James Schneeloch

Lecture for physics 407
Thursday 12-5-2013
Cover & Thomas Ch. 2
Sections 1-6

Introduction to Classical Information Theory

- It is a mathematical theory of Communication

Shannon, 1948 "A mathematical theory of Communication"
$\longrightarrow$ over 65,000 citations

(EPR paradox paper has ~11,000 citations)
(Bell inequality paper has ~7900 citations)

Uses:

$\rightarrow$ Understanding limits of communicating data
$\rightarrow$ data compression
$\rightarrow$ statistical inference
$\rightarrow$ cryptography

How to quantify information?

- One symbol can have different meanings
- We quantify information in the context of communication (in bits)

★ We can't say how many bits a symbol has; we can say how many bits it takes to convey that symbol to others

What is a bit?

- The information conveyed in the answer of a yes/no question.

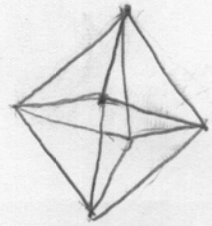
(20 questions $\rightarrow$ 20 bits)

- Bits are written as...

(on or off) (0 or 1) (bright or dark)
(red or green) (horizontal or vertical) etc...

ex

Let's say you have an 8-sided Loaded die.



Random Variable $\underline{X} = \{X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8\}$

: X_i = "lands on side (i)" | $P(X_i)$ = Probability that it lands on side (i)

$$P(\underline{X}) = \left\{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}, \frac{1}{64}, \frac{1}{128}, \frac{1}{128}\right\}$$

- Someone rolls the die and hides the outcome.
  - You want to figure out the outome
  - You may ask only yes/no questions.

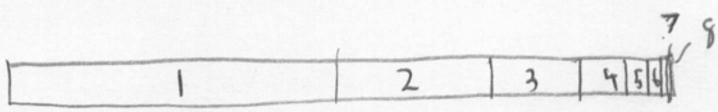
① How many questions (bits) does it take to be sure of the outcome of this roll?

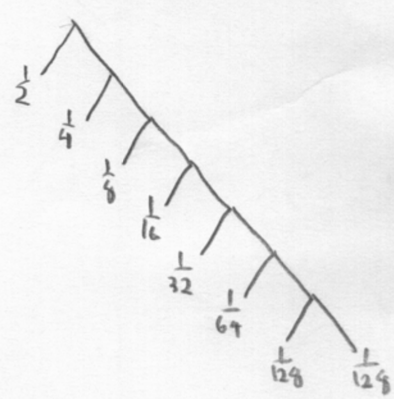
1	2	3	4
5	6	7	8

$3 = \log_2(8)$

3 bits are needed

② How many questions (bits) does it take on average per roll over many rolls?





$$\langle \text{\# of questions} \rangle = \frac{1}{2}(1) + \frac{1}{4}(2) + \frac{1}{8}(3) + \frac{1}{16}(4) + \frac{1}{32}(5) + \frac{1}{64}(6) + \frac{2}{128}(7) = \frac{127}{64} \sim 1.98 \text{ bits}$$

(less than 3)

* Tangent:

- How many bits does it take to be sure of the thermodynamic microstate of a liter of water at room temperature?

$$S = k_B \ln(\Omega)$$

$$\# \text{ of bits} = \log_2(\Omega) = \frac{\ln(\Omega)}{\ln(2)} = \frac{S}{k_B \ln(2)}$$

$$S_{H_2O} \sim 6\, J/_{mol \cdot K} \text{ at } 20\text{-}25^\circ C$$

There are about 55 mol of H_2O in 1 liter of it.

so $\# \text{ of bits} \sim \underline{3.5 \times 10^{25} \text{ bits}}$

- The highest information transfer rate over an optical fiber is currently about 10^{15} bits/s (with a 12-core fiber)
  - At this rate, how long would it take to transfer the infomation of the microstate of that liter of water?

$$\text{Total time} = \frac{3.5 \times 10^{25} \text{ bits}}{10^{15} \text{ bits}/s} \sim 3.5 \times 10^{10} \text{ seconds}$$

or ~ <u>1100 years</u>

(1 billion seconds is ~32 years)

Entropy

The Shannon entropy (in bits)	$H_2(\underline{X}) \equiv -\sum_{x_i \in \underline{X}} P(x_i) \log_2(P(x_i))$

convention: $0 \log_b 0 \equiv 0$ since $\lim_{z \to 0} z \log_b z = 0$

$H_2(\underline{X}) \to$ The minimum average number of bits needed to communicate the outcome of $\underline{X}$.

$\to$ A measure of the inherent uncertainty in the outcome of $\underline{X}$.

Entropy can be measured in different bases

$b=2$ "bits"
$b=3$ "trits"
$b=e$ "nats"

$$H_b(\underline{X}) = -\sum_{x_i \in \underline{X}} P(x_i) \log_b(P(x_i))$$

Note: A "trit" is the amount of information conveyed by answering a 3-answer question
(more, less, same)
(here, there, neither)

[Let's just use bits $\to$ $H(\underline{X}) \Rightarrow H_2(\underline{X})$ (by convention)]

(Useful form) $$H(\underline{X}) = \left\langle \log\left(\frac{1}{P(\underline{X})}\right)\right\rangle_{P(\underline{X})}$$

Elementary properties

① $H(\underline{X}) \geq 0$.

$P(x_i) \in [0,1] \Rightarrow \log_b\left(\frac{1}{P(x_i)}\right) \geq 0$, so $H(\underline{X}) \geq 0$

② $H_b(\underline{X}) = (\log_b a) H_a(\underline{X})$ (change of base formula for logarithms)

$\log_b P = \log_b a \log_a P$

eg. Biased coin toss

$$\underline{X} = \{X_1, X_2\} \qquad X_1 = \text{"heads"} \quad X_2 = \text{"tails"}$$

$$P(\underline{X}) = \{p, 1-p\}$$

what is $H(\underline{X})$?

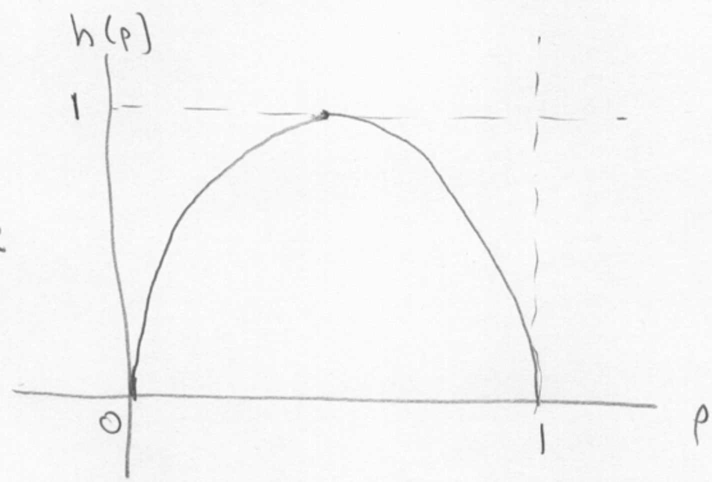
$$H(\underline{X}) = -\sum_{x_i \in \underline{X}} P(X_i) \log(P(X_i))$$

$$\underline{H(\underline{X}) = -p\log_2(p) - (1-p)\log_2(1-p)} \equiv \boxed{\text{the binary entropy function} \quad h(p)}$$

As $p \to 0$ or $p \to 1$

$h(p) \to 0$

"A certain outcome requires no bits to communicate."



If $p = \frac{1}{2}$, we need on average 1 bit per coin toss.

---

eg. 8-sided loaded die

$$P(\underline{X}) = \left\{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}, \frac{1}{64}, \frac{1}{128}, \frac{1}{128}\right\}$$

what is $H(\underline{X})$?

$$H(\underline{X}) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{4}\log\left(\frac{1}{4}\right) - \cdots$$

$$= \frac{1}{2}\log(2) + \frac{1}{4}\log(4) + \cdots$$

$$= \frac{1}{2}(1) + \frac{1}{4}(2) + \frac{1}{8}(3) + \frac{1}{16}(4) + \cdots$$

$$= \frac{127}{64} \text{ bits} \quad \text{(just like before)}$$

Joint and Conditional Entropy

(Marginal Entropy) $H(\underline{X}) = \left\langle \log\left(\frac{1}{P(\underline{X})}\right) \right\rangle_{P(\underline{X})}$ "Average # of bits you need to communicate the outcome of $\underline{X}$"

Joint Entropy

$$H(\underline{X},\underline{Y}) \equiv - \sum_{X_i Y_j \in (\underline{X}\,\underline{Y})} P(X_i, Y_j) \log\left(P(X_i, Y_j)\right) = \left\langle \log\left(\frac{1}{P(\underline{X},\underline{Y})}\right) \right\rangle_{P(\underline{X},\underline{Y})}$$

$H(\underline{X},\underline{Y}) \longrightarrow$ "Average # of bits you need to communicate the outcomes of both $\underline{X}$ and $\underline{Y}$"

Note: If $\underline{X}$ and $\underline{Y}$ are independent, ~~th~~

then $H(\underline{X},\underline{Y}) = H(\underline{X}) + H(\underline{Y})$

Conditional Entropy

$$H(\underline{Y}|\underline{X}) \equiv - \sum_{X_i Y_j \in (\underline{X},\underline{Y})} P(X_i, Y_j) \log\left(P(Y_j|X_i)\right) = \left\langle \log\left(\frac{1}{P(\underline{Y}|\underline{X})}\right) \right\rangle_{P(\underline{X},\underline{Y})}$$

Note:

$$H(\underline{Y}|\underline{X}) = \sum_{X_i \in \underline{X}} P(X_i)\, H(\underline{Y}|\underline{X}=X_i)$$

where $H(\underline{Y}|\underline{X}=X_i) = - \sum_{Y_j \in \underline{Y}} P(Y_j|X_i) \log\left(P(Y_j|X_i)\right)$

Note: $H(\underline{X}|\underline{Y}) \neq H(\underline{Y}|\underline{X})$

Note: From Bayes' Rule

$$\boxed{H(\underline{X},\underline{Y}) = H(\underline{X}) + H(\underline{Y}|\underline{X})}$$

Chain rule of the joint entropy

because $\log\left(\frac{1}{P(A,B)}\right) = \log\left(\frac{1}{P(A)P(B|A)}\right) = \log\left(\frac{1}{P(A)}\right) + \log\left(\frac{1}{P(B|A)}\right)$

ex) Let $\underline{X}, \underline{Y}$ have the following joint distribution

$\underline{X} \backslash \underline{Y}$	a	b	c	d
a	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$
b	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$
c	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
d	$\frac{1}{4}$	0	0	0

marginals	a	b	c	d
$\underline{X}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
$\underline{Y}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$

What are $H(\underline{X})$, $H(\underline{Y})$, $H(\underline{X},\underline{Y})$, $H(\underline{X}|\underline{Y})$, and $H(\underline{Y}|\underline{X})$?

(skip in lecture)

$$H(\underline{X}) = -\sum_{x_i} P(x_i)\log(P(x_i)) = -\frac{1}{4}\log\left(\frac{1}{4}\right) - \frac{1}{4}\log\left(\frac{1}{4}\right) - \frac{1}{4}\log\left(\frac{1}{4}\right) - \frac{1}{4}\log\left(\frac{1}{4}\right)$$

$H(\underline{X}) = 2$ bits

$H(\underline{Y}) = \frac{7}{4}$ bits

$H(\underline{X},\underline{Y}) = \frac{27}{8}$ bits

$H(\underline{X}|\underline{Y}) = H(\underline{X},\underline{Y}) - H(\underline{X}) = \frac{11}{8}$ bits

$H(\underline{Y}|\underline{X}) = \frac{13}{8}$ bits

note:

$H(\underline{X}|\underline{Y}) \neq H(\underline{Y}|\underline{X})$

Mutual Information

(The Shannon mutual information)

$$\boxed{H(\underline{X}:\underline{Y}) \equiv \sum_{x_i y_j \in \underline{X}\,\underline{Y}} P(x_i, y_j)\log\left(\frac{P(x_i y_j)}{P(x_i)P(y_j)}\right)}$$

$$H(\underline{X}:\underline{Y}) = \left\langle \log\left(\frac{P(\underline{X},\underline{Y})}{P(\underline{X})P(\underline{Y})}\right)\right\rangle_{P(\underline{X},\underline{Y})}$$

$$H(\underline{X}:\underline{Y}) = \left\langle \log\left(\frac{1}{P(\underline{X})}\right) - \log\left(\frac{1}{P(\underline{X}|\underline{Y})}\right)\right\rangle_{P(\underline{X},\underline{Y})}$$

$$\boxed{H(\underline{X}:\underline{Y}) = H(\underline{X}) - H(\underline{X}|\underline{Y})}$$

$$= H(\underline{Y}) - H(\underline{Y}|\underline{X})$$

$\Rightarrow$ "average number of bits communicated about the outcome of $\underline{Y}$ by communicating the outcome of $\underline{X}$.

Note: From $H(\underline{X}|\underline{Y}) = H(\underline{X},\underline{Y}) - H(\underline{Y})$

$$\boxed{H(\underline{X}:\underline{Y}) = H(\underline{X}) + H(\underline{Y}) - H(\underline{X},\underline{Y})}$$

Elementary properties

1. $H(\underline{X}:\underline{Y}) = H(\underline{Y}:\underline{X})$
2. $H(\underline{X}:\underline{X}) = H(\underline{X})$, because $H(\underline{X},\underline{X}) = H(\underline{X})$

Conditional Mutual Information

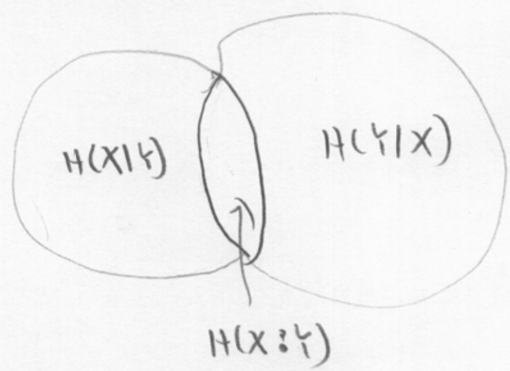
$$H(\underline{X}:\underline{Y}|\underline{Z}) \equiv \sum_{x_i y_j z_k \in \underline{X}\underline{Y}\underline{Z}} p(x_i, y_j, z_k) \log\left(\frac{p(x_i y_j | z_k)}{p(x_i|z_k)\, p(y_j|z_k)}\right)$$

$$H(\underline{X}:\underline{Y}|\underline{Z}) = \left\langle \log\left(\frac{p(\underline{X},\underline{Y}|\underline{Z})}{p(\underline{X}|\underline{Z})\, p(\underline{Y}|\underline{Z})}\right)\right\rangle_{p(\underline{X},\underline{Y},\underline{Z})}$$

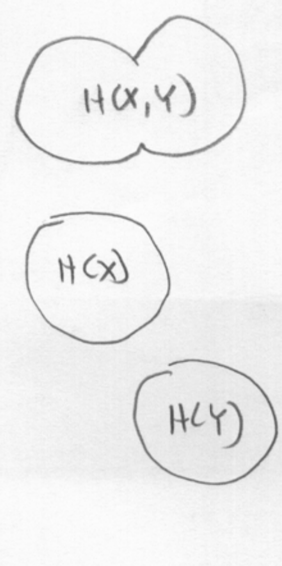
$$\boxed{H(\underline{X}:\underline{Y}|\underline{Z}) = H(\underline{X}|\underline{Z}) + H(\underline{Y}|\underline{Z}) - H(\underline{X},\underline{Y}|\underline{Z})}$$

"Average # of bits communicated about the outcome of $\underline{Y}$ by communicating the outcome of $\underline{X}$, given that the outcome of $\underline{Z}$ is known"

(Entropy Venn Diagram)?



"amount of uncertainty)



Relative Entropy (for any 2 distributions $P(\underline{X}), Q(\underline{X})$)

$$D(P(\underline{X}) \| Q(\underline{X})) \equiv \sum_{x_i \in \underline{X}} P(x_i) \log\left(\frac{P(x_i)}{Q(x_i)}\right)$$

(a.k.a. Kullback - Leibler divergence)

- It is a measure of divergence between two probability distributions
- It is a measure of inefficiency of coding the outcomes according to $Q(\underline{X})$, when true distribution is $P(\underline{X})$

Convention?

$0 \log\left(\frac{0}{0}\right) = 0$

$0 \log\left(\frac{0}{q}\right) = 0$

$p \log\left(\frac{p}{0}\right) \to \infty$

Vocab:

code: The system of assignment of code "words" to each outcome of $\underline{X}$.

code word: Sequence of (binary) digits assigned to particular outcome of $\underline{X}$.

$H(\underline{X})$ = Minimum average length of (binary) codeword to describe outcome of $\underline{X}$ with true distribution $P(\underline{X})$

☆ If we know the true distribution $P(\underline{X})$, we can ideally construct a code of average length $H(\underline{X})$

- If we constructed a code for $Q(\underline{X})$, when the distribution really was $P(\underline{X})$, our average codeword length would be longer by $D(P(\underline{X}) \| Q(\underline{X}))$

$D(P(\underline{X}) \| Q(\underline{X}))$ is **not** a distance measure

$$D(P(\underline{X}) \| Q(\underline{X})) \neq D(Q(\underline{X}) \| P(\underline{X}))$$

$$D(P(\underline{X}) \| Q(\underline{X})) + D(Q(\underline{X}) \| R(\underline{X})) \ngeq D(P(\underline{X}) \| R(\underline{X}))$$

Look up triangle inequality for distance metrics

$$\mathcal{D}(P(\underline{X}) \| Q(\underline{X})) = \left\langle \log\left(\frac{P(\underline{X})}{Q(\underline{X})}\right)\right\rangle_{P(\underline{X})}$$

Note:

$\left[\text{The mutual information is also a relative entropy}\right] \rightarrow H(\underline{X} : \underline{Y}) = \mathcal{D}(P(\underline{X},\underline{Y}) \| P(\underline{X})P(\underline{Y}))$

Conditional Relative Entropy

$$\boxed{\mathcal{D}(P(\underline{Y}|\underline{X}) \| Q(\underline{Y}|\underline{X})) \equiv \sum_{X_i Y_j \in (\underline{X}\,\underline{Y})} P(X_i\, Y_j) \log\left(\frac{P(Y_j | X_i)}{Q(Y_j | X_i)}\right)}$$

$$\mathcal{D}(P(\underline{Y}|\underline{X}) \| Q(\underline{Y}|\underline{X})) = \left\langle \log\left(\frac{P(\underline{Y}|\underline{X})}{Q(\underline{Y}|\underline{X})}\right)\right\rangle_{P(\underline{X},\underline{Y})}$$

Note: With Bayes' Rule

$$\left\langle \log\left(\frac{P(\underline{X},\underline{Y})}{Q(\underline{X},\underline{Y})}\right)\right\rangle_{P(\underline{X},\underline{Y})} = \left\langle \log\left(\frac{P(\underline{X})P(\underline{Y}|\underline{X})}{Q(\underline{X})Q(\underline{Y}|\underline{X})}\right)\right\rangle_{P(\underline{X},\underline{Y})}$$

$$= \left\langle \log\left(\frac{P(\underline{X})}{Q(\underline{X})}\right)\right\rangle_{P(\underline{X},\underline{Y})} + \left\langle \log\left(\frac{P(\underline{Y}|\underline{X})}{Q(\underline{Y}|\underline{X})}\right)\right\rangle_{P(\underline{X},\underline{Y})}$$

so that

$$\boxed{\mathcal{D}(P(\underline{X},\underline{Y}) \| Q(\underline{X},\underline{Y})) = \mathcal{D}(P(\underline{X}) \| Q(\underline{X})) + \mathcal{D}(P(\underline{Y}|\underline{X}) \| Q(\underline{Y}|\underline{X}))}$$

Chain Rule for Relative entropy

Jensen's Inequality and its Consequences

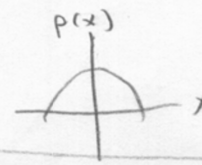
"the inequality about convex functions"

Definition:

- A function $f(x)$ is <u>convex</u> in the interval $x \in [a,b]$ if for any $(x_1, x_2) \in [a,b]$ and for any $\lambda \in [0,1]$

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$

$\rightarrow$ $f(x)$ is <u>strictly convex</u> if equality here implies $\lambda = 0$ or 1.



Note: If $\frac{\partial^2 f}{\partial x^2} \geq 0$ over $x \in [a,b]$

then $f(x)$ is convex over $x \in [a,b]$

Jensen's Inequality:

If $f(\underline{X})$ is a convex function of random variable $\underline{X}$

then $$\langle f(\underline{x}) \rangle_{p(\underline{x})} \geq f\left(\langle \underline{x} \rangle_{p(\underline{x})}\right)$$

ex/

$\langle x^2 \rangle \geq \langle x \rangle^2$ because $f(x) = x^2$ is a convex function

$\langle -\log(x) \rangle \leq -\log(\langle x \rangle)$ because $f(x) = -\log(x)$ is a concave function

Consequence:

For any two distributions $P(\underline{X})$ and $Q(\underline{X})$

$$\boxed{\begin{array}{l} D(P(\underline{X})\|Q(\underline{X})) \geq 0 \\ \text{Information inequality} \end{array}}$$

Proof:

$$-D(P(\underline{X})\|Q(\underline{X})) = -\sum_{x_i \in \underline{X}} P(x_i)\log\left(\frac{P(x_i)}{Q(x_i)}\right)$$

concave

$$= \sum_{x_i \in \underline{X}} P(x_i)\log\left(\frac{Q(x_i)}{P(x_i)}\right)$$

with Jensen's inequality

$$\leq \log\left(\sum_{x_i \in \underline{X}} \cancel{P(x_i)}\frac{Q(x_i)}{\cancel{P(x_i)}}\right) = \log\left(\sum_{x_i \in \underline{X}} Q(x_i)\right)$$

$$= \log(1) = 0$$

$$-D(P(\underline{X})\|Q(\underline{X})) \leq 0$$

so $\underline{D(P(\underline{X})\|Q(\underline{X})) \geq 0}$

Similarly...

$$\boxed{D(P(\underline{Y}|\underline{X})\|Q(\underline{Y}|\underline{X})) \geq 0}$$

Consequenses of Information Inequality

$$\boxed{\mathcal{D}(P(\underline{X}) \| Q(\underline{X})) \geq 0} \to \boxed{H(\underline{X}) \leq \log(N)}$$

since

$\log(N) + H(\underline{X}) = \mathcal{D}(P(\underline{X}) \| U(\underline{X}))$

: $U(\underline{X})$ = uniform distribution

$$\downarrow$$

$$\boxed{H(\underline{X} : \underline{Y}) \geq 0}$$

$\to H(\underline{X} | \underline{Y}) \leq H(\underline{X})$

$\to H(\underline{X}, \underline{Y}) \leq H(\underline{X}) + H(\underline{Y})$

$$\mathcal{D}(P(\underline{Y}|\underline{X}) \| Q(\underline{Y}|\underline{X})) \geq 0 \to H(\underline{Y}|\underline{X}) \leq \log(N)$$

$$\downarrow$$

$$H(\underline{X} : \underline{Y} | \underline{Z}) \geq 0$$

$\to H(\underline{X} | \underline{Y}\,\underline{Z}) \leq H(\underline{X}, \underline{Y})$

$\to H(\underline{X}, \underline{Y} | \underline{Z}) \leq H(\underline{X}|\underline{Z}) + H(\underline{Y}|\underline{Z})$

Other useful

Chain Rules

$$H(\underline{W}\,\underline{X}\,\underline{Y}\,\underline{Z}) = H(\underline{W}) + H(\underline{X}|\underline{W}) + H(\underline{Y}|\underline{W}\,\underline{X}) + H(\underline{Z}|\underline{W}\,\underline{X}\,\underline{Y})$$

$$H(\underline{W} : \underline{X}\,\underline{Y}\,\underline{Z}) = H(\underline{W} : \underline{X}) + H(\underline{W} : \underline{Y}|\underline{X}) + H(\underline{W} : \underline{Z}|\underline{X}\,\underline{Y})$$