KL Divergence 완전정복!
-PRML을 바탕으로-

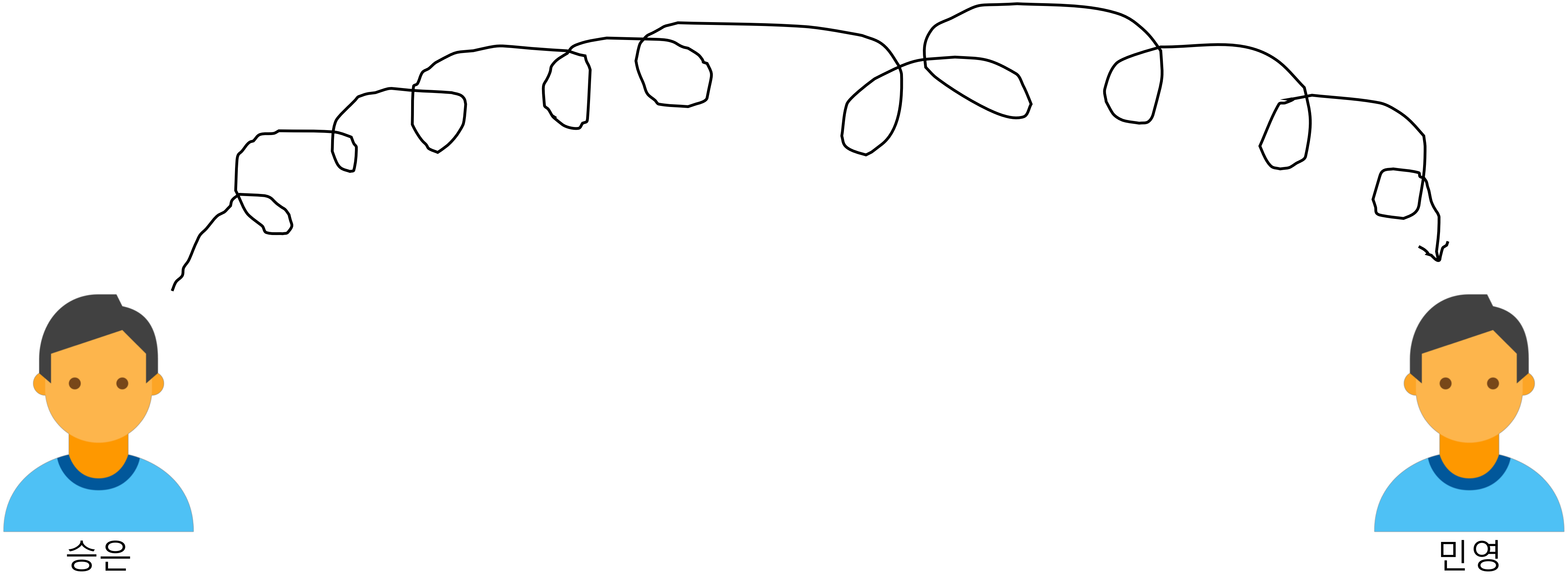
팡요랩

2019.01.20

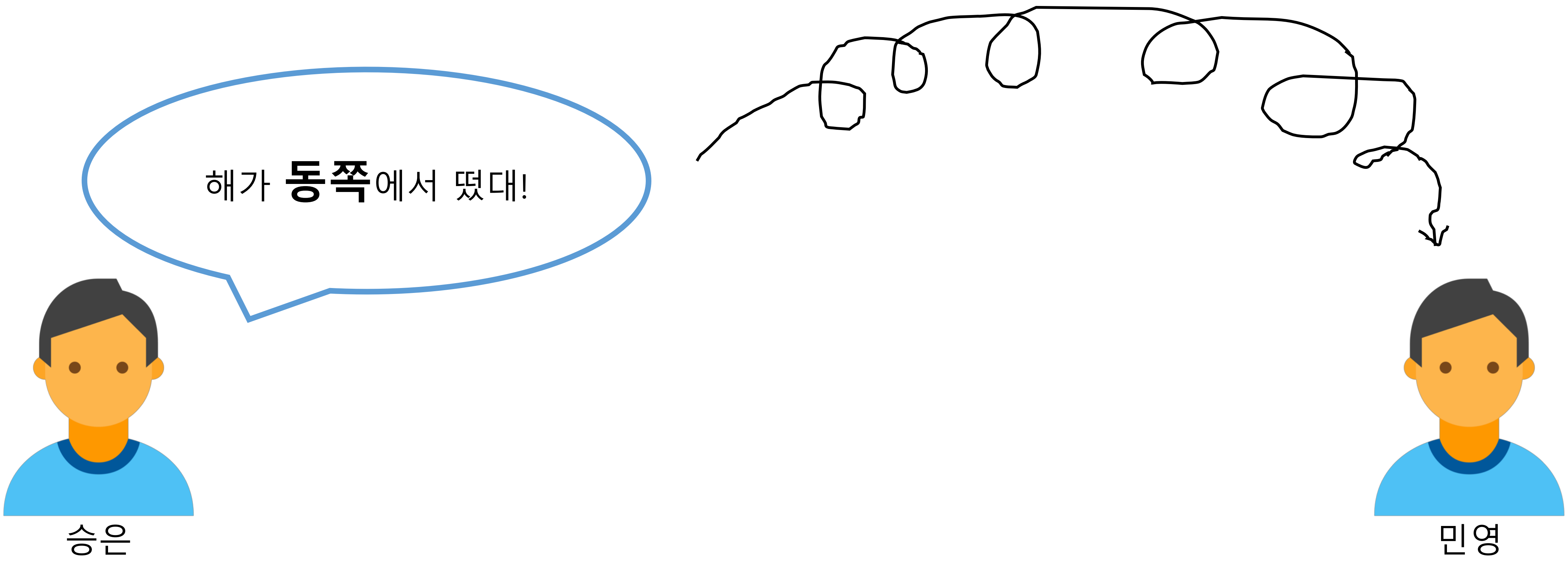
오늘 다룰 주제 : Information Theory

- 왜 엔트로피는 $H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i).$ 형태인가?
- KL Divergence는 어떤 의미인가?
- KL Divergence와 Cross-Entropy의 차이는?
- Mutual Information 은 무엇인가?

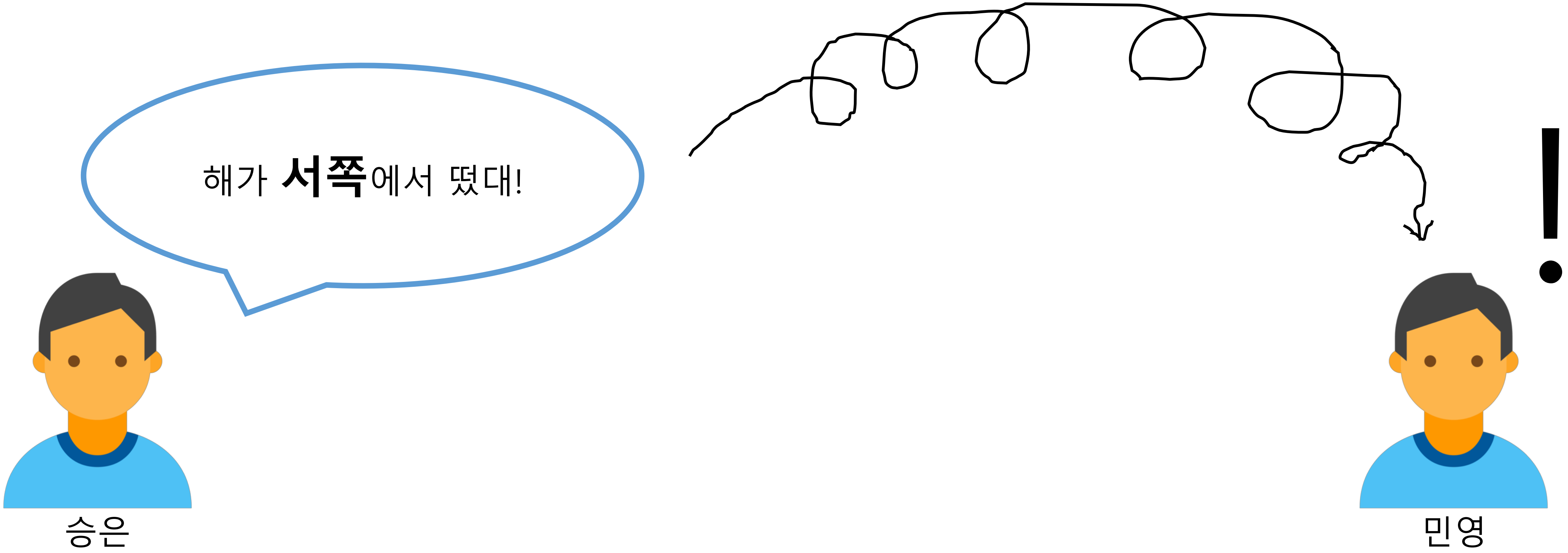
승은은 민영에게 통신을 하고 있다.



뻔한 이야기는 정보량이 거의 없다.
민영은 관심도 없다.



뻔하지 않은 이야기는 정보량이 크다.



h의 첫 번째 조건

- 확률 변수(Random variable) X
  1. X는 East, West 두 가지 값을 가질 수 있음.

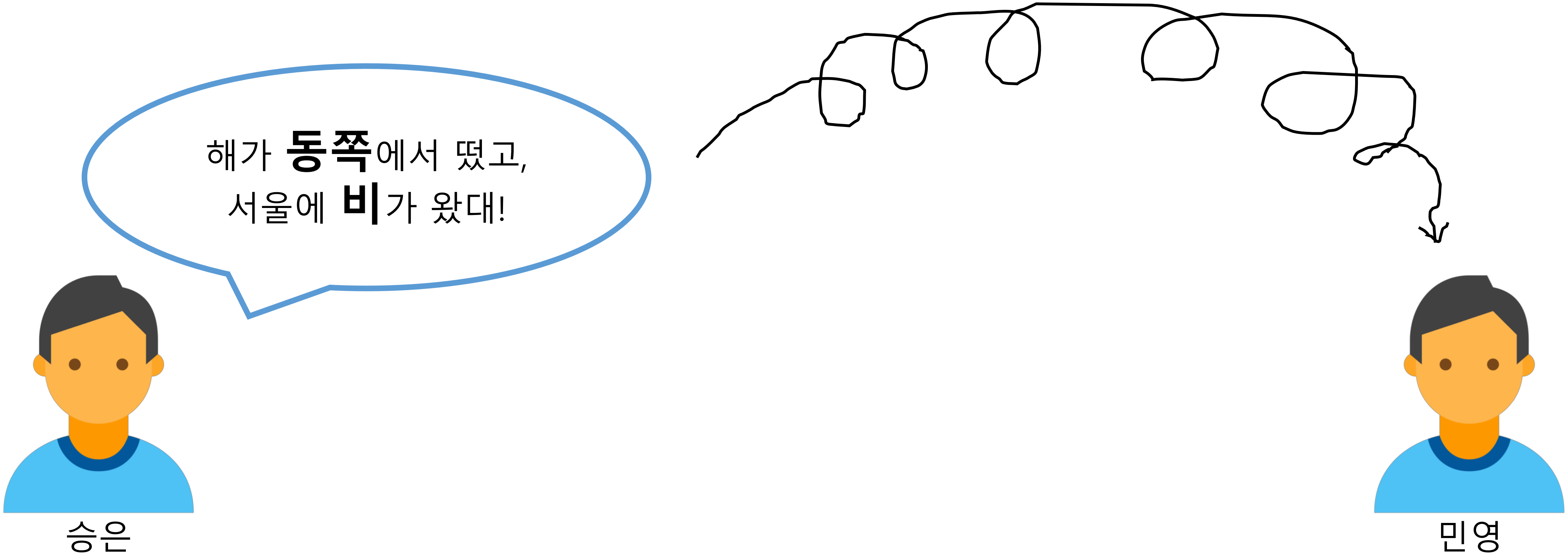
- X 의 정보량 h(x) 는 p(x)에 대한 함수.   즉, h=f(p)

- $p(east) = 0.99999999$, $p(west) = 0.00000001$

- $h(west) > h(east)$ 여야 함.

- p(x)와 h(x)는 monotonic한 관계여야 한다. 즉, f는 단조 감소 함수.

두 사건을 알려주면..?



h의 두 번째 조건

- 확률 변수(Random variable) X, Y
  1. X는 East, West 두 가지 값
  2. Y는 Rain, Not Rain 두 가지 값
  3. X,Y 는 독립

- $h(x, y) = h(x) + h(y)$
- $p(x, y) = p(x) * p(y)$

즉, f(p(x,y)) = f(p(x) * p(y)) = f(p(x)) + f(p(y))

- 이를 만족하는 f는 => log 함수.
- 첫 번째 조건과 결합하면

$$\boldsymbol{h(x) = -log_2 p(x)}$$

예시

$$\boldsymbol{h(x) = -log_2 p(x)}$$

- $h(east) = -log_2 p(east) = -log_2(0.99999999) = 0.000000014$
- $h(west) = -log_2 p(west) = -log_2(0.00000001) = 26.5754247591$

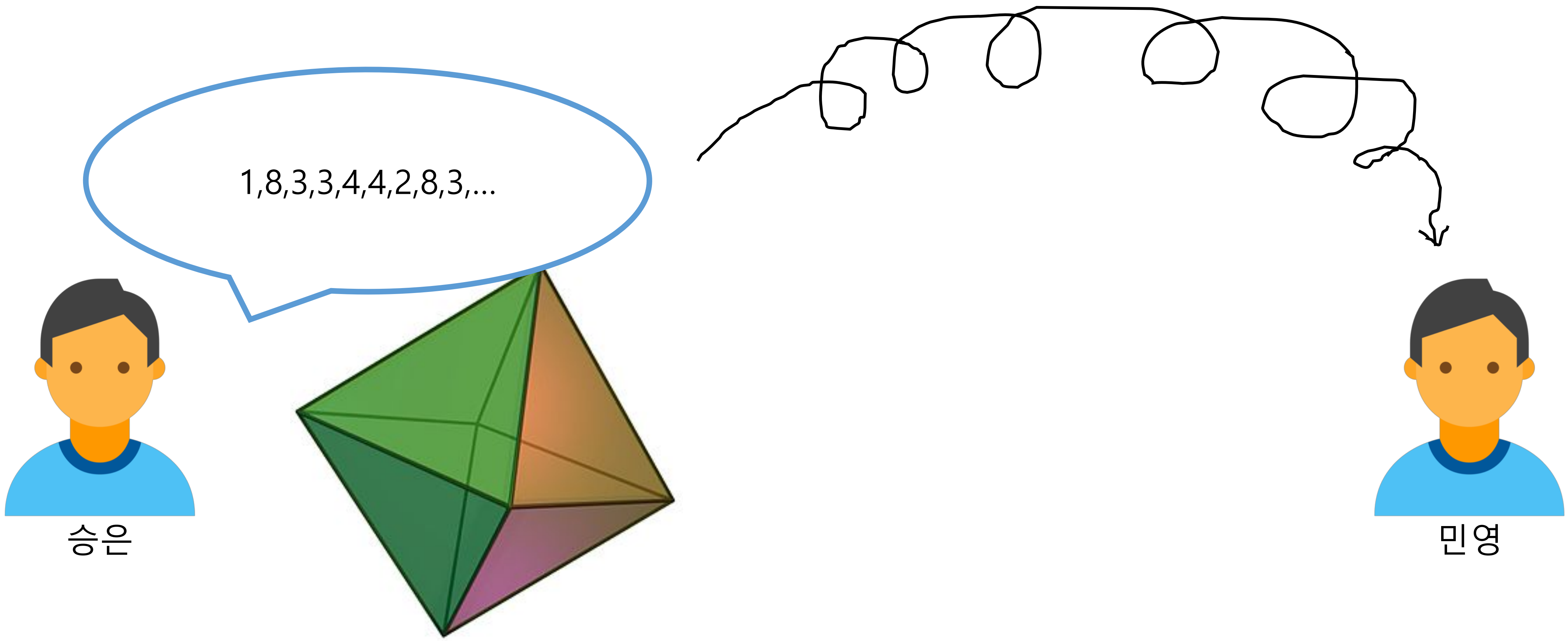
그러면 평균적인 정보량은 ..?

- $p(east) * h(east) + p(west) * h(west) = 0.99999999 \ * 0.000000014 + 0.00000001 \ * 26.6$
- 보다 일반적으로는,

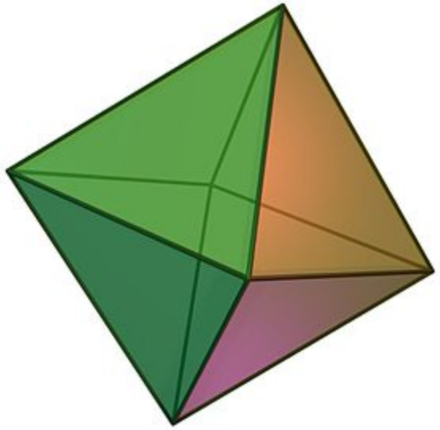
$$H[X] = -\sum_x p(x) log_2 p(x) = E_p[-log_2 p(x)]$$

- 이 값이 바로 ENTROPY !!

정 8 면체 주사위



1번 주사위

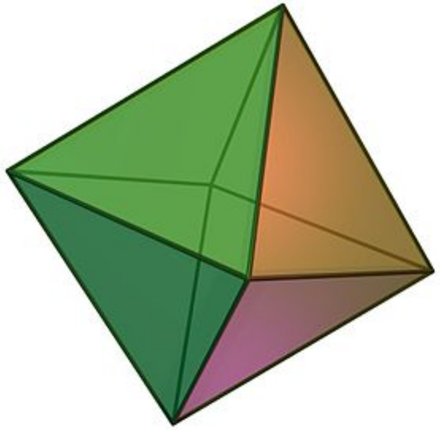


8개의 면의 확률이 균일함.
-> (1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8)

이때 엔트로피 값을 구해보면

$$\mathrm{H}[x] = -8 \times \frac{1}{8} \log_2 \frac{1}{8} = 3 \text{ bits.}$$

2번 주사위



8개의 면의 확률이 불균일함.
-> (1/2, 1/4, 1/8, 1/16, 1/64, 1/64, 1/64, 1/64)

이때 엔트로피 값을 구해보면

$$\mathrm{H}[x] = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{8}\log_2\frac{1}{8} - \frac{1}{16}\log_2\frac{1}{16} - \frac{4}{64}\log_2\frac{1}{64} = 2 \text{ bits.}$$

8개의 값을 각각 0, 10, 110, 1110, 111100, 111101, 111110, 111111 로 coding 하면

$$\text{average code length} = \frac{1}{2}\times 1 + \frac{1}{4}\times 2 + \frac{1}{8}\times 3 + \frac{1}{16}\times 4 + 4\times\frac{1}{64}\times 6 = 2 \text{ bits}$$

Entropy 의 몇 가지 특징

- Continuous Variable 인 경우

$$\mathrm{H}[\mathbf{x}] = \lim_{\Delta \to 0} \left\{ \sum_i p(x_i)\Delta \ln p(x_i) \right\} = -\int p(x) \ln p(x)\, \mathrm{d}x$$

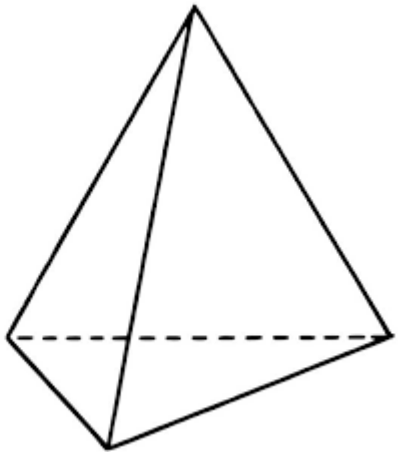
- Entropy 는 Average Coding Length의 Lower Bound!

- Entropy Maximize ?
  1. Discrete variable : Uniform
  2. Continuous variable : Gaussian

- Entropy Minimize ?
  1. 한 점에 확률이 다 몰려 있는 경우! -> 0이 된다.

드디어 KL Divergence …

승은이는 바보.



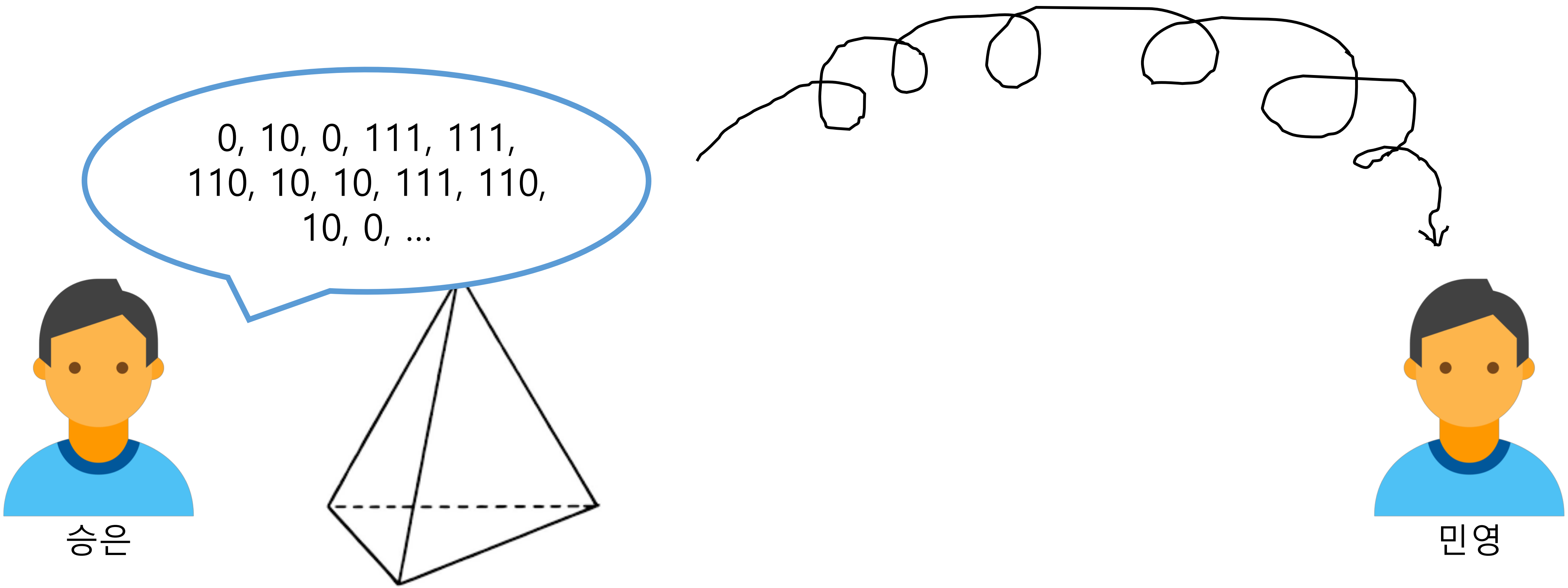
- 실제 주사위 4 면의 확률분포
  p = (1/4, 1/4, 1/4, 1/4)

- 승은이가 생각한 4 면의 확률분포
  q = (1/2, 1/4, 1/8, 1/8)

- 그래서 승은이는 4 상태를 각각 0, 10, 110, 111로 코딩하였다.

- 실제 최적의 코딩은 00, 01, 10, 11 이다.

정 4 면체 주사위



- 이 때 평균 coding length 는

$$\frac{1}{4} * 1 + \frac{1}{4} * 2 + \frac{1}{4} * 3 + \frac{1}{4} * 3 = 2.25$$

- 즉, $-\sum_x p(x) log_2 q(x)$

$$=- \frac{1}{4} * log_2(0.5) - \frac{1}{4} * log_2(0.25) - \frac{1}{4} * log_2(0.125) - \frac{1}{4} * log_2(0.125) = 2.25$$

- 승은이가 p를 정확하게 모델링 했을 경우

$$-\sum_x p(x) log_2 p(x) =$$

$$-\frac{1}{4} * log_2(0.25) - \frac{1}{4} * log_2(0.25) - \frac{1}{4} * log_2(0.25) - \frac{1}{4} * log_2(0.25) = 2$$

- 모델링한 q가 p와 다르기 때문에 발생한 추가 비용!

=> 2.25 – 2 = 0.25

- 즉, 모델링 오류때문에 발생한 추가 비용은

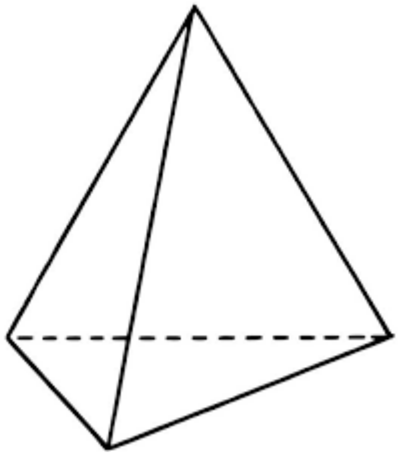
$$\left(-\sum_{x} p(x) log_2 q(x)\right) - \left(-\sum_{x} p(x) log_2 p\,(x)\right) = -\sum_{x} p(x) log_2 \frac{q(x)}{p(x)}$$

- Continuous variable의 경우는

$$\begin{aligned} \mathrm{KL}(p\|q) &= -\int p(\mathbf{x}) \ln q(\mathbf{x})\, \mathrm{d}\mathbf{x} - \left(-\int p(\mathbf{x}) \ln p(\mathbf{x})\, \mathrm{d}\mathbf{x}\right) \\ &= -\int p(\mathbf{x}) \ln \left\{\frac{q(\mathbf{x})}{p(\mathbf{x})}\right\} \mathrm{d}\mathbf{x}. \end{aligned}$$

p와 q가 바뀌면 ...?

p와 q가 바뀌면 ...?



승은

- 실제 주사위 4 면의 확률분포
  $p \ = (1/2, 1/4, 1/8, 1/8)$
- 승은이가 생각한 4 면의 확률분포
  $q \ = (1/4, 1/4, 1/4, 1/4)$
- 그래서 승은이는 4 상태를 각각 00, 01, 10, 11로 코딩하였다.
- 실제 최적의 코딩은 0, 10, 110, 111 이다.

- 이 때 평균 coding length 는

$$\frac{1}{2}*2+\frac{1}{4}*2+\frac{1}{8}*2+\frac{1}{8}*2=2.0$$

- 즉, $-\sum_x p(x)log_2q(x)$

$$=-\ \frac{1}{2}*log_2(0.25)-\frac{1}{4}*log_2(0.25)-\frac{1}{8}*log_2(0.25)-\frac{1}{8}*log_2(0.25)=2.0$$

- 승은이가 p를 정확하게 모델링 했을 경우

$$-\sum_x p(x)log_2p(x) =$$

$$-\ \frac{1}{2}*log_2(0.5)-\frac{1}{4}*log_2(0.25)-\frac{1}{8}*log_2(0.125)-\frac{1}{8}*log_2(0.125)=1.75$$

- 모델링한 q가 p와 다르기 때문에 발생한 추가 비용!

=> 2 – 1.75 = 0.25

KL Divergence의 몇 가지 특징

- $KL(p|q) \neq KL(q|p)$

- $KL(p|q) \geq\ 0\quad,\ \ (p = q$ 일때 0 만족)
    1. 로그 함수의 Convexity 이용하여 증명 가능.

Cross-Entropy

$$\begin{aligned} \mathrm{KL}(p\|q) &= \boxed{-\int p(\mathbf{x}) \ln q(\mathbf{x})\, \mathrm{d}\mathbf{x}} - \left(-\int p(\mathbf{x}) \ln p(\mathbf{x})\, \mathrm{d}\mathbf{x}\right) \\ &= -\int p(\mathbf{x}) \ln \left\{ \frac{q(\mathbf{x})}{p(\mathbf{x})} \right\} \mathrm{d}\mathbf{x}. \end{aligned}$$

Cross Entropy

- $KL(p|q) = H(p,q) - H(p)$

Classification 할때 왜 Loss 함수는 Cross Entropy 일까?

- p(모분포, 정답)를 근사하기 위해 q(뉴럴넷)를 만들었음.
- H(p)는 q와 무관함. 즉, q의 parameter로 미분하면 사라짐.
- 그래서 H(p,q)를 loss 함수로 씀. -> KL 을 쓰는 것과 마찬가지.

Mutual Information

- x와 y가 Independent 면 p(x,y) = p(x) * p(y)

- 만일 Independent가 아니라면, KL Divergence를 이용하여 p(x)*p(y) 가 p(x,y)에 얼마나 가까운에 대한 idea를 얻을 수 있다.

$$
\begin{aligned}
\mathrm{I}[\mathbf{x}, \mathbf{y}] & \equiv \mathrm{KL}(p(\mathbf{x}, \mathbf{y}) \| p(\mathbf{x}) p(\mathbf{y})) \\
& = -\iint p(\mathbf{x}, \mathbf{y}) \ln \left(\frac{p(\mathbf{x}) p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})}\right) \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y}
\end{aligned}
$$

- $I(x, y) \geq 0$, x, y가 독립일 때 등호 성립