1.

RWTH AACHEN UNIVERSITY

G. Alirezaei

$$X \quad \Rightarrow \quad H(X) = -\sum_{i=1}^{m} p_i \cdot \log p_i$$

$$\mathrm{Prob}(X = x_i) = p_i \quad , \quad i = 1 \dots m$$

Theorem 2.1.8.

a) $0 \overset{(i)}{\leq} H(X) \overset{(ii)}{\leq} \log m$

"=" in (i) $\Leftrightarrow$ X has a singleton dist., i.e.,
$\exists x_i : \mathrm{Prob}(X = x_i) = 1$

"=" in (ii) $\Leftrightarrow$ X is uniformly dist., i.e.,

$\mathrm{Prob}(X = x_i) = \frac{1}{m} \quad \forall i = 1, \dots, m$

b) $0 \overset{(i)}{\leq} H(X|Y) \overset{(ii)}{\leq} H(X)$

"=" in (i) $\Leftrightarrow$ $\mathrm{Prob}(X = x_i \mid Y = y_j) = 1 \quad \forall i,j$
with $\mathrm{Prob}(X = x_i, Y = y_j) > 0$, i.e.,
X is totally dependent on Y.

"=" (ii) $\Leftrightarrow$ X, Y are stoch. independent.

c) $H(X) \overset{(i)}{\leq} H(X, Y) \overset{(ii)}{\leq} H(X) + H(Y)$

"=" in (i) $\Leftrightarrow$ Y is totally dependent on X

"=" in (ii) $\Leftrightarrow$ X, Y are stoch. independent.

d) $H(X | Y, Z) \leq \min\{H(X|Y), H(X|Z)\}$ ┘

Proof:

a) (ii) $$H(X) = -\sum_{i=1}^{m} p_i \cdot \log p_i = \sum_{i=1}^{m} p_i \log \frac{1}{p_i}$$

$$\overset{\text{Lemma 2.1.6}}{\leq} \log \left(\sum_{i=1}^{m} p_i \cdot \frac{1}{p_i} \right)$$

$$= \log \left(\sum_{i=1}^{m} 1 \right) = \log m$$

(i) obvious

b) Exercise

c) (i) By the chain rule, Th. 2.1.5.:

$$H(X,Y) = H(X) + \underbrace{H(Y|X)}_{\geq 0} \geq H(X)$$

equality follows from b) (ii).

(ii) $$0 \overset{(b)}{\leq} H(X) - H(X|Y) \overset{\text{Th. 2.1.5}}{=} H(X) - [H(X,Y) - H(Y)]$$

$$\Rightarrow \quad H(X) + H(Y) \geq H(X,Y)$$

with equality from b)(i).

d) Analogous to b)(ii). ■.

Ref. Cover & Thomas, Mathar.

RWTH AACHEN UNIVERSITY

Definition 2.1.9.

Let X, Y, Z discrete r.v.

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

is called mutual information or synentropy of X and Y.

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

is called conditional mutual information of X and Y given Z.

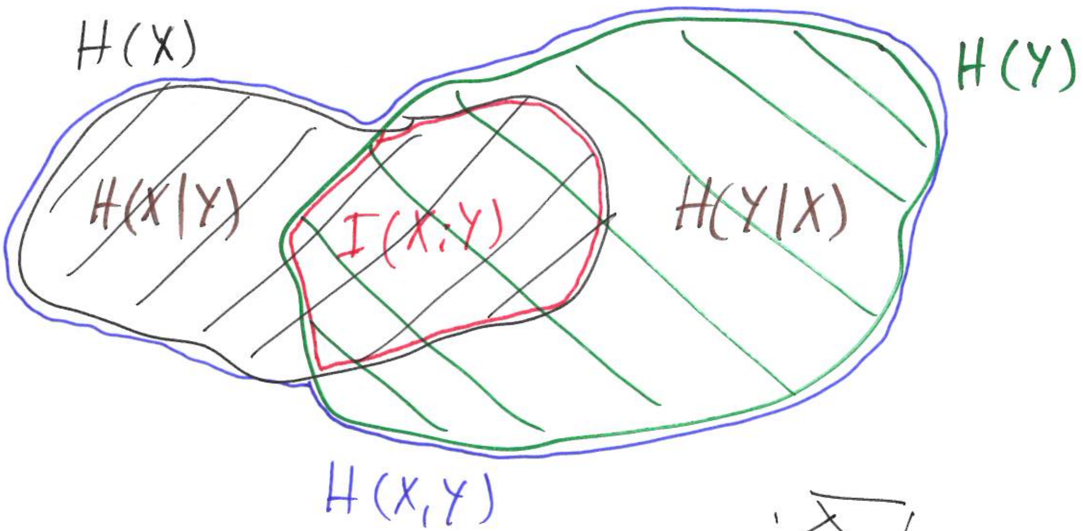
Interpretation: $I(X;Y)$ is the reduction in uncertainty about X when Y is given or the amount of information about X provided by Y.

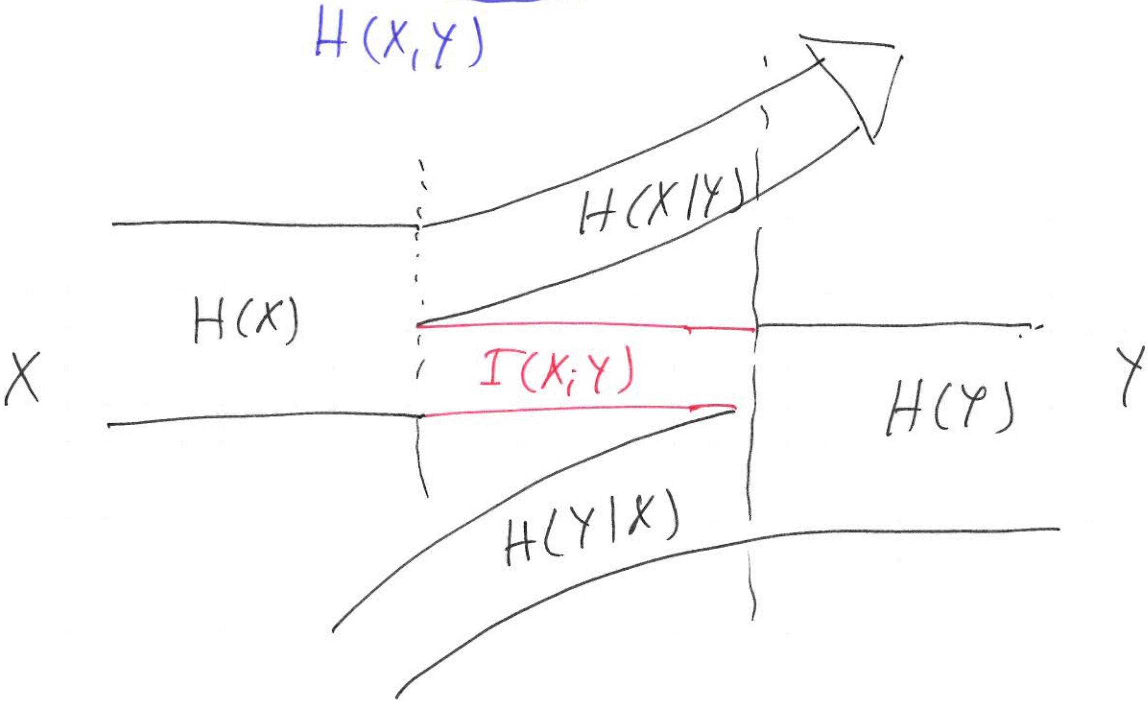
see. Figures on the next page.

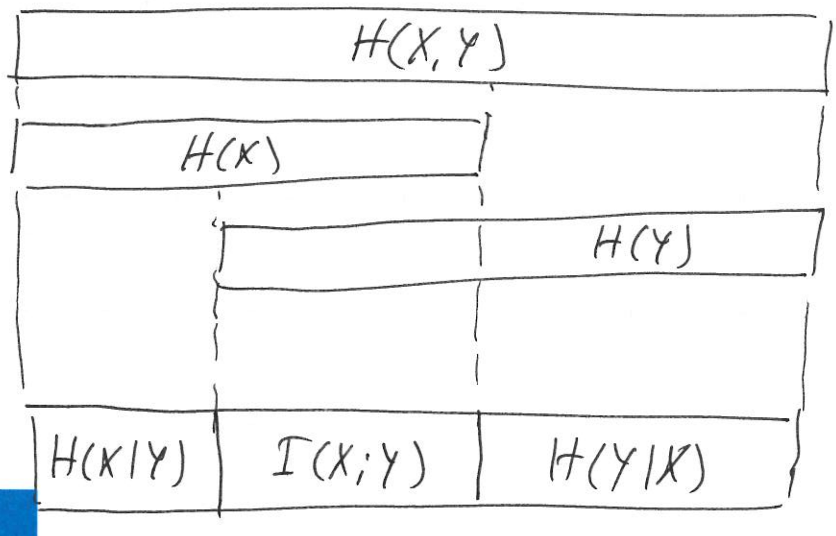
$$I(X;Y) = H(X) - H(X|Y)$$

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

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$


H(X)
H(Y)
H(X|Y)
I(X;Y)
H(Y|X)
H(X,Y)


H(X|Y)
H(X)
X
I(X;Y)
H(Y)
Y
H(Y|X)


H(X,Y)
H(X)
H(Y)
H(X|Y)
I(X;Y)
H(Y|X)

Note: by Thm 2.1.8 b) $I(X;Y) \geq 0$.

By definition it holds that:

$$I(X;Y) = \underbrace{-\sum_i p(x_i) \cdot \log p(x_i)}_{= H(X)} + \sum_{i,j} p(x_i, y_j) \cdot \log p(x_i \mid y_j)$$

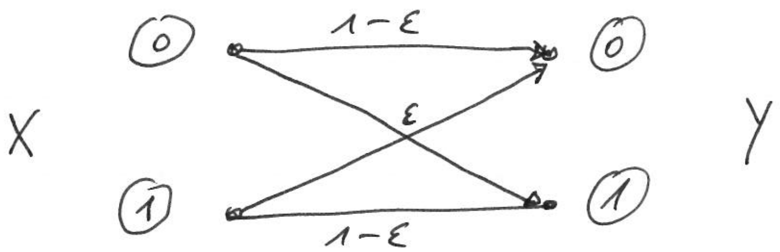
$$= -\sum_{i,j} p(x_i, y_j) \cdot \log p(x_i) + \sum_{i,j} p(x_i, y_j) \cdot \log p(x_i \mid y_j)$$

$$= \sum_{i,j} p(x_i, y_j) \cdot \log \frac{p(x_i \mid y_j)}{p(x_i)}$$

$$= \sum_{i,j} p(x_i, y_j) \cdot \log \frac{p(x_i, y_j)}{p(x_i) \cdot p(y_j)} \qquad (*)$$

which shows symmetry in X and Y.

Example 2.1.10. Binary symmetric channel (BSC)



symbol error with probability ε, $0 \leq \varepsilon \leq 1$.

Hence:

$$P(Y=0 \mid X=0) = P(Y=1 \mid X=1) = 1-\varepsilon$$

$$P(Y=0 \mid X=1) = P(Y=1 \mid X=0) = \varepsilon$$

Assume that input symbols are uniformly distributed: $P(X=0) = P(X=1) = 1/2$.

Then for the joint distributions:

$$P(X=0, Y=0) = P(Y=0 \mid X=0) \cdot P(X=0) = (1-\varepsilon) \cdot \frac{1}{2}$$

$\vdots$

X \ Y	0	1	
0	$\frac{1}{2}(1-\varepsilon)$	$\frac{\varepsilon}{2}$	$\frac{1}{2}$
1	$\frac{\varepsilon}{2}$	$\frac{1}{2}(1-\varepsilon)$	$\frac{1}{2}$
	$1/2$	$1/2$	

Further: $P(X=0 \mid Y=0) = \dfrac{P(X=0, Y=0)}{P(Y=0)} = 1-\varepsilon$

$$P(X=1 \mid Y=1) = 1-\varepsilon$$

$$P(X=0 \mid Y=1) = P(X=1 \mid Y=0) = \varepsilon$$

RWTH AACHEN UNIVERSITY

For $\log = \log_2$

$$\Rightarrow \quad H(X) = H(Y) = -\tfrac{1}{2}\cdot\log\tfrac{1}{2} - \tfrac{1}{2}\cdot\log\tfrac{1}{2} = 1 \text{ bit.}$$

$$H(X,Y) = 1 - (1-\varepsilon)\cdot\log(1-\varepsilon) - \varepsilon\cdot\log\varepsilon$$

$$H(X|Y) = -(1-\varepsilon)\cdot\log(1-\varepsilon) - \varepsilon\cdot\log\varepsilon$$

$$0 \le I(X;Y) = 1 + (1-\varepsilon)\cdot\log(1-\varepsilon) + \varepsilon\cdot\log\varepsilon \le 1$$

Def. 2.1.11 (Kullback-Leibler divergence)

Let $p = (p_1, \dots, p_m)$, $q = (q_1, \dots, q_m)$ be stoch. vectors.

$$D(p \| q) = \sum_{i=1}^{m} p_i \cdot \log \frac{p_i}{q_i}$$

is called KL divergence between p and q (or relative entropy.).

$D(p\|q)$ measures the divergence (distance, dissimilarity) between distributions p and q. However, it is not a metric, neither symmetric nor satisfies the triangle inequality.

It measures how difficult it is for p to pretend it were q.

Thinking the Future
Zukunft denken

RWTH AACHEN UNIVERSITY

Theorem 2.1.12.

a) $D(p\|q) \geq 0$ with "=" iff $p=q$.

b) $D(p\|q)$ is convex in the pair (p,q).

c) $I(X;Y) = D\Big(\big(p(x_i, y_j)\big)_{i,j} \,\|\, \big(p(x_i)\cdot p(y_j)\big)_{i,j}\Big)$ ⌟

proof.

a) By definition and Cor. 2.1.8.

b) use the log-sum inequality Lemma 2.1.7.
Let p, r, and q, s be stoch. vectors.
For all $i = 1, \ldots, m$ it holds that

$$(\lambda p_i + (1-\lambda)\cdot r_i)\cdot \log \frac{\lambda p_i + (1-\lambda)\cdot r_i}{\lambda q_i + (1-\lambda)\cdot s_i}$$

$$\leq \lambda p_i \log \frac{\lambda p_i}{\lambda q_i} + (1-\lambda)\cdot r_i \log \frac{(1-\lambda)\cdot r_i}{(1-\lambda)\cdot s_i}$$

summing over all $i = 1, \ldots, m$, it follows $\forall \lambda \in [0,1]$

$$D(\lambda p + (1-\lambda) r \,\|\, \lambda q + (1-\lambda) s) \leq \lambda D(p\|q) + (1-\lambda)\cdot D(r\|s)$$

c) By definition. ∎

Note: $D(p\|q) \neq D(q\|p)$

Thinking the Future
Zukunft denken