Information Theory

📖 Cover, Thomas M. *Elements of information theory.* John Wiley & Sons, 1999.

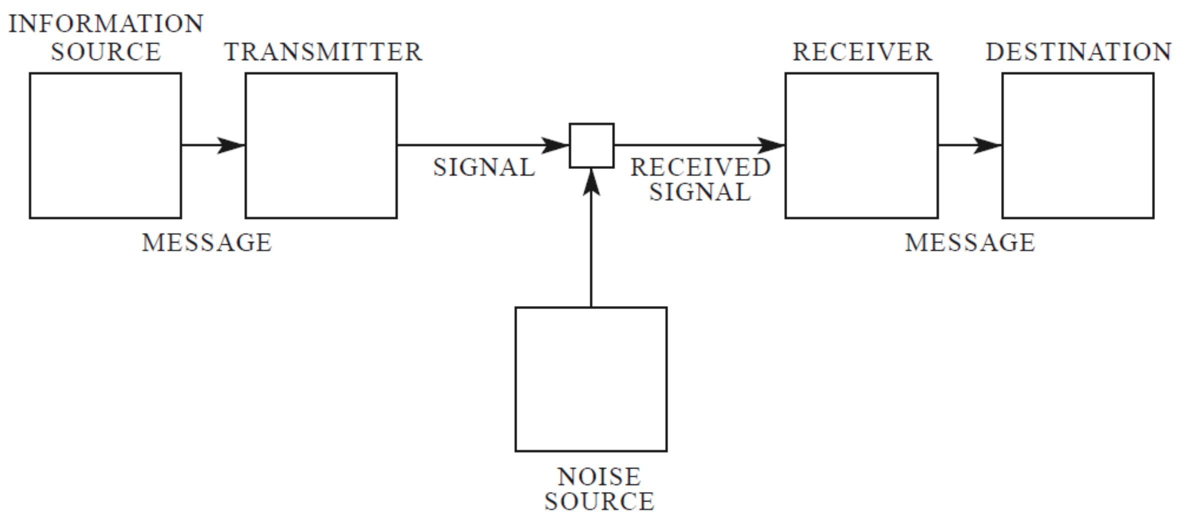


Fig. 1—Schematic diagram of a general communication system.

이상적인 channel은 transmitter로부터 receiver까지 오류 없이 정보가 전달되는 channel이지만, 일반적인 channel은 noise channel임

Probability

Event : 일어날 수 있는 어떤 사건, 확률 변수가 가질 수 있는 값, $X = a$

- ▼ 확률 변수가 주어짐 ⇔ 확률 변수를 앎 ⇔ 확률 분포를 앎

  ex. 동전을 던졌을 때 나올 수 있는 면에 대한 확률 변수 X

    확률 변수 X∈{앞면, 뒷면}

    확률 분포는 P(X=앞면)=1/2, P(X=뒷면)=1/2

결과로 도출될 수 있는 모든 event를 모은 집합이 sample space이고, '어떤 event가 얼마나 나올 수 있는가?'에 대한 확률이 주어졌을 때 event에 대한 변수가 확률 변수이고 이 확률 변수의 events에 대한 확률이 확률 분포를 이룸

⇒ 확률 변수를 다루는 것은 어떤 event의 확률을 다루는 것이 아니라 변수의 확률 분포를 다루는 것

확률 질량 함수(probability mass function) : 이산 확률 변수의 확률 분포에 따른 함수

- ▼ 확률 변수 X를 사용한 함수 $f(X)$도 변수에 대한 확률 분포를 갖는 확률 변수임

  확률 변수

  사건 공간에서 가측 공간으로의 가측 함수

  ex. $f(X) = \begin{cases} 1 & \text{if } X = \text{Head} \\ 0 & \text{if } X = \text{Tail} \end{cases}$ 이때, X는 event를 나타내는 변수

확률 변수 $f(X)$의 평균 $\mathbb{E}[f(X)] := \sum P(X = x) f(X = x)$

Joint distribution

P(X,Y) : joint distribution

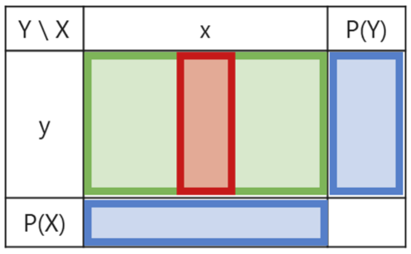
P(X), P(Y) : marginal distribution

→ joint distribution이 주어지면
marginal distribution을 구할 수 있으나
역은 성립하지 않음

P(Y|X=x) : conditional distribution

Y는 R.V.이고, 조건부의 x는 value

X=2인 행에 대한 확률은
X=2일 때의 Y에 대한 조건부 분포를 나타냄



- ▼ ex

Y \ X	0	1	2	P(Y)
0	1/4	1/8	1/8	1/2
1	1/4	0	1/4	1/2
P(X)	1/2	1/8	3/8	

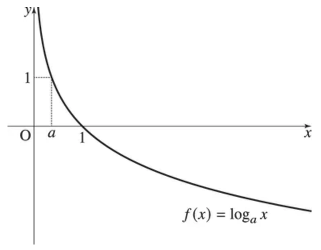
Information and Uncertainty

- **Information : level of surprise**
  - → 낮은 확률 = 높은 entropy
    얼마나 적은 확률을 가지고 있으며, 알게 되었을 때 얼마나 놀라움을 주는가
  - X: random variable, $X \in \mathcal{X} = \{x_1, \dots, x_n\}$
- The information of an event $X = x_i$ is defiend by $I(x_i) := \log \frac{1}{P(X=x_i)} = -\log p(x_i)$
  정보는 $P(X = x_i)$에만 depend한다. 즉, $I(x_i) = I(p(x_i)) = -\log p(x_i)$

💡 왜 $\log$를 사용하여 정보를 나타내는가?

정보는 다음을 만족함

1. 확률이 낮으면 정보가 많고, 확률이 높으면 정보가 적음
2. $0 \le P(X = x_i) \le 1$
   $I(x_i) \ge 0 \quad \because 0 < p(x) \le 1$
   (확률이 0인 event는 $\mathcal{X}$에 속하지 않음)
3. Fundamental Axioms (axiomatic apprach)
   $I(p)$: information measure
   a. $I(p) \ge 0$
   b. $p = 1 \quad \Rightarrow \quad I(p) = 0$
   c. For tow independent events $P(X = x_i), P(Y = y_i)$,
      $I(X = x_i, Y = y_j) = I(X = x_i) + I(Y = y_j)$
      because, $P(X = x_i,\ Y = y_i) = P(X = x_i)(Y = y_i)$
4. The information measure $I(p)$ is continuous



log의 밑을 2로 사용하면 정보량의 비트 수를 표현할 수 있음

- 분포의 information = $\mathbb{E}[I(X)] = \sum_x P(X = x) I(X = x) = -\sum_x p(x) \log p(x)$
  "확률 변수를 앎 ⇔ 확률 분포를 앎"이기 때문에 R.V.의 information을 아는 것은 분포의 information을 아는 것과 같음
  분포에 대한 information은 각각의 $X = x_i$에 대한 information을 물어보는 것이 아니기 때문에, 모든 x에 대한 information의 기댓값을 구하여 분포의 information을 정의함

Mutual information

- For two R.V. $X \leftarrow \{x_1, x_2, \dots, x_m\}$, $Y \leftarrow \{y_1, y_2, \dots, y_m\}$,
  (case 1) X, Y are independent
    $Y = y_j$의 발생이 $X = x_i$에 대한
    어떠한 정보도 제공하지 않음
  (case 2) X, Y are fully dependent
    $Y = y_j$의 발생이 $X = x_i$의 발생을 결정
    → 통신 과정에서 Y=y를 전송 받으면
      실제로 보내진 값 X가 x였음을 보장할 수 있음

(case 2)의 그림

Y \ X	x	P(Y)
y	0 ... 0 1 0 ... 0	
P(X)		

- $I(X;Y)$: mutual information (average of mutual information $X = x_i, Y = y_j$) — for R.V.s
  $I(X;Y) = \sum_{x,y} p(x,y) I(x,y)$
  I(x,y)는 X=x,Y=y일 때의 상호 정보량을 나타냄
  일반적으로 joint information을 정의한 기호가 없기 때문에
  고정된 x_i, y_j에 대한 mutual information 표기를 ;이 아닌 ,로 사용함
- $I(X;Y) := \sum_{x,y} p(x,y)[I(x) - I(x|y)] = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$
  (위 식의 첫 등식에 나오는 $p(x,y)$는 $x = x_i, y = y_j$에 대한 $p(x_i, y_j)$임.
  $I(x_i; y_j) = I(x_i) - I(x_i|y_j))$ — for event
  Y=y임을 알게 됨으로써 줄어드는 x에 대하여 모르는 정도(정보량)
  상호 정보량이 높으면 Y를 알 때 X를 알 확률이 높아짐
  $I(x) - I(x|y) = \log \frac{1}{p(x)} - \log \frac{1}{p(x|y)} = \log \frac{p(x|y)}{p(x)} = \log \frac{p(x,y)}{p(x)p(y)}$
  $I(X;Y) = I(Y;X)$
- $I(X;Y) = H(X) - H(X|Y)$

pf) $I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$
$= \sum_{x,y} p(x,y) \log \frac{1}{p(x)} - \sum_{x,y} p(x,y) \log \frac{p(y)}{p(x,y)}$
$= \sum_{x} p(x) \log \frac{1}{p(x)} - \sum_{x,y} p(x,y) \log \frac{1}{p(x|y)}$
$= H(X) - H(X|Y)$

네 번째 등호 넘어갈 때 아래 참고

$H(X|Y) = \sum_{y} p(y) H(X|Y=y)$
$= \sum_{y} p(y) \sum_{x} p(x|y) \log \frac{1}{p(x|y)}$
$= \sum_{x,y} p(x|y)p(y) \log \frac{1}{p(x|y)} = \sum_{x,y} p(x,y) \log \frac{1}{p(x|y)}$

▼ 변형

$H(X) = H(X|Y) + I(X;Y)$; X에 대한 불확실성은 Y를 알 때 X에 대하여 모르는 정도와 Y를 알 때 X를 아는 정도의 합

좋은 채널 = 상호 정보량이 높은 채널

▼ Ex 1

1. independent : $p(x|y) = p(x) \Rightarrow I(x;y) = 0$
2. fully dependent : $p(x|y) = 1 \Rightarrow I(x;y) = I(x)$

▼ Ex 2 (*binary symmetric channel*)

p : error rate

(case 1)

X	0	1
P(X)	1	0

→

Y	0	1
P(Y)	1-p	p

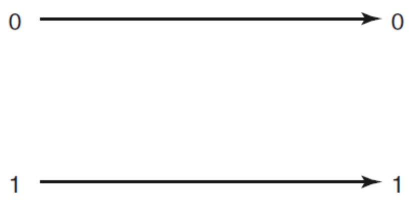


FIGURE 1.3. Noiseless binary channel. $C = 1$ bit.

(case 2)

X	0	1
P(X)	1/2	1/2

→

Y	0	1
P(Y)	1/2	1/2

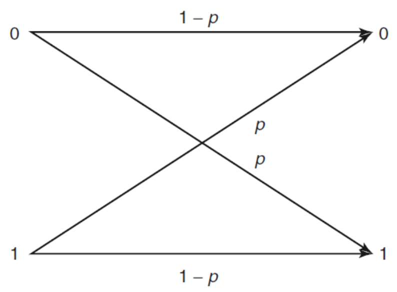


FIGURE 1.5. Binary symmetric channel.

$P(Y=0) = P(Y=0|X=0)P(X=0) + P(Y=0|X=1)P(X=1)$
$= (1-p)(1/2) + p(1/2) = 1/2$

$P(Y=1) = P(Y=1|X=0)P(X=0) + P(Y=1|X=1)P(X=1)$
$= p(1/2) + (1-p)(1/2) = 1/2$

$I(X=0;Y=0) = \log \frac{p(x|y)}{p(x)} = \log \frac{p(y|x)}{p(y)} = \log \frac{1-p}{1/2} = \log 2(1-p) = \log(1-p)$

$I(X=0;Y=1) = \log \frac{p(x|y)}{p(x)} = \log \frac{p(y|x)}{p(y)} = \log \frac{p}{1/2} = \log 2p = \log p$

→ $p \to 0$일 때 $\log p \to -\infty$; 개별 상호 정보량은 음수가 나오는 경우도 있으나, 평균을 구하면 0보다 큰 값으로 보정됨

Conditional Mutual Information

$I(X;Y \mid Z) = H(X|Z) - H(X \mid Y, Z)$

▼ **Theorem (Chain rule for mutual information)**

$I(X_1, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y \mid X_{i-1}, \ldots, X_1)$

pf) $I(X_1, \ldots, X_n; Y)$
$= H(X_1, \ldots, X_n) - H(X_1, \ldots, X_n \mid Y)$
$= \sum_{i=1}^{n} H(X_i \mid X_{i-1}, \ldots, X_1) - \sum_{i=1}^{n} H(X_i \mid X_{i-1}, \ldots, X_1, Y)$
$= \sum_{i=1}^{n} [H(X_i \mid X_{i-1}, \ldots, X_1) - H(X_i \mid Y, X_{i-1}, \ldots, X_1)]$
$= \sum_{i=1}^{n} H(X_i; Y \mid X_{i-1}, \ldots, X_1)$

Entropy

The Shannon entropy of R.V. $X \sim p(x)$ is defined by

$$H(X) := -\sum_{x} p(x) \log p(x) \qquad ; \log \frac{1}{p(x)} \text{의 기대값(평균)}$$

▼ 확률 변수의 (확률 분포의) 평균 information

$H(X) := \mathbb{E}[I(X)]$

- 어떤 분포를 따르는 확률 변수를 표현할 수 있는 최소 비트 수
- 데이터를 최대로 압축 때의 크기 (바른 복원을 보장할 수 있어야 함)

▼ Example 1.1.2 (*x의 확률이 uniform하지 않은 경우 code word의 길이가 줄어드는 상황*)

Example 1.1.2 Suppose that we have a horse race with eight horses taking part. Assume that the probabilities of winning for the eight horses are $\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right)$. We can calculate the entropy of the horse race as

$$H(X) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{8}\log\frac{1}{8} - \frac{1}{16}\log\frac{1}{16} - 4\frac{1}{64}\log\frac{1}{64} = 2 \text{ bits.} \quad (1.3)$$

x_i	1	2	3	4	5	6	7	8
$p(x_i)$	1/2	1/4	1/8	1/16	1/64	1/64	1/64	1/64
code	0	10	110	1110	111100	111101	111110	111111

평균 코드 길이 = entropy ; $\mathbb{E}[\text{code len}] = \mathbb{E}[f(X)]$

▼ Example 1.1.3

X	0	1
$P(X)$	p	$1-p$

$H(X) = H(p)$

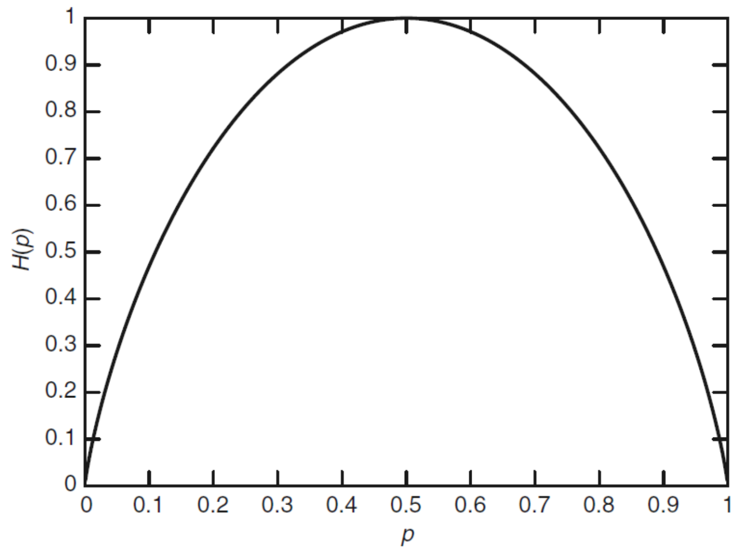


FIGURE 2.1. $H(p)$ vs. p.

Joint entropy

R.V. $Z := (X,Y) \sim p(z) = p(x,y) := P[X = x \text{ and } Y = y]$

$H(Z) = H(X,Y) = \sum_x \sum_y p(x,y) \log \frac{1}{p(x,y)} = \mathbb{E}_{x,y}[-\log(X,Y)]$

Conditional entropy

$H(Y|X)$; $X = x_i$의 x_i가 변화할 때 Y의 엔트로피

💡 Shannon entropy : 정보량의 평균

$H(Y|X) := \sum P(X = x_i) H(Y|X = x_i)$

Mutual entropy

▼ $H(X;Y) := \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$

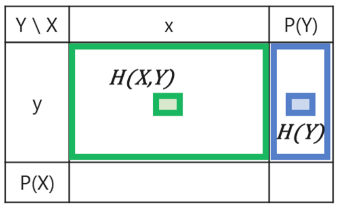
Y=y임을 알게 됨으로써 줄어드는 x에 대하여 모르는 정도(정보량)

$I(x) - I(x|y) = \log \frac{1}{p(x)} - \log \frac{1}{p(x|y)} = \log \frac{p(x|y)}{p(x)} = \log \frac{p(x,y)}{p(x)p(y)}$

Theorem (Chain Rule)

$H(X,Y) = H(Y) + H(X|Y)$

joint dist (X,Y)의 불확실성
= Y에 대한 불확실성 + Y를 알 때 X에 대한 불확실성



▼ 증명

$$\begin{aligned}
&H(X,Y) \\
&= -\sum_{x,y} \log p(x,y) \\
&= -\sum_{x,y} p(x,y) \log p(x)p(y|x) \\
&= -\sum_{x,y} p(x,y) \log p(x) - \sum_{x,y} p(x,y) \log p(y|x) \\
&= -\sum_x p(x) \log p(x) - \sum_{x,y} p(x,y) \log p(y|x) \\
&= H(X) + H(Y|X)
\end{aligned}$$

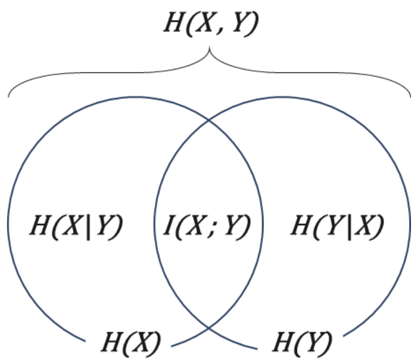
- **Corollary (Diagram about Entropy & M.I.)**

  $H(X,Y \mid Z) = H(X|Z) + H(X \mid Y, Z)$

  $I(X;Y) = H(X) - H(X|Y)$
  $= I(Y;X) = H(Y) - H(Y|X)$

  $H(X,Y) = H(X) + H(Y|X)$
  $= H(Y,X) = H(Y) + H(X|Y)$

  $I(X;X) = H(X)$



Chain rule for entropy

$X_1, X_2, \ldots, X_n$ are R.V.s $\sim p(x_1, \ldots, x_n)$

$H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|Xi-1, \ldots, X_1)$

- pf 1

  $H(X_1, X_2) = H(X_1) + H(X_2|X_1)$

  $H(X_1, X_2, X_3) = H(X_1) + H(X_2, X_3|X_1)$
  $= H(X_1) + H(X_2|X_1) + H(X_3|X_2, X_1)$ (by corollary)

  ...

  $H(X_1, \ldots, X_n)$
  $= H(X_1) + H(X_2, \ldots, X_n|X_1)$
  $= H(X_1) + H(X_2|X_1) + H(X_3|X_2, X_1) + \cdots + H(X_n|X_{n-1}, \ldots, X_1)$
  $= \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1)$

- pf 2

  $p(x_1, x_2) = p(x_1)p(x_2|x_1)$

  $p(x_1, x_2, x_3) = p(x_1, x_2)p(x_3|x_1, x_2) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)$

  ...

  $p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p(x_i|x_{i-1}, \ldots, x_1)$

  Then,

  $$\begin{aligned} H(X_1, \ldots, X_n) &= -\sum_{x_1, \ldots, x_n} p(x_1, \ldots, x_n) \log p(x_1, \ldots, x_n) \\ &= -\sum_{x_1, \ldots, x_n} p(x_1, \ldots, x_n) \log \prod_{i=1}^{n} p(x_i|x_{i-1}, \ldots, x_1) \\ &= -\sum_{x_1, \ldots, x_n} p(x_1, \ldots, x_n) \sum_{i=1}^{n} \log p(x_i|x_{i-1}, \ldots, x_1) \\ &= -\sum_{i=1}^{n} \sum_{x_1, \ldots, x_n} p(x_1, \ldots, x_n) \log p(x_i|x_{i-1}, \ldots, x_1) \\ &= -\sum_{i=1}^{n} \sum_{x_1, \ldots, x_i} p(x_1, \ldots, x_i) \log p(x_i|x_{i-1}, \ldots, x_1) \\ &= -\sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1) \end{aligned}$$

Relative entropy

- $D(p \parallel q) = \sum_i p(x_i) \log \frac{p(x_i)}{q(x_i)}$ — p와 q의 차이

  $\neq D(q \parallel p)$ (in general)

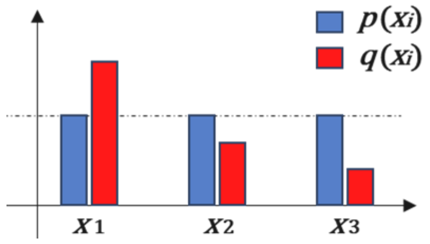
  - A measure of the distance between two distributions p and q
  - p가 중심이 되어 바라볼 때 q의 분포가 얼마나 가까운가

  R.V. distribution

  - $X \sim p \ \wedge \ \widetilde{X} \sim q$
  - $\chi = \{x_1, \ldots, x_n\}$ (the same sample space)
    ⇒ 서로 다른 sample space에서 정의된 변수는
    비교 대상이 될 수 없음

  $q(x_i) := P(\widetilde{X} = x_i)$

  $p(x_i) := P(X = x_i)$



- $\sum p(x) \log \frac{1}{p(x)} + \sum p(x) \log \frac{p(x)}{q(x)} = \sum p(x) \log \frac{1}{q(x)}$

  p : true distribution (unknown) $\wedge$ q : approximation (guessing) of p

  If we can construct a “good code” for $X \sim p$, then the average cod length = $H(p)$

Instead, we use (of mis-use) a code for $X \sim q$,
then the average code length = $\sum p(x) \log \frac{1}{q(x)}$ > $H(p)$

⇒ We need $H(p) + D(p \parallel q)$ bits on the average.

현실에서 code word를 부여하여 코딩할 때 필요한 평균 비트 수는
최소 요구 비트 수 + 잘못 부여함에 의한 추가 비트 수

▼ ex. Code word

x_i	1	2	3	4	5	6	7	8
$p(x_i)$	1/2	1/4	1/8	1/16	1/64	1/64	1/64	1/64
code	0	10	110	1110	111100	111101	111110	111111

sequence of R.V.s: $X_1, X_2, \ldots, X_n$ where $X_i \sim P(X)$

⇒ 변수의 절반은 1, 나머지 절반은 2, ... 같은 방식으로 값이 나올 것을 예상할 수 있음

주어진 상황에서 $H(X) = -\sum p(x) \log p(x)$이고, 엔트로피는 event에 대한 information의 평균이며 동시에 최적화된 (minimized) code length의 평균임

결국 information이자 code length인 $\log \frac{1}{p(x)}$가 변수를 뽑을 때마다 계산됨
⇒ $X_1, \ldots, X_n$ 중 절반은 code length가 1, 나머지 절반은 2, ... 가 됨
$H(X) = 2$이므로 X_1부터 X_n까지 표현하기 위한 비트는 평균적으로 $2 \cdot n$ bits가 필요함

R.V.의 sequence를 가지고 있으면 이 sequence를 어느 정도의 길이로 표현할 수 있는가는 분포에 의해 정해지고, 최적화된 code length인 엔트로피를 가지고 있는 것이 Shannon's entropy 값임

▼ 설명

known $X_1, X_2, \ldots, X_n \sim p(x) \Rightarrow \log \frac{1}{p(x)}$; code length $\Rightarrow \sum p(x) \log 1 p(x)$

확률 분포를 알면 최적화된 code word의 길이를 구할 수 있음

반대로 fixed $p(x)$를 모르는 경우, $q(x)$로써 code word의 길이를 guessing하는 과정이 필요함

$q(x) \Rightarrow \log \frac{1}{q(x)}$

⇒ 최적화 code word 길이 = 실제 나올 확률 × 부여한 코드 길이 ⇒ $\sum q(x) \log \frac{1}{q(x)}$

$p(x)$를 모르는 채 $X_1, X_2, \ldots, X_n$을 뽑아서 average length를 계산할 수 있음

실제로 구해서 계산하게 되는 값은 $\sum p(x) \log \frac{1}{p(x)}$이므로 두 값의 차이를 알 수 있음

$D(p \parallel q) := \sum p(x) \log \frac{1}{q(x)} - \sum p(x) \log \frac{1}{p(x)} = \sum p(x) \log \frac{p(x)}{q(x)}$

▼ Mutual information과의 관계

Mutual inforamtion

$I(x_i\,; y_j) = I(x_i) - I(x_i|y_j) = \log \frac{1}{p(x_i)} - \log \frac{1}{p(x_i|y_i)} = \log \frac{p(x_i, y_j)}{p(x_i)p(y_j)}$

$$\begin{aligned} I(X;Y) &= \sum p(x_i, y_j) I(x_i\,; y_j) \\ &= \sum p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)p(y_j)} = D(p(x,y) \parallel p(x)p(y)) \text{ — joint || product} \end{aligned}$$

$\therefore I(X;Y) = 0 \Leftrightarrow p(x,y) = p(x)p(y)$ for $\forall x, y$ — independent

⇒ joint 분포와 product의 분포의 차이가 0

💡 $0 \log \frac{0}{0} \to 0 \quad 0 \log \frac{0}{q} \to 0 \quad p \log \frac{p}{0} \to \infty$

define $p(x), q(x)$ for $\exists x \in \chi$ s.t. $p(x) > 0, q(X) = 0$
then $D(p \parallel q) = \infty$

▼ ex. $D(p \parallel q) \neq D(q \parallel q)$

$\chi = \{0, 1\}$

x	0	1
p(x)	1-r	r
q(x)	1-s	s

$$\begin{aligned} H(p) &= -\sum p(x) \log p(x) \\ &= -(1-r)\log(1-r) \\ &\quad -r \log r \end{aligned}$$

$H(p) \leftarrow$

$H(q) \leftarrow$

$H(q) = -(1-s)\log(1-s) = s \log s$

For $r = \frac{1}{2}, s = \frac{1}{4}$,

$D(p \parallel q) = \sum p(X) \log \frac{p(x)}{q(x)} = \frac{1}{2} \log \frac{1/2}{3/4} + \frac{1}{2} \log \frac{1/2}{1/4} \approx 0.2075$

and $D(q \parallel q) = \frac{3}{4} \log \frac{3/4}{1/2} + \frac{1}{4} \log \frac{1/4}{1/2} \approx 0.1887$

Jensen's inequality

convex 또는 concave를 판단하기 위해서는 조건을 만족하도록 구간을 설정해 주어야 함

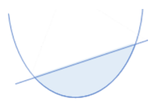
Def: Convex

$f:(a,b)\to\mathbb{R}$ is convex if for every $x_1,x_2\in(a,b)$,

$f(\lambda x_1+(1-\lambda)x_2)\le\lambda f(x_1)+(1-\lambda)f(x_2)$ where $0\le\forall\lambda\le 1$

(구간 내 x=a일 때, 곡선 위 point ≤ 구간 내 x=a일 때, 직선 위의 point)

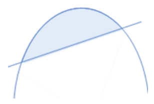
NOTE: $\lambda x_1+(1-\lambda)x_2\Rightarrow x_1$,$x_2$를 $\lambda:1-\lambda$로 내분하는 점



convex

Def: Concave

$f:(a,b)\to\mathbb{R}$ is concave if $-f$ is convex



concave

▼ **Theorem**

$f\in C^2(f',f''$ exist and are continuous) and $f''(x)\ge 0$ on (a,b), then f is convex

pf: $f(x)=f(x_0)+f'(x_0)(x-x_0)+(f''(x^*)/2)(x-x_0)^2,\quad x^*\in[x_0,x]$

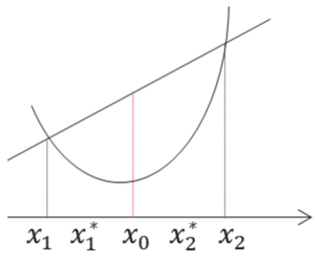
$(f''(x^*)/2)(x-x_0)^2$이 $y=ax^2+bx+c, a>0$ 꼴로 convex

x_1,x_2 : arbitrarily given points

Let $x_0:=\lambda x_1+(1-\lambda)x_2\ \ (0\le\lambda\le 1)$

$$f(x_1)=f(x_0)+f'(x_0)(x_1-x_0)$$
$$+(f''(x_1^*)/2)(x_1-x_0)^2$$

(x_1^* lines between x_0,x_1)



$$f(x_2)=f(x_0)+f'(x_0)(x_2-x_0)+(f''(x_2^*)/2)(x_2-x_0)^2$$

(x_2^* lines between x_0,x_1)

Since $f''(x_1^*)\ge 0, f''(x_2^*)\ge 0$,

1. $f(x_1)\ge f(x_0)+f'(x_0)(x_1-x_0)$
2. $f(x_2)\ge f(x_0)+f'(x_0)(x_2-x_0)$

$$\begin{aligned}
&\lambda f(x_1)+(1-\lambda)f(x_2)\\
&\quad\ge\lambda f(x_0)+\lambda f'(x_0)(x_1-x_0)+(1-\lambda)f(x_2)+(1-\lambda)f'(x_0)(x_2-x_0)\\
&\quad=f(x_0)f'(x_0)[\lambda x_1-\lambda x_0+(1-\lambda)x_2-(1-\lambda)x_0]\\
&\quad=f(x_0)+f'(x_0)[\lambda x_1+(1-\lambda)x_2-x_0]\\
&\quad=f(x_0)+f'(x_0)[x_0-x_0]=f(x_0)\\
&\quad=f(\lambda x_1+(1-\lambda)x_2)
\end{aligned}$$

$\therefore f$ is convex

▼ **Thm: Jensen's inequality**

f : convex and X : R.V. then $\mathbb{E}[f(X)]\ge f(\mathbb{E}[X])$

pf: We use mathematical induction.

Let $\chi_2=\{x_1,x_2\},\dots,\chi_k=\{x_1,x_2,\dots,x_n\}$ where $|\chi_i|=i$ for all $2\le i\le k$

case 1) $\chi=\{x_1,x_2\}$

x	x_1	x_2
$p(x)$	p_1	p_2

$$\begin{aligned}
\mathbb{E}[f(X)]&=p_1f(x_1)+p_2f(x_2)\\
&=p_1f(x_1)+(1-p_1)f(x_2)\\
&\ge f(p_1x_1+(1-p_1)x_2)\text{ since } f \text{ is convex}\\
&=f(\mathbb{E}[X])
\end{aligned}$$

case 2) Suppose that Theorem is true for $|\chi|\le k-1$

$X\sim\{\ \chi:=\{x_1,\dots,x_k\}$ and $p_i:=P[X=x_i]\ \ (i=1,\dots,k)\ \}$ are given.

Define a distribution X' with $\chi_{k-1}:=\{x_1,\dots,x_{k-1}\}$
where $p_i':=P(X=x_i)=\frac{p_i}{1-p_k}\ \ (i=1,\dots,k-1)$

$$\begin{aligned}
\mathbb{E}[f(X)]&=\sum p_if(x_i)=p_kf(x_k)+\sum_{i=1}^{k-1}p_if(x_i)\\
&=p_kf(x_k)+(1-p_k)\sum_{i=1}^{k-1}\frac{p_i}{1-p_k}f(x_i)\\
&=p_kf(x_k)+(1-p_k)\sum_{i=1}^{k-1}p_i'f(x_i)\\
&=p_kf(x_k)+(1-p_k)\mathbb{E}[f(X')]\\
&\ge p_kf(x_k)+(1-p_k)f\left(\sum_{i=1}^{k-1}p_i'x_i\right)\text{ for } |\chi|\le k-1\\
&\ge f\left(p_kx_k+(1-p_k)\sum_{i=1}^{k-1}p_i'x_i\right)\text{ by } f \text{ : convex}
\end{aligned}$$

$= f\left(p_k x_k + (1-p_k)\sum_{i=1}^{k-1} \frac{p_i}{1-p_k} x_i\right)$
$= f\left(\sum_{i=1}^{k} p_i x_i\right)$
$= f(\mathbb{E}[X])$

▼ Theorem: Information inequality

$D(p \parallel q) \geq 0 \quad , \quad I(X;Y) \geq 0 \quad , \quad H(X) \leq \log|\chi|$

$p(x), q(x)$ are p.m.f. (probability mass function) on $x \in \chi$

1. $D(p \parallel q) \geq 0$ (equality holds $\iff$ $p(x) \equiv q(x)$)
   - ▼ pf

     $supp(f)$: support of f

     $supp(f) := \overline{\{x|f(x)>0\}} = \{x|f(x)>0\}$ where x is discrete

     Let $A := \{x|p(x)>0\}$ (i.e. $A := supp(p)$)

     $-D(p \parallel q) = -\sum_{x\in\chi} p(x)\log\frac{p(x)}{q(x)} = -\sum_{x\in A} p(x)\log\frac{p(x)}{q(x)}$
     $= \sum_{x\in a} p(x)\log\frac{q(x)}{p(x)} \quad \cdots \quad \mathbb{E}\left[\log\frac{q(X)}{p(X)}\right]$
     $\leq \log\left[\sum_{x\in A} p(x)\frac{q(x)}{p(x)}\right]$ by Jensen's inequality
     $= \log\left[\sum_{x\in A} q(x)\right]$
     $\leq \log\left[\sum_{x\in\chi} q(x)\right] = \log 1 = 0$

     $\therefore \ D(p \parallel q) \geq 0$
2. $I(X;Y) \geq 0$ (equality holds $\iff$ X, Y are independent
   - ▼ pf

     $I(X;Y) = D\left(p(x,y) \parallel p(x)q(y)\right) \geq 0$
3. X: R.V. $x \in \chi$, $H(X) \leq \log|\chi|$ (entropy는 distribution이 uniform일 때 최대)
   - ▼ pf

     Let $u(x) := \frac{1}{|\chi|}$ (constant function on χ)

     $u(x)$ is uniform p.m.f. over χ.

     Then for any R.V. $X \sim P$

     $D(p \parallel q) = \sum p(x)\log\frac{p(x)}{u(x)}$
     $= \sum p(x)\log|\chi| + \sum p(x)\log p(x)$
     $= \log|\chi| - \sum p(x)\log\frac{1}{p(x)}$ since $\log|\chi|$ is constant and $\sum p(x) = 1$
     $= \log|\chi| - H(X) \geq 0$

     $\therefore H(X) \leq \log|\chi|$

     ⇒ entropy of X is smaller than Entropy of uniform disribution

▼ Theorem: Conditioning reduces entropy

$H(X|Y) \leq H(X)$

pf: $0 \leq I(X;Y) = H(X) - H(X|Y)$

$\therefore \ H(X) \geq H(X|Y)$

ex. $H(X) = H\left(\frac{1}{8}, \frac{7}{8}\right)$ or

$= H(p)$ where $p = \frac{1}{8}, q = 1-p$

$H(X|Y=2) = H\left(\frac{1}{2}, \frac{1}{2}\right) = 1$

$\therefore H(X) \ll H(X|Y=2)$

⇒ 정리에 모순인 것처럼 보이지만,

정리는 average에 대한 내용이고

예제는 single term $Y=2$라는

특정 상황에 대한 결과임

$Y \setminus X$	1	2	$P(Y)$
1	0	3/4	3/4
2	1/8	1/8	1/4
$P(X)$	1/8	7/8	

▼ Theorem: independence bound

$H(X_1,\ldots,X_n) \leq \sum_{i=1}^{n} H(X_i)$ (equality ⇔ independence)

pf: By the Chain rule,

$H(X_1,\ldots,X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1},\ldots,X_1)$
$\leq \sum_{i=1}^{n} H(X_i)$ by previous theorems

Log-sum inequality and its Applications

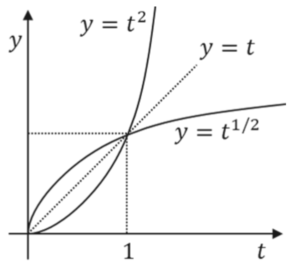
Theorem: Log-sum inequality

▼ For non-negative numbers $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$,

$\sum_{i=1}^n a_i \left(\log \frac{a_i}{b_i}\right) \geq \left(\sum_{i=1} n a_i\right) \cdot \log\left(\sum_{i=1}^n a_i / \sum_{i=1}^n b_i\right)$ (equality ⇔ $\frac{a_i}{b_i}$ is constant)

pf: case 1) For $a_i \geq 0, b_i > 0$.

Let $f(t) := t \log t$, then $f(t)$ is convex.



> 💡 $t \times a > t \quad \text{as} \quad t \to \infty \quad \text{where} \quad a > 1$

Since $(\log t)' = \left(\frac{\log_e t}{\log_e 2}\right)' = \frac{1}{\log_e 2} \cdot \frac{1}{t}$,

$f'(t) = \log t + \log e$

$f''(t) = \frac{1}{\log_e 2} \cdot \frac{1}{t} = \frac{\log e}{t} > \frac{1}{t} > 0$

$\therefore\ f''(t) > 0 \ \Rightarrow\ f$ is convex

By Jensen's inequality $\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$ for convex f,

$\sum \alpha_i f(t_i) \geq f\left(\sum \alpha_i t_i\right)$

for $X \leftarrow t_i, P_i \leftarrow \alpha_i$ where $\sum \alpha_i = \sum p_i = 1, \alpha_i \geq 0, p_i \geq 0$

Set $\alpha_i = \frac{b_i}{\sum_j b_j}$ because we need to property $\sum_i \alpha_i = \sum_i \frac{b_i}{\sum_j b_j} = \frac{1}{\sum_j b_j}\left(\sum_i b_i\right) = 1$.

set $t_i = \frac{a_i}{b_i}$

Then we obtain $\sum_i \alpha_i f(t_i) = \sum_i \frac{b_i}{\sum_j b_j} f\left(\frac{a_i}{b_i}\right) = \sum_i \frac{b_i}{\sum_j b_j} \frac{a_i}{b_i} \log \frac{a_i}{b_i}$

$= \sum_i \frac{a_i}{\sum_j b_j} \log \frac{a_i}{b_i} = \frac{1}{\sum_j b_j} \sum_i a_i \log \frac{a_i}{b_i}$

$\geq f\left(\sum_i \alpha_i t_i\right)$ by Jensen's inequality

$= \left[\sum_i \left(\frac{b_i}{\sum_j b_j}\right) \frac{a_i}{b_i}\right] \log \left[\sum_i \left(\frac{b_i}{\sum_j b_j}\right) \frac{a_i}{b_i}\right]$

$= \frac{1}{\sum_j b_j} \sum_i a_i \log \left(\frac{1}{\sum_j b_j} \sum_i a_i\right)$

In this inequality, $\frac{1}{\sum_j b_j} \sum_i a_i \log \frac{a_i}{b_i} \geq \frac{1}{\sum_j b_j} \sum_i a_i \log \left(\frac{1}{\sum_j b_j} \sum_i a_i\right)$

$\therefore\ \sum_i a_i \log \frac{a_i}{b_i} \geq \sum_i a_i \log \left(\frac{\sum_i a_i}{\sum_i b_i}\right)$

Applying Log-sum inequality

▼ $D(p \parallel q) \geq 0$ where p, q are p.m.f.

Relatibe entropy $D(p \parallel q) := \sum_x p(x) \log \frac{p(x)}{q(x)} \geq \sum_x p(x) \log \frac{\sum_x p(x)}{\sum_x q(x)} = \sum_x p(x) \log 1 = 0$

$\therefore\ D(p \parallel q) \geq 0$ (equality ⇔ $\frac{p(x)}{q(x)} = 1$ for all x)

▼ **Thm: Convexity of relative entropy**

$D(\lambda p_1 + (1-\lambda)p_2 \parallel \lambda q_1 + (1-\lambda)q_2) \leq \lambda D(p_1 \parallel q_1) + (1-\lambda) D(p_2 \parallel q_2)$

$D(p \parallel q)$ is convex in the pair (p, q) where p, q are p.m.f.
i.e. If $(p_1, q_1), (p_2, q_2)$ are two pairs of p.m.f. and $0 \leq \lambda \leq 1$,
then $D(\lambda p_1 + (1-\lambda)p_2 \parallel \lambda q_1 + (1-\lambda)q_2) \leq \lambda D(p_1 \parallel q_1) + (1-\lambda) D(p_2 \parallel q_2)$

pf: First, we check $\lambda p_1 + (1-\lambda)p_2$ and $\lambda p_1 + (1-\lambda)p_2$ are p.m.f.

It's trivial!

Next, $D(\lambda p_1 + (1-\lambda)p_2 \parallel \lambda q_1 + (1-\lambda)q_2)$

$:= \sum_x \left[\lambda p_q(x) + (1-\lambda)p_2(x)\right] \log \frac{\lambda p_1 + (1-\lambda)p_2}{\lambda q_1 + (1-\lambda)q_2}$

$\leq \sum_x \left[\sum_i \{\lambda p_1 + (1-\lambda)p_2\} \log \frac{\lambda p_1 + (1-\lambda)p_2}{\lambda q_1 + (1-\lambda)q_2}\right]$

$= \sum_x \left[\lambda p_1(x) \log \frac{\lambda p_1(x)}{\lambda q_1(x)} + (1-\lambda)p_2(x) \log \frac{(1-\lambda)p_2(x)}{(1-\lambda)q_2(x)}\right]$

$= \lambda \sum_x p_1(x) \log \frac{p_1(x)}{q_1(x)} + (1-\lambda) \sum_x p_2(x) \log \frac{p_2(x)}{q_2(x)}$

$= \lambda D(p_1 \parallel q_1) + (1-\lambda) D(p_2 \parallel q_2)$

즉, 두 p.m.f.를 combination한 뒤 relative entropy는 각각의 relative entropy의 평균보다 작거나 같음

▼ **Thm. Concavity of entropy**

$H(p)$ is a concave function of p.m.f. p

pf 1: Let $p(x)$ is p.m.f., $x \in \chi = \{x_1, \dots, x_n\}$ with $|\chi| = n$

and $u(x) := \frac{1}{|\chi|}$ is uniform distribution

$H(p) = \log|\chi| - D(p \parallel u)$

💡 $D(p \parallel u) = \sum p(x) \log \frac{p(x)}{u(x)} = \sum p(x) \log \frac{1}{u(x)} - \sum p(x) \log \frac{1}{p(x)}$

Since $\log \frac{1}{u(x)} = \log|\chi|$ is constant, $D(p \parallel u) = \log|\chi| - H(p)$.

Thus $D(p \parallel u)$ is convex.

$\Rightarrow -D(p \parallel u)$ is concave

$H(p) = \log|\chi| - D(p \parallel u)$

$\therefore\ H(p)$ is concave

pf 2:

Channel Capacity

- 한번에 (일정 시간 내에) 최대로 전송할 수 있는 데이터

💡 Communication channel
output이 확률적으로 input에 의존하는 시스템

▼ Ex (*noisy four-symbol channel*)

error rate $p = 1/2$

(case 1) $P(X = x)$가 uniform한 경우

X \ Y	1	2	3	4	P(X)
1	1/8	1/8	0	0	1/4
2	0	1/8	1/8	0	1/4
3	0	0	1/8	1/8	1/4
4	1/8	0	0	1/8	1/4
P(Y)	1/4	1/4	1/4	1/4	

Y를 받았을 때 X를 하나로 결정할 수 없음

→ decoding 실패. $I(X;Y) \neq 1$

noisy n-symbol channel

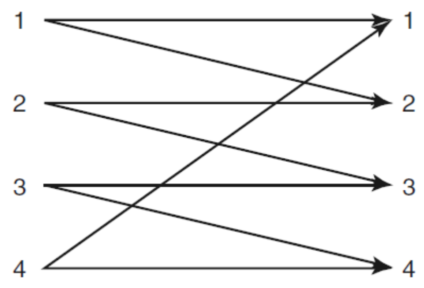


FIGURE 1.4. Noisy channel.

(case 2) $P(X = x)$가 uniform하지 않은 경우

X \ Y	1	2	3	4	P(X)
1	1/4	1/4	0	0	1/2
2	0	0	0	0	0
3	0	0	1/4	1/4	1/2
4	0	0	0	0	0
P(Y)	1/4	1/4	1/4	1/4	

Y를 받으면 X를 결정할 수 있음

→ decoding 성공. $I(X;Y) = 1$

4개의 정보를 전할 수 있는
2비트 채널을 가지고 있지만,

실제로는 1과 3 두 정보만
주고받는 1비트 채널로 사용함

→ 채널 용량 $C = 1$ bit

더 좋은 방법이 있다면 C 수정

▼ $C := \max I(X;Y)$

The capacity is the maximum tate at which we can send information over the channel

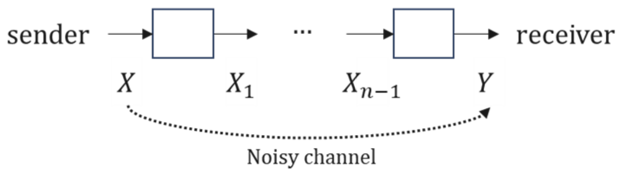
채널의 기본적인 설정(입력 가능한 값들의 범위, error rate 등)은 코딩을 통해 바꿀 수 없음

코딩으로 변경 가능한 값은 $p(x)$

이를 조절하여 가장 큰 C를 얻는 distribution으로 채널을 구성해야 효율이 가장 좋음

Data Processing inequality

Data flow



▼ Def: Markov Chain

X, Y, Z : R.V.s

$X \to Y \to Z$ (i.e. Markov chain) ⇔ $p(x,y,z) = p(x)p(y|x)p(z|y)$

and we also can denote as $X \longleftrightarrow Y \longleftrightarrow Z$

> 💡 For $X_1, X_2, \dots, X_n$,
> $X_1 \to X_2 \to \cdots X_{n-1} \to X_n \to \cdots$ is Markov chain when
> $P(X_n|Xn-1, \dots, X_1) = P(X_n|X_{n-1})$

Def ⇔ Z depends only on Y — (1)

⇔ Z is conditionally independent of X given Y — (2)

i.e. $P(X, Z \mid Y) = P(X|Y)P(Z|Y)$

> 💡 Conditional probability
> $p(x,y) = p(x)p(y|x)$
> $p(x,y,z) = p(x)p(y,z \mid x) = p(x)p(y|z)p(z \mid y, x)$
> ...
> $p(x_1, x_2, \dots, x_n) = \prod_{i=1}^{n} p(x_i \mid x_{i-1}, \dots, x_1)$

(1) For Markov chain $X \to Y \to Z$,

we need to show $p(x,y,z) = p(x)p(y|x)p(z \mid y, x) = p(x)p(y|x)p(z|y)$

By the definition, $p(z| \, y, x) = p(z|y)$ is trivial.

> 💡 Z depends on X only through Y
> 즉, X의 변화는 Y를 통해서만 Z에 반영됨
> ⇔ X와 Z가 depend하지만, Y를 통해서만 depend하므로 Y에만 depend하 보임
> ⇔ Z가 X로부터 depend하는 정도가 이미 Y에 반영되어 있음

Thus, $p(x,y,z) = p(x)p(y|x)p(z|y)$

(2) Given Y,

$p(x, z \, |y) = p(x,y,z)/p(y)$ and by the property of Markov chain,

$= [p(x)p(y|x)p(z|y)]/p(y) = [p(x,y)/p(y)]p(z|y) = p(x|y)p(z|y)$

Therefore, X, Z are independent given Y

By (1), (2), $X \to Y \to Z \iff X, Z$ are independent given Y

and it also can be represented as Z, X are independent given Y

then, $\iff Z \to Y \to X$ holds.

We may write $X \longleftrightarrow Y \longleftrightarrow Z$

Then if $Z = f(Y)$, $X \to Y \to f(Y)$?

직관적으로 생각할 때, Z는 Y가 결정되는 순간 같이 결정됨

i.e., Y determines Z completely.

It means Z depends only Y and is the definition of Markov chain.

So, this chain is holds.

> 💡 R.V.s X, Y, Z가 다음 중 하나라도 만족하면 Markov chain임
> - $P(Z|Y) = P(Z| \, Y, X)$
> - $P(X,Y,Z) = P(X)P(Y|X)P(Z|Y)$
> - $P(Z|Y)P(X|Y) = P(X, Z \, |Y)$

▼ ex

$X \xrightarrow[channel\,1]{BSC} Y \xrightarrow[channel\,2]{BSC} Z,$

error rate of channel1 = p_1, error rate of channel2 = p_2

BSC channel 1

Input data

X	0	1
$P(X)$	3/4	1/4

given $X = 0$

Y	0	1
$P(Y\|X=0)$	$1-p_1$	p_1

given $X = 1$

Y	0	1
$P(Y\|X=1)$	p_1	$1-p_1$

Joint distribution

$X \setminus Y$	0	1	$P(X)$
0	$\frac{3}{4}(1-p_1)$	$\frac{3}{4}p_1$	3/4
1	$\frac{1}{4}p_1$	$\frac{1}{4}(1-p_1)$	1/4
$P(Y)$	$\frac{3}{4}-\frac{1}{2}p_1$	$\frac{1}{4}-\frac{1}{2}p_1$	

Check: $P(X|Y)$

given Y (빈 칸 채우기)

given $Y = 0$

Y	0	1
$P(X\|Y=0)$		

given $Y = 1$

Y	0	1
$P(X\|Y=1)$		

$Z = f(Y) := 2Y - 1$ 라면

Joint distribution

$p_1 = 1/4$일 때 Y의

input data

$P(Y=0) = 5/8$

$P(Y=1) = 3/8$

$Y \setminus Z$	-1	1	$P(Y)$
0	5/8	0	5/8
1	0	3/8	3/8
$P(Z)$	5/8	3/8	

Y의 분포를 알면 Z의 분포가 fix됨

(X의 분포를 모르고 Y 분포만 알아도 고정됨, 개입할 여지가 없어짐)

Theorem: Data Processing Inequality

▼ $X \to Y \to Z \implies I(X;Y) \geq I(X;Z)$

💡 **Probability**

$p(x_1, \dots, x_n) = \prod_{i=1}^{n} p(x_i \,|\, x_{i-1}, \dots, x_1)$

Entropy

$$\begin{aligned} H(x_1, \dots, x_n) &= \textstyle\sum_{i=1}^{n} H(x_i \,|\, x_{i-1}, \dots, x_1) \\ &= H(x_1) + H(x_2|x_1) + H(x_3 \,|\, x_2, x_1) + \cdots \end{aligned}$$

Mutual information

$I(X;Y) = H(X) - H(X|Y)$

$I(X_1, \dots, X_n \,;\, Y) = \sum_{i=1}^{n} I(X_i; Y \mid X_{i-1}, \dots, X_1)$

pf: Since $I(X;\; Y, Z) = I(Y, Z\,; X) = I(Z, Y\,; X)$

$= I(Z; X) + I(Y; X\,|Z) = I(X; Z) + I(X; Y\,|Z)$

$= I(X;\; Z, Y) = I(X; Y) + I(X; Z\,|Y),$

$I(X;Z) + I(X;Y\,|Z) = I(X;Y) + I(X;Z\,|Y).$

$I(X;Z\,|Y) = 0$ by property of Markov Chain

- X, Z are conditionally independent given $Y \Rightarrow I(X;Z\,|Y) = 0$

$I(X;Y) = I(X;Z) + I(X;Y\,|Z) \geq I(X;Z)$ $\quad\quad \because I(A;B) \geq 0$

💡 직관적으로 생각해볼 때, X로부터 Y를 아는 것이 X로부터 Z를 아는 것보다 쉬움 (알 수 있는 정보가 더 많음). 이때 등호가 성립한다는 것은 Y를 거쳐 Z에서 X가 영향을 미치는 정도에 손실이 없음을 의미함

▼ Corollary

$Z = g(Y) \implies I(X;Y) \geq I(X;g(Y))$

pf: $X \longrightarrow Y \longrightarrow g(Y)$

$Y \rightarrow g(Y)$는 최대 1-1 대응 관계이므로 $g(Y)$의 정보는 Y보다 가질 수 있는 값의 범위가 작음

⇒ trivial!

▼ Corollary

$X \longrightarrow Y \longrightarrow Z \implies I(X;Y|Z) \leq I(X;Y)$

pf: Since $I(X;\ Y, Z) = I(X;Y) + I(X;Z|Y) = I(X;Y)$

and $I(X;\ Y, Z) = I(X;Z) + I(X;Y|Z)$,

$\Rightarrow\ I(X;Y) = I(X;Z) + I(X;Y|Z) \geq I(X;Y|Z) \quad \because I(X;Z) \geq 0$

💡 $X \longrightarrow Y \longrightarrow Z$
The dependency of X and Y is decreased (or unchanged) by the observation of a downstream (to receiver) R.V. Z
i.e., $I(X;Y) \geq I(X;Y|Z)$

▼ ex

case1) X_1, X_2, X_3 : independent R.V.s

case2) X_1 : dice {1, 2, ... , 6}, X_2 : biased coin {0, 1}, $X_3 = 2X_2 - 1$

X_1	1	...	6	X_2	0	1
$P(X_1)$	1/6		1/6	$P(X_2)$	$1-\theta$	θ

$\theta := \frac{1}{X_1} \in \{1, \frac{1}{2}, \frac{1}{3}, \ldots, \frac{1}{6}\}$

$X_1 = 2$ ⇒ fair coin toss

Which case is a Markov chain as $X \longrightarrow Y \longrightarrow Z$?

Definition : $p(x,y,z) = p(x)p(y|x)p(z|y)$

and we can lead $P(Z|Y)P(X|Y) = P(X, Z|Y)$

⇒ given Y, $X$and Z are independent from Markov chain

case1) and case2) are both satisfied the definition of Markov chain

case1) $P(X_1)P(X_3) = P(X_1, X_3)$, $P(X_1, X_3|X_2) = P(X_1|X_2)P(X_3|X_2)$

case2) $X \longrightarrow Y \longrightarrow g(Y)$ 형태이면 Markov chain

X_2 depends X_1, X_3 is determined by X_2.

따라서 X_3는 X_1에 depend하지만, X_2를 통해서 영향을 받음

(X_1은 X_2를 통해서가 아닌 방법으로 X_3에 영향을 줄 수 없음)

Caution!

$I(X;Y|Z) > I(X;Y)$ but $I(X;Y|Z) \neq I(X;Y)$ is possible

▼ ex: X, Y are fair coin tosses (independent)

$Z := X + Y$ — Not Markov chain

X	0	1
$P(X)$	1/2	1/2

Y	0	1
$P(Y)$	1/2	1/2

$X \setminus Y$	0	1	$P(X)$
0	1/4	1/4	1/2
1	1/4	1/4	1/2
$P(Y)$	1/2	1/2	

$X, Y \setminus Z$	0	1	2	$P(X,Y)$
0, 0	1/4	0	0	1/2
0, 1	0	1/4	0	1/2
1, 0	0	1/4	0	1/4
1, 1	0	0	1/4	1/4
$P(Z)$	1/4	1/2	1/4	

X	0	1
$P(X\|Z=1)$	1/2	1/2

- $I(X;Y)=0$ since X, Y are independent
- $I(X;Y|Z) = H(X|Z) - H(X|Y,Z)$
- $H(X|Y,Z)=0$ since given Y, Z, $X = Z - Y$ determined

$$\begin{aligned} I(X;Y|Z) &= H(X|Z) \\ &= P(Z=0)H(X|Z=0) \quad \longrightarrow P(X=0)=1 \quad (X=0, Y=0) \\ &\quad +P(Z=1)H(X|Z=1) \\ &\quad +P(Z=2)H(X|Z=2) \quad \longrightarrow P(X=1)=1 \quad (X=1, Y=1) \\ &= P(Z=1)H(X|Z=1) = \tfrac{1}{2}H\left(\tfrac{1}{2}, \tfrac{1}{2}\right) = 1/2 \end{aligned}$$

즉, $I(X;Y) = 0 < I(X;Y|Z) = 1/2$

Markov chain이면 부등호가 반대여야 하므로 $X \longrightarrow Y \longrightarrow Z$는 Markov chain이 아님

Sufficient Statistics

- Given $\theta \longrightarrow X \longrightarrow T(X)$, if $I(\theta; T(X)) = I(\theta; X)$, then $T(X)$ is called sufficient for θ.
  - $\{f_\theta(x)\}$: family of p.mp.f. indexed by θ ; 1-parameter family

    θ가 결정되면 f가 결정되고 x에 의한 값을 얻을 수 있는 확률 질량 함수

    (θ : index parameter, θ에 대하여 확률적인 부분이 있다면 R.V.)

    → parameter 하나가 결정되면 함수의 변수 X의 확률 분포를 알 수 있음

    ex:

X	0	1
$P(X)$	$1-\theta$	θ

$P(X=1)=\theta$

  $X \sim f_\theta(x)$: X is a R.V. from a distribution in this family $\{f_\theta(x)\}$

  coin마다 θ가 정해져 있음

  $\{f_\theta(x)\} \longrightarrow f_\theta(x) \longrightarrow X$

  family → p.m.f. → R.V.

  $X \longrightarrow T(X)$: statistic, X로부터 얻은 통계량, funtion of X

  Then, $\theta, X, T(X)$ form a Markov chain. i.e., $\theta \longrightarrow X \longrightarrow T(X)$

  $\Rightarrow \theta \longrightarrow X \longrightarrow T(X) \implies I(\theta; T(X)) \le I(\theta; X)$

  If $I(\theta; T(X)) = I(\theta; X)$, then $T(X)$ is called sufficient for θ.

  실험 데이터를 해석할 때,

  - 실험 대상이 family로부터 실험 대상을 거쳐 실험 데이터를 얻기까지

    ex: 현실에서 실험을 수행하면 family에 대한 정보는 알 수 없고 실험 결과만 알 수 있음

    동전을 10번 던져서 8번이 H가 나오면,

    H가 나올 확률이 $\theta = 4/5$인 동전을 사용했다고 추정하는 것이 타당함.

    물론 실제로 사용한 동전이 H가 4/5의 확률로 나오는 동전인지는 알 수 없음

    H가 나올 확률이 1/3인데 던지는 방식에 의해 H만 나왔을 수도 있음

    이렇듯 n번의 시행으로부터 얻은 데이터에서 $T(X)$를 구하고

    개별 시행의 결과는 시행 할 때마다 다르겠지만 $T(X)$를 통해 n번 중 H가 나오는 횟수의 기대값 θ를 추정할 수 있다고 요약 가능함

    ⇒ sufficient statistics

    = 통계량을 잘 잡으면 몇 회의 시행이 있었는지,

    몇 번 원하는 결과가 나왔는지 모르더라도 θ를 추정할 수 있음

  information loss가 하나도 없음을 보이기 위한 가정으로 쓰임

- Def
  - Informal version

    Informmaly, $T(X)$ is called sufficient for θ if it contains all information in X about θ.

    즉, θ에 대해서 X가 가지고 있는 information을 $T(X)$가 모두 가지고 있음

  - 1st formal version

    $T(X)$ is said to be a sufficient statistic relative to $\{f_\theta(x)\}$

    if X is independent of θ given $T(X)$ for any distributioni $f_\theta(x)$.

즉, $T(X)$를 알고 있을 때 X와 θ가 indepent함

▼ X와 θ의 depend한 정보는 모두 $T(X)$에 속해있음

X와 θ는 독립이 아님 $\Rightarrow$ X와 θ가 독립이면 서로의 정보를 가지고 있지 않음
(직관적으로 독립이면 서로 무관하기 때문에 영향을 주고 받고를 따질 의미가 없어 보임)

주어진 $T(X)$에 대해서 알고 있을 때 X와 θ가 독립이라는 말은
X와 θ가 주고받는 정보를 $T(X)$가 모두 가져갔음을 의미함

$\Rightarrow T(X)$에 속한 X, θ의 정보를 제외하면 X와 θ는 독립임

- 2nd formal version

If $\theta \longrightarrow X \longrightarrow T(X)$ is Markov chain then $\theta \longrightarrow T(X) \longrightarrow X$

▼ $T(X)$가 θ와 X를 연결해주는 모든 정보를 가지고 있음

$\Leftrightarrow$ θ와 X를 연결해주는 $T(X)$가 주어지면 X와 θ가 indepent함

$\Leftrightarrow$ $T(X)$를 알고 있을 때 X와 θ가 indepent함

💡 If $T(X)$ is a sufficient statistic, $I(\theta\,;T(X)) = I(\theta\,;X)$

▼ i.e., No information loss
2nd formal version definition shows that $I(\theta\,;T(X)) \geq I(\theta\,;X)$.
And in general, $\theta \longrightarrow X \longrightarrow T(X)$ then $I(\theta\,;T(X)) \leq I(\theta\,;X)$.
Therefore, if $T(X)$ is a sufficient statistic, $I(\theta\,;T(X)) = I(\theta\,;X)$.

💡 Where $X_1, \dots, X_n \sim f_\theta(x)$, $Y := g(X_1, \dots, X_n)$ is sufficient for θ

▼ if $P(X_1 = x_1, \dots, X_n = x_n \,|Y = y)$ does not depend on θ.

Let $X_1, \dots, X_n$ be R.V.s from a p.m.f. with parameter θ
i.e., $X_1, \dots, X_n \sim f_\theta(x)$
and $Y := g(X_1, \dots, X_n)$ is a funtion of $X_1, \dots, X_n$
i.e., Y is deterministic for $X_1, \dots, X_n$
Then Y is sufficient for θ,
if the conditional probability $P(X_1 = x_1, \dots, X_n = x_n \,|Y = y)$
does not depend on θ.

Since $Y = g(X_1, \dots, X_n)$, $\theta \longrightarrow Y \longrightarrow (X_1, \dots, X_n)$.
$\Rightarrow P(X_1 = x_1, \dots, X_n = x_n \,|Y = y)$
$= P(X_1 = x_1, \dots, X_n = x_n \,|Y = y, \Theta = \theta)$
즉, 주어진 Y에 대하여 θ와 $(X_1, \dots, X_n)$가 independent.

▼ ex. Bernoulli distribution

X	0	1
$P(X)$	$1-\theta$	θ

For sample space χ, $x \in \chi = \{0, 1\}$.

$$P_\theta(X = x) = \begin{cases} \theta & x = 1 \\ 1-\theta & x = 0 \end{cases} \quad \text{(p.m.f.)}$$
$$= \theta^x (1-\theta)^{1-x}$$

Let $X_1, \dots, X_n$ be R.V.s sample from a Bernoulli distribution with θ.

Bernoulli : i.i.d.(independent and identically distributed)

i.e., $P(X_1 = x_1, \dots, X_n = x_n) = P(X_1 = x_1) \cdots P(X_n = x_n)$
$$= \theta^{x_1}(1-\theta)^{1-x_1} \cdots \theta^{x_n}(1-\theta)^{1-x_n}$$
$$= \textstyle\prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i}$$

Introduce a statistic $Y := T(X_1, \dots, X_n) = \sum_{i=1}^n X_i$.

$\Rightarrow$ counting 1's out of $x_1, \dots, x_n$ $\Leftrightarrow$ counting the number of Heads in n tosses

$Y \sim B(n, \theta)$: Binomial ditribution

$\Rightarrow P(Y = k) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$

Consider the conditional probabiltiy

$$P(X_1 = x_1, \dots, X_n = x_n) = \frac{P(X_1 = x_1, \dots, X_n = x_n \text{ and } Y = k)}{P(Y = k)}$$
$$= \frac{\prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i}}{\binom{n}{k}\theta^k(1-\theta)^{n-k}} \quad \text{where } \textstyle\sum_{i=1}^n x_i = k$$
$$= \frac{\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}}{\binom{n}{k}\theta^k(1-\theta)^{n-k}}$$

$$= \frac{\theta^k(1-\theta)^{n-k}}{\binom{n}{k}\theta^k(1-\theta)^{n-k}}$$
$$= \frac{1}{\binom{n}{k}}$$

$\Rightarrow P(X_1 = x_1, \dots, X_n = x_n) = \frac{1}{\binom{n}{k}}$ does not depend on θ,

which means that given $Y = k \, (= \sum_{i=1}^n x_i)$,

the individual values of the x_i's cannot provide additional information about θ.

즉, Y 값이 주어지면 x_i 각 값은 (각 x_i가 0인지 1인지는) θ에 대한 추가적인 정보를 제공하지 않음

- ex. Poisson distribution

p.m.f. $p(k) = \frac{\lambda^k e^{-\lambda}}{k!}$ for $k = 0, 1, 2, \dots$ (λ를 알면 모든 parameter가 결정됨)

Let $X_1, \dots, X_n$ be sample from a Poisson distiribution with λ and $Y = \sum_{i=1}^n X_i$ (=통계량)

※ r : average rate of the event

t : time interval ($\lambda = rt$: 주어진 시간 t동안 event가 발생하는 횟수의 평균, 빈도)

$\Rightarrow P(k \text{ events in interval } t) = \frac{(rt)^k e^{-rt}}{k!}$

$Y = \sum_{i=1}^n X_i \sim$ Poisson dist with $\lambda y = n\lambda$

- Y is sufficient for λ.

Consider the conditional probability

$$P(X_1 = x_1, \dots, X_n = x_n \,|Y = y) = \frac{P(X_1 = x_1, \dots, X_n = x_n \text{ and } Y = y)}{P(Y = y)}$$
$$= \frac{P(X_1 = x_1, \dots, X_n = x_n)}{P(Y = y)} \quad \text{where } \sum x_i = y$$

(note that $\sum x_i \neq y \Rightarrow P(*) = 0$)

$$= \frac{\prod_{i=1}^n \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}}{\frac{(n\lambda)^y e^{-n\lambda}}{y!}}$$
$$= \frac{\frac{\lambda^{x_1 + \cdots + x_n} e^{-\lambda}}{x_1! x_2! \cdots x_n!}}{\frac{(n\lambda)^y e^{-n\lambda}}{y!}}$$
$$= \frac{\frac{\lambda^y e^{-\lambda}}{x_1! \cdots x_n!}}{\frac{(n\lambda)^y e^{-n\lambda}}{y!}}$$
$$= \frac{y!}{n^y x_1! \cdots x_n!} \quad \text{does not depend on } \lambda.$$

- Thm: Factorization Theorem

$Y = g(X_1, \dots, X_n)$ is sufficient for θ
$\iff P(x_1, \dots, x_n \,|\, \theta) = \phi(g(x_1, \dots, x_n) \,|\, \theta) h(x_1, \dots, x_n)$

Let $X_1, \dots, X_n$ be i.i.d. R.V.s from p.m.f. $f_\theta(x) \in \{f_\theta(x)\}$,

$Y = g(X_1, \dots, X_n)$ is sufficient for θ

$\iff P(x_1, \dots, x_n \,|\, \theta) = \phi(g(x_1, \dots, x_n) \,|\, \theta) h(x_1, \dots, x_n)$

where ϕ depends on x_i's only through g (i.e., Y) and h does not depend on θ.

θ가 주어졌을 때 각 X_i가 값 x_i를 가질 확률에 대하여

θ에 의존하는 값들의 확률은 ϕ, 즉 $g(x_1, \dots, x_n) = Y$에 몰려있음 (h는 θ에 independent)

$\Rightarrow Y$로 표현하지 않는 값들은 θ에 의존하지 않음

- Remark

💡 $P(x_1, \dots, x_n) := P(X_1 = x_1, \dots, X_n = x_n)$
$P(X_1, \dots, X_n)$가 $x_1, \dots, x_n$을 어떻게 결정하느냐에 따라 달라짐을 반영한 표기임

💡 $P(*; \theta) = P(*|\theta)$이기 위해서는 θ가 R.V.여야 함
우항의 표기를 사용하고 싶지만, θ의 분포를 고려하고 싶지 않음
좌항의 θ는 determined θ 또는 fixed θ, indexed by θ이고
우항의 θ는 p.m.f.를 갖는 R.V.임

- ex: For Bernoulli distribution,

$P(x_1, \dots, x_n \,; \theta) = \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i}) = \theta^{\sum x_i}(1-\theta)^{n - \sum x_i}$

Let $y = \sum_{i=1}^n x_i$,

$\theta^{\sum x_i}(1-\theta)^{n-\sum x_i} = \theta^y(1-\theta)^{n-y} \times 1 := \phi(y \,; \theta) \times h(x_1, \dots, x_n)$

⇒ Bernoulli distribution에 대하여

$P(x_1, \ldots, x_n\,; \theta) = \phi(y\,; \theta) \times h(x_1, \ldots, x_n)$를 만족하므로

$Y = \sum X_i$는 θ에 대하여 sufficient statistic임

▼ ex: For Poisson distribution,

$p(x_1, \ldots, x_n\,; \lambda) = \frac{\lambda^{x_1+\cdots+x_n} e^{-n\lambda}}{x_1! \cdots x_n!} = \lambda^y e^{-n\lambda} \times \frac{1}{x_1! \cdots x_n!}$

⇒ Poisson distribution에 대하여

$P(x_1, \ldots, x_n\,; \lambda) = \phi(y\,; \lambda) \times h(x_1, \ldots, x_n)$를 만족하므로

$Y = \sum X_i$는 λ에 대하여 sufficient statistic임

Minimal sufficient statistic

Def: $T(X)$ is a minimal sufficient statistic relative to $\{f_\theta(x)\}$
if it is a function of every other sufficient statistic U.

즉, sufficient statistic이면서 다른 sufficient statistic로 표시되는 함수

$\theta \longrightarrow T(X) \longrightarrow U(X) \longrightarrow X$

💡 A minimal sufficient statistic maximally compresses the information about θ in the sample.
Other sufficient statistics may contain additional irrelevant information.

▼ ex

- $Y = X_1 + \cdots + X_n$ is minimal sufficient statistic for θ in the independent coin toss example.

  Let $Y_{odd} = X_1 + X_3 + \cdots + X_{2n-1}$, $Y_{even} = X_2 + X_4 + \cdots + X_{2n}$ and
  $\widetilde{Y} = (Y_{odd}, Y_{even})$, then $\widetilde{Y}$ is a sufficient statistic

  $Y = g(\widetilde{Y}) = g(Y_{odd}, Y_{even}) = Y_{odd} + Y_{even}$

  ⇒ Y와 비교할 때, $Y_{odd} + Y_{even}$는 불필요하게 세부적으로 나뉜 함수임

- 학교에서 전교 수학점수 평균을 알길 바랄 때,
  - 전체 데이터로부터 각 반의 평균 점수를 구하고 이것들을 모아서 다시 전체 평균을 구하는 것보다
  - 전체 데이터로부터 한 번에 전체 평균을 구하는 것이 훨씬 효율적임
    (반 평균이라는 불필요한 정보가 포함되지 않음)