Information Theory and Statistics

태그	
생성일	@2024년 4월 8일 오전 2:45

1. Method of Types

1-1. Basic Concepts

1-1-1. Type

1. Notation
    a. $x = [x_1, x_2, ..., x_n]$ = A sequence whose length is n
    b. $N(a|x)$ = The number of occurrences of symbol a in x
    c. P_x = (probability mass function of type) = (Relative proportion of occurrences of each symbol of X) $= \frac{N(a|x)}{n}$
2. Type
    a. Empirical probability distribution of a specific sequence x as a sample
    → sequence를 sequence에 속하는 alphabet의 출현 빈도 및 sequence의 길이를 기준으로 분류하겠다는 의도
    b. Components
        i. Sample Space : Set of sequences whose length is n
        ii. Event Space : Collection of set of sequences which has same occurrences in each sequence for all symbols
            1. example: aab, baa
            2. stationary + sequence를 하나의 표본 집단으로 생각 (하나의 표본이 아니라)
        iii. Random Variable = Probability Mass Function of Type

1-1-2. Probability Simplex

1. Probability Simplex = $\{(x_1, x_2, ..., x_m)|x_1 + x_2 + ... + x_m = 1, x_i \geq 0(i = 1, 2, 3..., m)\}$

1-1-3. Set of Types with denominator n

1. Sequence의 길이가 n인 type들의 집합 (Probability Distribution들의 집합)

1-1-4. Type Class

1. Set of Types with denominator n의 원소인 type들에 대응하는 길이가 n인 sequence들의 집합

1-2. Properties

1-2-1. Size of Set of Types with denominator n

1. Notation
    a. $P_n =$ A set of types with denominator n
    b. χ = A set of alphabets
2. Theorem
    a. $|P_n| \leq (n+1)^{|\chi|}$: type set의 cardinality는 sequence의 길이 n의 polynomial이다.

1-2-2. Probability of Sequence

1. Notation
    a. $X_1, ..., X_n$ = a sequence of i.i.d random variables (alphabet을 event로 함)

b. $X_i \sim Q$ ($i = 1, ..., n$) = 각 Random Variable이 따르는 identical distribution

c. $Q^n(x)$ = $X_1, ..., X_n$의 특정 instance에 대응하는 sequence x에 대한 pmf

d. P_x = 특정 sequence x에 대한 type

2. Theorem

a. $Q^n(x) = 2^{-n(H(P_x)+D(P_x||Q))}$ (for any P)

1-2-3. Size of a Type Class ($T(P)$)

1. Notation

a. P_n = A set of types with denominator n

b. $P \in P_n$ = A type as a element of P_n

c. χ = A set of alphabets

d. $T(P)$ = A set of sequences corresponding to the given type P

2. Theorem

a. $\frac{1}{(n+1)^{|\chi|}} 2^{nH(P)} \leq T(P) \leq 2^{nH(P)}$ (for any P)

1-2-4. Probability of Type Class

1. Notation

a. P_n = A set of types with denominator n

b. $P \in P_n$ = A type as a element of P_n

c. $T(P)$ = A type class for type P = A set of sequences whose type is P

d. $Q(x)$ = Any probability distribution for alphabet χ

e. $Q^n(x)$ = Probability of a squence whose length is n and alphabet is χ

f. $Q^n(T(P))$ = Probability of a type class of sequences whose type is P (calculated as the sum of $Q^n(x)(x \in T(P))$

2. Theorem

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

2. Universal Source Coding

2-1. Basic Concepts

2-1-1. Encoder / Decoder

1. Code Rate(R)

   a. Entropy Rate = n의 값이 증가함에 따라 Entropy가 증가하는 정도

   b. n(=sequence의 길이)의 값이 증가함에 따라 Code의 개수가 증가하는 정도 (= code 개수의 sequence 길이에 대한 평균)

2. Fixed Code Rate: Code Rate의 값이 고정되어있다고 가정하는 상황

3. Encoder: n개의 symbol들로 이루어진 특정 Sequence를 특정 code로 mapping하는 함수

   a. $f_n : \chi^n \rightarrow \{1, 2, ..., 2^{nR}\}$

4. Decoder: 특정 code를 n개의 symbol들로 이루어진 특정 sequence로 mapping하는 함수

   a. $\phi_n : \{1, 2, ..., 2^n\} \rightarrow \chi^n$

5. Error of Probability: Decoding된 Sequence가 Encoding 전의 Sequence와 다르다고 할 때 즉 Error가 발생 했다고 할 때 특정 Sequence가 관측될 확률

   a. $P_e^{(n)} = Q^n(X^n|\phi_n(f_n(X^n)) \neq X^n)$

2-1-2. Universality of Code

1. Property of a rate R block code

2. Condition

   a. f_n and ϕ_n do not depend on Q, the distribution for symbol

   b. Under the condition, if $R > H(Q)$, then $\lim_{n\rightarrow\infty} P_e^{(n)} = 0$

i. $R > H(Q)$

1. 임의의 probability distribution에 대하여 sequence의 길이가 증가함에 따라 증가하는 code의 개수의 기댓값 > distribution Q에 대하여 개별 symbol에 필요한 code 개수의 기댓값
2. symbol을 encoding하기에 충분한 code의 개수가 주어짐을 통계적으로 보장하기 위한 조건
   a. sequence의 길이가 1 증가할 때 마다 사용할 수 있는 code의 개수가 증가하는 정도가
   sequence의 길이가 1 증가할 때 마다 해당 slot에 들어갈 수 있는 symbol의 개수보다 큼을 보장하기 위한 조건

2-2. Theorem: Existence of Universal Code

1. Notation
   a. n = Sequence의 길이 = Sequence에 포함된 symbol의 개수
   b. R = code rate = n(Sequence의 길이)의 증가량에 대한 code의 개수 증가량
   c. 2^{nR}= Sequence의 길이가 n이고 code rate이 R인 상황에서 총 code의 개수
   d. a sequence of ($2^{nR}, n$) universal source code = Sequence의 길이가 n이고 code rate이 R이며 총 code의 개수가 2^{nR}인 상황에서 universaility 조건을 만족하는 code들의 집합
2. 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 $H(Q) < R$
3. Proof
   a. Assumption
      i. The rate R is fixed
      ii. $R_n = R - |\chi| \frac{\log(n+1)}{n}$
      iii. $A = \{x \in \chi^n | H(P_x) \leq R_n\}$
      iv. $T(P) =$ type P의 type class = type P에 대응하는 sequence들의 집합
   b. Proof

i. Upper Bound of $|A|$

1. $|A| = \sum_{P \in P_n | H(P) \leq R_n} |T(P)|$
2. $\leq \sum_{P \in P_n | H(P) \leq R_n} 2^{nH(P)}$
3. $\leq \sum_{P \in P_n | H(P) \leq R_n} 2^{nR_n}$
4. $\leq 2^{nR_n} \sum_{P \in P_n | H(P) \leq R_n} 1 = 2^{nR_n} * (Size\ of\ the\ type\ set)$
5. $\leq 2^{nR_n}(n+1)^{|\chi|}$ (By the theorem of "size of the type set")
6. $= 2^{[|\chi| \log(n+1) + nR_n]}$
7. $= 2^{n(R_n + |\chi| \frac{\log(n+1)}{n})}$
8. $= 2^{nR_n}$ (By definition of R_n)

ii. Encoding / Decoding Scheme

1. A의 원소의 개수보다 code의 수가 많으므로 indexing이 가능함
2. indexing 방식으로 encoding function($f_n(x)$)을 구성 (code를 index로 mapping)
   a. $f_n(x) = \begin{cases} index\ of\ x\ in\ A\ if\ x \in X \\ 0 \end{cases}$
3. decoding function (index를 code로 mapping)

iii. Universality of Encoding/Decoding Scheme

1. Assumption
   a. $X_1, X_2, ..., X_n$ = i.i.d Random Variables
   b. Q = Probability Distribution that each of the random variables $X_1, X_2, ..., X_n$ depends on
   c. $H(Q) < R$
2. Proof
   a. Upper Bound of Error Probability
      i. $P_\epsilon^{(n)} = 1 - Q^{(n)}(A)$
      ii. $= \sum_{P|H(P) > R_n} Q^{(n)}(T(P))$
      iii. $\leq \sum_{P|H(P) > R_n} \max_{P|H(P) > R_n} Q^{(n)}(T(P))$

iv. $\leq \max_{P|H(P)>R_n} Q^{(n)}(T(P)) \sum_{P|H(P)>R_n} 1$

v. $\leq \max_{P|H(P)>R_n} Q^{(n)}(T(P))(n+1)^{|\chi|}$

vi. $\leq (n+1)^{|\chi|}(2^{-n \min_{P|H(P)>R_n} D(P||Q)})$

b. $H(P) > R_n > H(Q)$ (for some n)

i. R_n은 R보다 작은 상태에서 R에 가까워짐 ($n \to \infty,\ R_n \to R$)

ii. $H(Q) < R$

iii. $H(Q) < R_n < R$을 만족하게 하는 임계값으로서 n_0가 존재

iv. $\min_{P|H(P)>R_n}$에 의해 $H(P) > R_n$이므로 $H(P) > R_n > H(Q)$

3. Large Deviation Theory

3-1. Sanov's Theorem

1. Notation

a. E = A subset of the set of probability mass functions

b. P^n = Set of types with denominator n

c. $E \cap P^n$ = Subset of the set of types with denominator n

d. $Q^n(E)$ = 길이가 n인 Sequence에 대한 set of types에 대한 부분 집합의 확률

i. $Q^n(E) = Q^n(E \cap P_n) = \sum_{x|P_x \in E \cap P_n} Q^n(x)$

e. $X_1, X_2, ..., X_n$ = a sequence of i.i.d random variables

f. $Q(x)$ = A distribution that each of random variables $X_1, ..., X_n$ depends on

g. $P^* = \arg\min_{P \in E} D(P||Q)$ = Distribution in E that is closest to Q in relative entropy

2. Theorem

a. $Q^n(E) = Q^n(E \cap P_n) \leq (n+1)^{|\chi|} 2^{-nD(P^*||Q)}$

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

3. Proof

a. Upper Bound of $Q^n(E)$

i. $Q^n(E) = \sum_{P \in E \cap P_n} Q^n(T(P))$

ii. $\leq \sum_{P \in E \cap P_n} 2^{-nD(P||Q)}$

iii. $\leq \sum_{P \in E \cap P_n} \max_{P \in E \cap P_n} 2^{-nD(P||Q)}$

iv. $= \sum_{P \in E \cap P_n} 2^{-n \min_{P \in E \cap P_n} D(P||Q)}$

v. $\leq \sum_{P \in E \cap P_n} 2^{-n \min_{P \in E} D(P||Q)}$

vi. $= \sum_{P \in E \cap P_n} 2^{-nD(P^*||Q)}$

vii. $\leq (n+1)^{|\chi|} 2^{-nD(P^*||Q)}$

b. Lower Bound of $Q^n(E)$

i. Additional Assumption for E

1. For all large n, there is a distribution in $E \cap P_n$ that is close to P^*

2. E is the closure of its interior → Thus, the interior must be non-empty for alla $n \geq n_0$

a. Limit Point

i. Notation

1. (X, T, d) = Metric Space

2. E = X의 subset

3. $N_r(p)$ = 기준 거리 r에 대한 기준점 p의 Neighborhood= $\{p' | p' \in X, d(p, p') \leq r\}$

4. $N_r'(p) = N_r(p) \setminus \{p\}$

ii. Definition: 다음의 조건을 만족하는 기준점 $p \in X$를 E의 limit point라고 함

1. $\forall r > 0$, $N_r'(p) \cap E \neq \emptyset$ = 임의의 양의 기준 거리 r에 대한 기준점 p의 스스로를 제외한 Neighborhood가 전체 집합

X와의 교집합을 가진다.

b. Interior Point

i. Notation

1. (X, T, d) = Metric Space
2. E = X의 subset
3. $N_r(p)$ = Point p의 기준 거리 r에 대한 Neighborhood = $\{p'|p' \in X, d(p,p') \leq r\}$

ii. Definition: 다음의 조건을 만족하는 기준점 $p \in X$를 E의 interior point라고 함

1. $\exists r > 0$, $N_r(p) \subseteq E$ = 기준 거리 r에 대한 기준점 p의 neighborhood가 X의 부분집합이다.

c. Exterior Point

i. Notation

1. (X, T, d) = Metric Space
2. $E \subseteq X$ = X의 부분 집합
3. $X \backslash E$ = X 중에서 E를 제외한 부분
4. $N_r(p)$ = 기준 거리 r에 대한 기준점 p의 neighborhood = $\{p'|p' \in X, d(p',p) \leq r\}$

ii. Definition: 다음의 조건을 만족하는 기준점 $p \in X$를 E의 exterior point라고 함 = 다음의 조건을 만족하는 기준점 $p \in X$를 E^c의 interior point라고 함

1. $\exists r > 0$, $N_r(p) \subseteq X \backslash E$ = 기준 거리 r에 대한 기준점 p의 neighborhood가 $X \backslash E$의 부분집합

d. open set

i. (X, T, d) = Metric Space

ii. $E \subseteq X$ = E는 X의 부분 집합

iii. Definition: 다음의 조건을 만족하는 집합 E를 open set이라고 함

1. E의 모든 원소로서의 점이 E의 interior point임

e. closed set

i. (X, T, d) = Metric Space

ii. $E \subseteq X$ = E는 X의 부분집합

iii. Definition: 다음의 조건을 만족하는 집합 E를 closed set이라고 함

1. E의 모든 limit point가 E의 원소임

f. closure

i. 모든 limit point들의 집합

g. interior

i. 모든 interior point들의 집합

ii. There are distributions P_n such that $P_n \in E \cap P_n$ and $D(P_n||Q) \to D(P^*||Q)$

iii. For each $n \geq n_0$, the followings are true

1. $Q^n(E) = \sum_{P \in E \cap P_n} Q^n(T(P))$
2. $\geq Q^n(T(P_n))$
3. $\geq \frac{1}{(n+1)^{|\chi|}} 2^{-nD(P_n||Q)}$

c. Convergence

i. $\liminf \frac{1}{n} \log Q^n(E) \geq \liminf(-\frac{|\chi| \log(n+1)}{n} - D(P_n||Q)) = -D(P^*||Q)$

4. Conditional Limit Theorem

4-1. Pythagorean Theorem

1. Notation

a. E = Closed convex set

b. P = Probability distribution의 집합

c. $E \subset P$ = P의 부분 집합

2. Condition
   a. $P^* \in E$ = Distribution that achieves the minimum distance to $Q \equiv D(P^*||Q) = \min_{P \in E} D(P||Q)$
   b. $Q \notin E$
3. Theorem
   a. $D(P||Q) \geq D(P||P^*) + D(P^*||Q)$ ($\forall P$)
4. Usecase
   a. Suppose that we have a sequence $P_n \in E$ that yields $D(P_n||Q) \rightarrow D(P^*||Q)$
   b. Then, $D(P_n||P^*) \rightarrow 0$ (P_n이 최적화되고 있다.)

4-2. L_1 distance

1. Notation
   a. P_1, P_2 = Probability Distributions
   b. χ = Set of symbols
   c. a = specific symbol
2. Definition: L_1 distance
   a. $||P_1 - P_2|| = \sum_{a \in \chi} |P_1(a) - P_2(a)|$
3. Lemma 11.6.1 : Lower bound of Relative Entropy with L_1 distance
   a. $D(P_1||P_2) \geq \frac{1}{2 \ln 2}(||P_1 - P_2||_1)^2$

4-3. Conditional Limit Theorem

1. Notation
   a. P= set of types with denominator n
   b. E = closed convex set as a subset of P

c. $X_1, X_2, ..., X_n$ ~ Q = i.i.d discrete random variables

d. $p^* = \min_{p \in E} D(p||Q)$

e. a = specific symbol

2. Theorem

a. $n \to \infty$, $P(X_1 = a | P_{X^n} \in E) \to p^*(a)$

b. The conditional distribution of X_1 is close to p^* for large n

3. Proof of Theorem

a. Preliminary

i. $S_t = \{p \in P | D(P||Q) \leq t\}$ (기준 거리가 t인 Pseudo Neighborhood 이자 subset of set of types)

1. $D(P||Q)$ is a convex function. Therefore the set S_t is convex

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

1. $D(P||Q)$ is strictly convex in P. Therefore P^* is unique

iii. $A = S_{D^*+2\delta} \cap E$ =

iv. $B = E - S_{D^*+2\delta} \cap E = E - A$

v. $A \cup B = E$

vi. $Q^{(n)}(B) =$ Subset of set of types with denominator n로서 B의 각 type p에 대응하는 type class $T(p)$들의 합집합에 대한 probability

b. Upper Bound of $Q^n(B)$

i. $Q^n(B) = \sum_{p \in E \cap P_n | D(P||Q) > D^*+2\delta} Q^n(T(p))$

ii. $\leq \sum_{p \in E \cap P_n | D(p||Q) > D^*+2\delta} 2^{-nD(p||Q)}$

iii. $\leq \sum_{p \in E \cap P_n | D(p||Q) > D^*+2\delta} 2^{-n(D^*+2\delta)}$

iv. $= 2^{-n(D^*+2\delta)} \sum_{p \in E \cap P_n | D(p||Q) > D^*+2\delta} 1$

v. $\leq 2^{-n(D^*+2\delta)} \sum_{p \in P_n} 1$

vi. $= 2^{-n(D^*+2\delta)} (n+1)^{|\mathcal{X}|}$

c. Lower Bound of $Q^n(A)$

i. $Q^n(A) \geq Q^n(S_{D^*+\delta} \cap E)$

ii. $= \sum_{p \in E \cap P_n | D(p||Q)} \frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-nD(P||Q)}$

iii. $\geq \sum_{p \in E \cap P_n | D(p||Q)} \frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-n(D^*+\delta)}$

iv. $\geq \frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-n(D^*+\delta)}$ (for sufficiently large n)

1. For sufficiently large n, $S_{D^*+\delta} \cap E \cap P_n \neq \emptyset$ (n을 sample의 크기로 봐야할 듯)

d. Upper Bound of $P(p_{X^n} \in B | p_{X^n} \in E)$

i. p_{X^n} = Sequence $X^n = (X_1, X_2, ..., X_n)$에 대한 type

ii. $P(p_{X^n} \in B | p_{X^n} \in E) = \frac{Q^n(B \cap E)}{Q^n(E)}$

iii. $\leq \frac{Q^{(n)}(B)}{Q^n(A)}$

iv. $\leq \frac{(n+1)^{|\mathcal{X}|} 2^{-n(D^*+2\delta)}}{\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-n(D^*+\delta)}}$

v. $= (n+1)^{2|\mathcal{X}|} 2^{-n\delta}$

e. Upper bound of $P(p_{X^n} \in B | p_{X^n} \in E)$ implies the followings

i. $n \to \infty$, $P(p_{X^n} \in B | p_{X^n} \in E) \to 0$

ii. $n \to \infty$, $P(p_{X^n} \in A | p_{X^n} \in E) \to 1$

f. All members of A are close to p^* in relative entropy

i. By definition

1. $D(p||Q) \leq D^* + 2\delta$

ii. By upper inequality and Pythagorean theorem,

1. $D(p||P^*) + D(p^*||Q) \leq D(p||Q) \leq D^* + 2\delta$

2. $D(p||P^*) + D^* \leq D(p||Q) \leq D^* + 2\delta$,

3. $D(p||P^*) \leq 2\delta$ (by definition, $D* = D(p^*||Q)$)

iii. $p_X \in A$ implies that $D(p_x||Q) \leq D^* + 2\delta$. Therefore, $D(p_x||P) \leq 2\delta$

g. $P(p_{X^n} \in A | p_{X^n} \in E) \to 1$. Therefore, $P(D(p_{X^n}||p^*) \leq 2\delta | p_{X^n} \in E) \to 1$ as $n \to \infty$

h. By Lemma 11.6.1 : Lower bound of Relative Entropy with L_1 distance

i. "The relative entropy is small" implies that L_1 distance is small

ii. "L_1 distance is small" implies that "$\max_{a \in \chi} |P_{X^n}(a) - P^*(a)|$ is small" (by sum $\geq$ max for non-negative values)

iii. Thus, $P(|P_{X^n}(a) - P^*(a)| \geq \epsilon | P_{X^n} \in E) \to 0$ as $n \to \infty$

   1. L_1 distance is small $\equiv |P_{X^n}(a) - P^*(a)| < \epsilon$

iv. Alternatively, this can be written as follows

   1. $P(X_i = a | P_{X^n} \in E) \to P^*(a)$ in probability, $a \in \chi$
   2. Since $X_1, ..., X_n$ is i.i.d, $P(X_1 = a | P_{X^n} \in E) \to P^*(a)$ in probability, $a \in \chi$

5. Hypothesis Testing

5-1. Hypothesis Testing

1. Intuitive definition: Decision problem between alternative explanations(hypothesis) for the data observed
2. Formal definition

   a. Notation

      i. X_1, X_2, ..., $X_n \sim Q(x)$ = I.I.D Random Variables

      ii. Probability Distributions

         1. Q = The unknown probability distribution that $X_1, ..., X_n$ depends on in real
         2. P_1, P_2 = The known probability distributions that $X_1, ..., X_n$ may depend on

      iii. Hypotheses

1. $H_1 : Q = P_1$ = The first hypothesis that $X_1, X_2, ..., X_n$ depends on the probability distribution P_1
2. $H_2 : Q = P_2$ = The second hypothesis that $X_1, X_2, ..., X_n$ depends on the probability distribution P_2

iv. Decision function

1. $g(x_1, x_2, .., x_n)$
    a. $g(x_1, x_2, ..., x_n) = 1$ = The first hypothesis H_1 is accepted
    b. $g(x_1, x_2, ..., x_n) = 2$ = The second hypothesis H_2 is accepted
2. Decision Region: Inverse image of g = Acceptance Region
    a. $A = \{(x_1, x_2, ..., x_n) | g(x_1, x_2, ..., x_n) = 1\} = g^{-1}(1)$
    b. $A^c = \{(x_1, x_2, ..., x_n) | g(x_1, x_2, ..., x_n) = 2\} = g^{-1}(2)$

v. Probabilities of Error → Recall / Precision과도 관련 있음

1. $\alpha = P(g(X_1, X_2, ..., X_n) = 2 | H_1\ is\ true) = P_1^n(A^c)$ (Type 1 Error -1 $- Precision$)
2. $\beta = P(g(X_1, X_2, ..., X_n) = 1 | H_2\ is\ true) = P_2^n(A)$ (Type 2 Error - $Recall$)

5-2. Neyman-Pearson lemma

1. Notation
    a. $X_1, ..., X_n \sim Q$ = I.I.D Random Variables
    b. Probability distributions
        i. Q = The unknown probability distribution that $X_1, ..., X_n$ depends on
        ii. P_1, P_2 = The known probability distributions that $X_1, ..., X_n$ may depend on
    c. Hypothesis
        i. $H_1 : Q = P_1$

ii. $H_2 : Q = P_2$

d. Decision function

i. $g(x_1, x_2, ..., x_n)$

1. $g(x_1, x_2, ..., x_n) = 1 \;\; (if \; \frac{P_1(x_1,x_2,...,x_n)}{P_2(x_1,x_2,...,x_n)} > T$)

2. $g(x_1, x_2, ..., x_n) = 2 \;\; (if \; \frac{P_1(x_1,x_2,...,x_n)}{P_2(x_1,x_2,...,x_n)} \leq T$)

ii. Decision Region = Acceptance Region

1. $A_n(T) = \{x^n | \frac{P_1(x_1,x_2,...,x_n)}{P_2(x_1,x_2,...,x_n)} > T\} = g^{-1}(1)$

2. $A_n^c(T) = \{x^n | \frac{P_1(x_1,x_2,...,x_n)}{P_2(x_1,x_2,...,x_n)} \leq T\} = g^{-1}(2)$

3. B_n = Any decision region/acceptance region other than A_n

e. Probabilities of Error

i. $\alpha^* = P(g(X_1, X_2, ..., X_n) = 2 | H_1 \; is \; true) = P_1^n(A_n^c(T))$

ii. $\beta^* = P(g(X_1, X_2, ..., X_n) = 1 | H_2 \; is \; true) = P_2^n(A_n(T))$

iii. $\alpha = P_1^n(B_n^c)$

iv. $\beta = P_2^n(B_n)$

2. Theorem

a. If $\alpha \leq \alpha^*$, then $\beta \geq \beta^*$ = Hypothesis 1의 Error Probability가 감소하면 Hypothesis 2의 Error Probability가 증가한다.

3. Proof

a. Notation

i. Acceptance Region

1. $A = A_n(T) = \{x^n | \frac{P(x_1,x_2,...,x_n)}{P_2(x_1,x_2,..,x_n)} > T\} \subseteq \chi^n$

2. $B \subseteq \chi^n$ = Any decision/acceptance region other than A_n

ii. Indicator Function

1. $\phi_A(x) = c_A * 1(x \in A)$(Indicator fucntion for Acceptance Region A)

2. $\phi_B(x) = c_B * 1(x \in B)$(Indicator function for Acceptance Region B)

iii. $x = (x_1, x_2, ..., x_n) \in \chi^n$

b. Base Inequality

i. $(\phi_A(x) - \phi_B(x))(P_1(x) - TP_2(x)) \geq 0$

ii. $[\phi_A(x) * P_1(x) - T * \phi_A(x) * P_2(x) - \phi_B(x) * P_1(x) + T\phi_B(x)P_2(x)] \geq 0$

c. $\sum_x [\phi_A(x) * P_1(x) - T * \phi_A(x) * P_2(x) - \phi_B(x) * P_1(x) + T\phi_B(x)P_2(x)] \geq 0$

i. $\sum_{x \in A} [\phi_A(x) * P_1(x) - T * \phi_A(x) * P_2(x)] - \sum_{x \in B} [\phi_B(x) * P_1(x) - T * \phi_B(x) * P_2(x)] \geq 0$

ii. $= \sum_{x \in A} [P_1(x) - T * P_2(x)] - \sum_{x \in B} [P_1(x) - T * P_2(x)]$

iii. $= \sum_{x \in A} [P_1(x)] - T * \sum_{x \in A} [P_2(x)] - \sum_{x \in B} [P_1(x)] + T * \sum_{x \in B} [P_2(x)]$

iv. $= P_1^n(A) - T * P_2^n(A) - P_1^n(B) + T * P_2^n(B)$

v. $= (1 - P_1^n(A^c)) - T * P_2^n(A) - (1 - P_1^n(B)) + T * (P_2^n(B))$

vi. $= (1 - \alpha^*) - T * \alpha^* - (1 - \alpha) + T * (\beta)$

vii. $= (\alpha - \alpha^*) + T(\beta - \beta^*)$

viii. $= T(\beta - \beta^*) - (\alpha^* - \alpha) \geq 0$ $(T \geq 0)$

d. $T(\beta - \beta^*) \geq (\alpha^* - \alpha)$ $(T \geq 0)$. Therefore, if $\alpha^* \geq \alpha$, then $\beta \geq \beta^*$

4. [Neyman-Pearson Lemma] means that the following test form (Likelihood Ratio Test) for two hypotheses is best

a. $\frac{P_1(X_1, X_2, ..., X_n)}{P_2(X_1, X_2, ..., X_n)} > T$

6. Chernoff-Stein Lemma

6-1. AEP for Relative Entropy

1. Notation

a. $X_1, X_2, ..., X_n \sim P_1(x)$: I.I.D random variables that depends on $P_1(x)$

b. $P_2(x)$: Any probability distribution other than $P_1(x)$ on sample space χ

2. Theorem

a. $\frac{1}{n} \log \frac{P_1(X_1, X_2, ..., X_n)}{P_2(X_1, X_2, ..., X_n)} \to D(P_1||P_2)$ in probability

6-2. Relative Entropy Typical Sequence

1. Notation

a. n = Fixed number of occurrences of symbols in a sequence

b. χ^n = Set of sequence with length n

c. $(x_1, x_2, ..., x_n) \in \chi^n$ = A sequence of symbols in χ

d. $A_\epsilon^{(n)}(P_1||P_2)$ = Relative Entropy Typical Set = A set of relative entropy typical sequences

2. Definition

a. Relative Entropy Typical Set($A_\epsilon^{(n)}(P_1||P_2)$) is the set which meets the following condition

i. $D(P_1||P_2) - \epsilon \leq \frac{1}{n} \log \frac{P_1(x_1, x_2, ..., x_n)}{P_2(x_1, x_2, ..., x_n)} \leq D(P_1||P_2) + \epsilon$

6-3. Basic Properties of Relative Entropy Typical Sequences

1. Notation

a. n = Fixed number of occurrences of symbols in a sequence

b. $(x_1, x_2, ..., x_n) \in \chi^n$ = A sequence of symbols in χ

c. $P_1(x_1, x_2, ..., x_n)$ = Probability #1 of a sequence of $x_1, x_2, ..., x_n$

d. $P_2(x_1, x_2, ..., x_n)$ = Probability #2 of a sequence of $x_1, x_2, ..., x_n$

e. $A_\epsilon^{(n)}(P_1||P_2)$ = A set of relative entropy typical sequences whose length is n

f. $P_1(A_\epsilon^{(n)}(P_1||P_2))$ = Probability of a relative entropy typical set = Sum of elements of relative entropy typical sequences whose length is n

2. Properties
    a. Bounds of P_2 for a relative entropy typical sequence
        i. $P_1(x_1, x_2, ..., x_n) * 2^{-n[D(P_1||P_2)+\epsilon]} \leq P_2(x_1, x_2, ..., x_n) \leq P_2(x_1, x_2, ..., x_n) * 2^{-n[D(P_1||P_2)-\epsilon]}$
    b. Convergence of P_1
        i. $P_1(A_\epsilon^{(n)}(P_1||P_2)) > 1 - \epsilon$
        ii. for sufficiently large n
    c. Bounds of P_2 for a relative entropy typical set
        i. Upper Bound
            1. $P_2(A_\epsilon^{(n)}(P_1||P_2)) < 2^{-n[D(P_1||P_2)-\epsilon]}$
            2. for sufficiently large n
        ii. Lower Bound
            1. $P_2(A_\epsilon^{(n)}(P_1||P_2)) > (1-\epsilon)2^{-n[D(P_1||P_2)+\epsilon]}$
            2. for sufficiently large n
3. Lemma 11.8.1: Lower Bound of P_2 distribution for a set of general sequence
    a. Notation
        i. χ = A set of symbols
        ii. χ^n = A set of sequences which consist of n symbols
        iii. $x_1, x_2, ..., x_n \in \chi^n$ = A sequence whose length is n
        iv. $B_n \subset \chi^n$ = A subset of sequences which consist of n symbols
        v. $P_1(B_n)$ = A probability distribution for a subset of sequences which consist of n symbols
        vi. $P_2(B_n)$ = A probability distributiion other than P_1 for a subset of sequences which consts of n symbols such that $D(P_1||P_2) < \infty$
    b. Theorem

i. $P_2(B_n) > (1-2\epsilon)2^{-n(D(P_1||P_2)+\epsilon)}$

6-4. Chernoff-Stein Lemma

1. Notation
    a. Sequence
        i. $X_1, X_2, ..., X_n \sim Q$ = I.I.D random variables that depends on Q
    b. Probability Distributions
        i. Q = The unknonw probability distribution that $X_1, ..., X_n$ depend on in real
        ii. P_1, P_2 = The known probability distributions that $X_1, ..., X_n$ may depend on such that $D(P_1||P_2) < \infty$
    c. Hypothesis
        i. $H_1 : Q = P_1$, $H_2 : Q = P_2$
    d. Decision Function
        i. $g(x_1, x_2, ..., x_n)$
            1. $g(x_1, x_2, ..., x_n) = 1$ (if H_1 is true)
            2. $g(x_1, x_2, ..., x_n) = 2$ (if H_1 is false)
    e. Acceptance Region = Inverse Image of g
        i. $A_n = \{(x_1, x_2, ..., x_n) | g(x_1, x_2, ..., x_n) = 1\} = g^{-1}(1)$
        ii. $A_n^c = \{(x_1, x_2, ..., x_n) | g(x_1, x_2, ..., x_n) \neq 1\} = g^{-1}(2)$
    f. Probabilities of Error
        i. $\alpha_n = P_1^n(A_n^c)$ = Type 1 Error Probability
        ii. $\beta_n = P_2^n(A_n)$ = Type 2 Error Probability
        iii. $\beta_n^\epsilon = \min_{A_n \subseteq \chi^n, \alpha_n < \epsilon} \beta_n$ ($0 < \epsilon < \frac{1}{2}$)
2. Theorem
    a. $\lim_{n \to \infty} \frac{1}{n} \log \beta_n^\epsilon = -D(P_1||P_2)$
3. Proof

a. [Part 1]: Exhibit a sequence of sets such that $\lim_{D(P_1||P_2)\to 0} \beta_n = 0 \equiv \lim_{D(P_1||P_2)\to 0} \log \beta_n = 0$ (lineary for $\log \beta_n$, exponentially for β_n)

   i. Notation

      1. A_n = A selected sequence of sets

      2. $A_n^\epsilon(P_1||P_2)$ = Relative entropy typical set from P_1 to P_2

   ii. Proof

      1. By assumption

         a. $A_n = A_\epsilon^{(n)}(P_1||P_2)$

      2. By the convergence property of P_1

         a. Convergence Property

            i. $P_1(A_\epsilon^{(n)}(P_1||P_2)) > 1 - \epsilon$ (for n sufficiently large)

         b. Convergence Property for complement

            i. $P_1(A_\epsilon^{(n)}(P_1||P_2)^c) < \epsilon$ (for n sufficiently large)

            ii. $\alpha_n < \epsilon$

      3. By the upper bound property of P_2

         a. $P_2(A_\epsilon^{(n)}(P_1||P_2)) < 2^{-n[D(P_1||P_2)-\epsilon]}$ (for sufficiently large n)

         b. $\log P_2(A_\epsilon^{(n)}(P_1||P_2)) < -n[D(P_1||P_2) - \epsilon]$ (for sufficiently large n)

         c. $\frac{1}{n} \log P_2(A_\epsilon^{(n)}(P_1||P_2)) < -[D(P_1||P_2) - \epsilon]$ (for sufficiently large n)

         d. $\frac{1}{n} \log P_2(A_\epsilon^{(n)}(P_1||P_2)) < -D(P_1||P_2) + \epsilon$ (for sufficiently large n)

         e. $\lim_{n\to\infty} \frac{1}{n} \log \beta_n = -D(P_1||P_2)$

      4. Therefore, if $A_n = A_\epsilon^{(n)}(P_1||P_2)$, $\lim_{n\to\infty} \frac{1}{n} \log \beta_n = -D(P_1||P_2)$ with $\alpha_n < \epsilon$

b. [Part 2]: No sequence of sets including A_n can have a lower exponent in type 2 error probability (β_n) than $-D(P_1||P_2)$

   i. Notation

1. $B_n = \{(x_1, ..., x_n) \in \chi^n | P_1(B_n) > 1 - \epsilon\}$ ($\epsilon > 0$)

ii. Proof

1. By lemma 11.8.1 (lower bound for P_2 distribution for a set of general sequences)
   a. $P_2(B_n) > (1 - 2\epsilon)2^{-n[D(P_1||P_2)+\epsilon]}$ ($\epsilon < \frac{1}{2}$)
   b. $\log P_2(B_n) > \log(1 - 2\epsilon) - n[D(P_1||P_2) + \epsilon]$
   c. $\frac{1}{n} \log P_2(B_n) > \frac{1}{n} \log(1 - 2\epsilon) - [D(P_1||P_2) + \epsilon]$
2. Lower Bound for $\lim_{n\to\infty} \frac{1}{n} \log P_2(B_n)$
   a. $\lim_{n\to\infty} \frac{1}{n} \log P_2(B_n) > \lim_{n\to\infty} \frac{1}{n} \log(1 - 2\epsilon) - [D(P_1||P_2) + \epsilon]$
   b. $\lim_{n\to\infty} \frac{1}{n} \log P_2(B_n) > -[D(P_1||P_2) + \epsilon]$

c. [Part 2]에 따라 $\frac{1}{n} \log P_2(B_n)$에 대하여 $-D(P_1||P_2)$ 보다 더 작은 극한 값은 존재하지 않음.

d. $\log$는 monontic increasing function이므로 따라서 이에 대응하는 β_n의 값들도 동일하며 이 β_n들이 최솟값이 됨