} D(P\|Q)}. \end{aligned}$$

where for P that $H(P) > R > H(Q)$?

References

- T. M. Cover and J. A. Thomas. Elements of Information Theory, Wiley, 2nd ed., 2006.
- Gallager (2008), Principles of Digital Communication, Cambridge University Press.
- Lecture notes for EE623: Information Theory (Fall 2024)