07

Dependent Random Variables

Notice

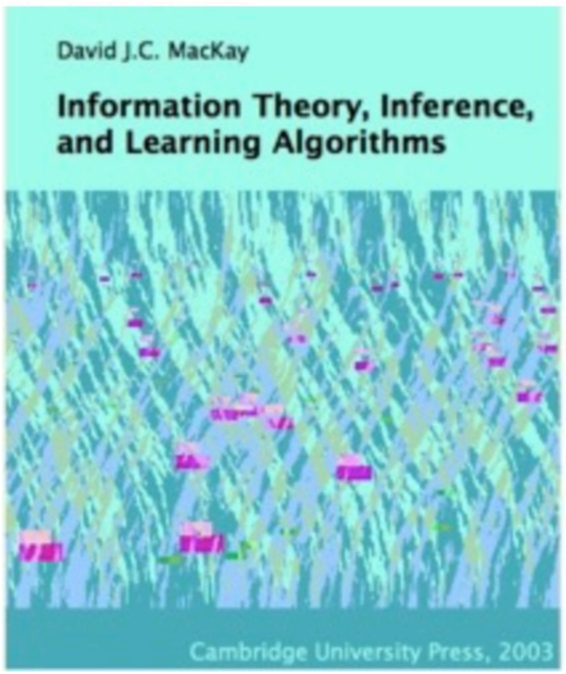
- **Author**
  - **João Moura Pires (jmp@fct.unl.pt)**

- **This material can be freely used for personal or academic purposes without any previous authorization from the author, provided that this notice is maintained/kept.**

- **For commercial purposes the use of any part of this material requires the previous authorization from the author.**

Bibliography

- **Many examples are extracted and adapted from:**



Information Theory, Inference, and Learning Algorithms
David J.C. MacKay
2005, Version 7.2

- **And some slides were based on Iain Murray course**
  - [**http://www.inf.ed.ac.uk/teaching/courses/it/2014/**](http://www.inf.ed.ac.uk/teaching/courses/it/2014/)

Table of Contents

- **More about entropy**
- **Mutual Information**
- **Mutual Information, Entropy and a Channel**
- **Few Demos**

More about entropy

The Joint Entropy

- The **joint entropy** of X, Y is

$$H(X,Y)=\sum_{xy\in A_x A_y} P(x,y)\log\frac{1}{P(x,y)}$$

- Entropy is **additive** for **independent** random variables

$$H(X,Y)=H(X)+H(Y) \quad iff \quad P(x,y)=P(x)P(y)$$

- The **marginal entropy of *X*** is another name for the entropy of X, $H(X)$

Conditional Entropy

- The **conditional entropy of *X* given $y = b_k$**

  is the entropy of the **probability distribution $P(x \mid y = b_k)$**.

$$H(X \mid y = b_k) = \sum_{x \in A_x} P(x \mid y = b_k) \log \frac{1}{P(x \mid y = b_k)}$$

- For each value of b_k we have, in general, a different value of $H(X \mid y = b_k)$

- The **conditional entropy of *X* given *Y*** is the **average over *y*** of the conditional entropy of *X* given *y*

$$H(X \mid Y) = \sum_{y \in A_y} P(y) H(X \mid y)$$

Conditional Entropy of X given Y

- The **conditional entropy of X given Y** is the **average over y** of the conditional entropy of X given y

$$H(X|Y) = \sum_{y \in A_y} P(y) H(X|y)$$

$$H(X|Y) = \sum_{y \in A_y} P(y) \left[\sum_{x \in A_x} P(x|y) \log \frac{1}{P(x|y)} \right]$$

$$H(X|Y) = \sum_{xy \in A_x A_y} P(x,y) \log \frac{1}{P(x|y)}$$

- This **measures the average uncertainty** that **remains about x** when **y is known**.

Chain Rule for information content

- Chain rule for information content

$$P(x,y) = P(x)P(y\,|\,x)$$
$$= P(y)P(x\,|\,y)$$

$$\log\frac{1}{P(x,y)} = \log\frac{1}{P(x)} + \log\frac{1}{P(y\,|\,x)}$$

$$h(x,y) = h(x) + h(y\,|\,x)$$

- The **information content** of ***x* and *y*** is
  - the **information content** of ***x* +**
  - the **information content** of ***y* given *x***.

Chain Rule for Entropy

- The joint entropy, conditional entropy and marginal entropy are related by

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

- The uncertainty of X and Y is
  - the uncertainty of X plus
  - the uncertainty of Y given X.
- Or
  - the uncertainty of Y plus
  - the uncertainty of X given Y.

Mutual Information

Chain Rule for Entropy

- The **mutual information** between X and Y is

$$I(X;Y) = H(X) - H(X \mid Y) \qquad I(X;Y) = H(Y) - H(Y \mid X)$$

and satisfies $I(X;Y) = I(Y;X)$, and $I(X;Y) \geq 0$.

- It measures the **average reduction in uncertainty about *x*** that results from **learning the value of *y***;
- or, the **average amount of information that *y*** conveys **about *x***.

The conditional mutual information

- The **conditional mutual information between X and Y given $z = c_k$**
  - is the mutual information between the random variables X and Y in the joint ensemble

    $P(x, y \mid z = c_k)$,

$$I(X;Y \mid z = c_k) = H(X \mid z = c_k) - H(X \mid Y, z = c_k)$$

- The **conditional mutual information between X and Y given Z**
  - is the average over z of $I(X; Y \mid z)$

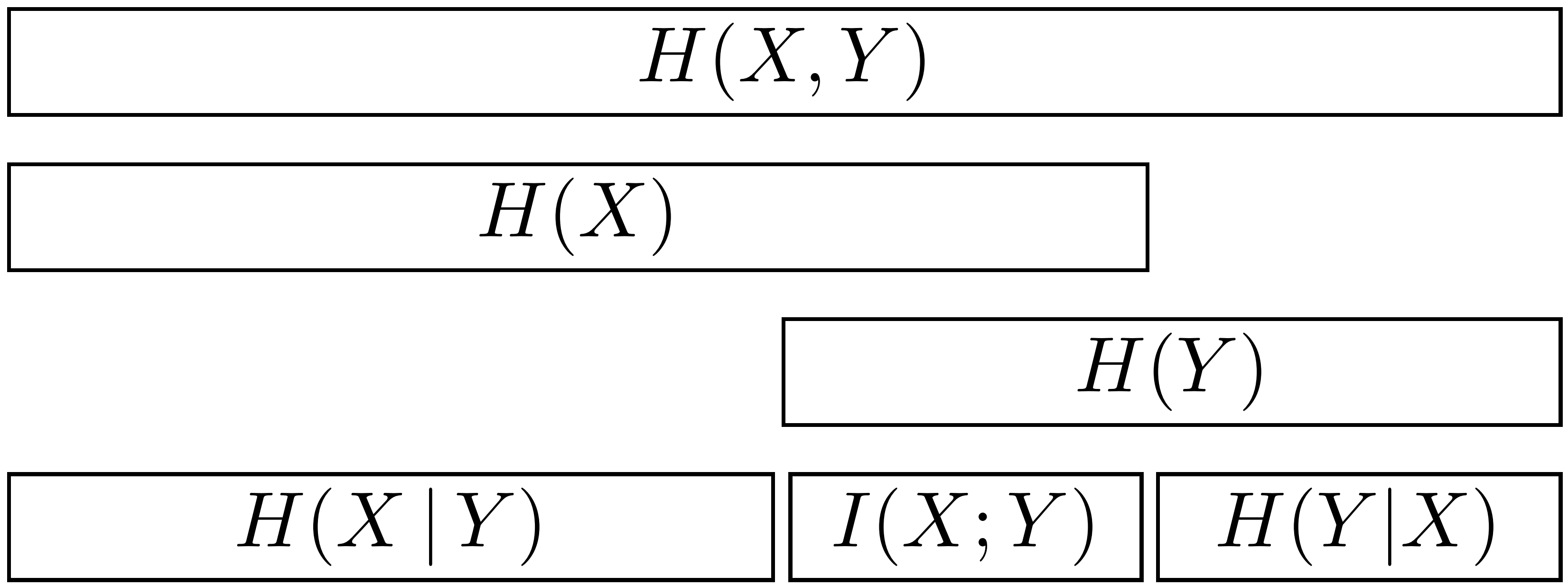
$$I(X;Y \mid Z) = H(X \mid Z) - H(X \mid Y, Z)$$

FACULDADE DE CIÊNCIAS E TECNOLOGIA UNIVERSIDADE NOVA DE LISBOA

No other ‘three-term entropies’ !

- No other ‘three-term entropies’ will be defined.
- For example:
  - expressions such as $I(X; Y ; Z)$ and $I(X \mid Y ; Z)$ **are illegal**.
- But you may put **conjunctions of arbitrary numbers of variables in each of the three spots** in the expression $I(X \,;\, Y \mid Z)$:
  - $I(A, B ; C, D \mid E , F)$ is fine !
  - it measures how much information on average c and d convey about a and b, assuming e and f are known.

Relationship between Mutual information and Entropies

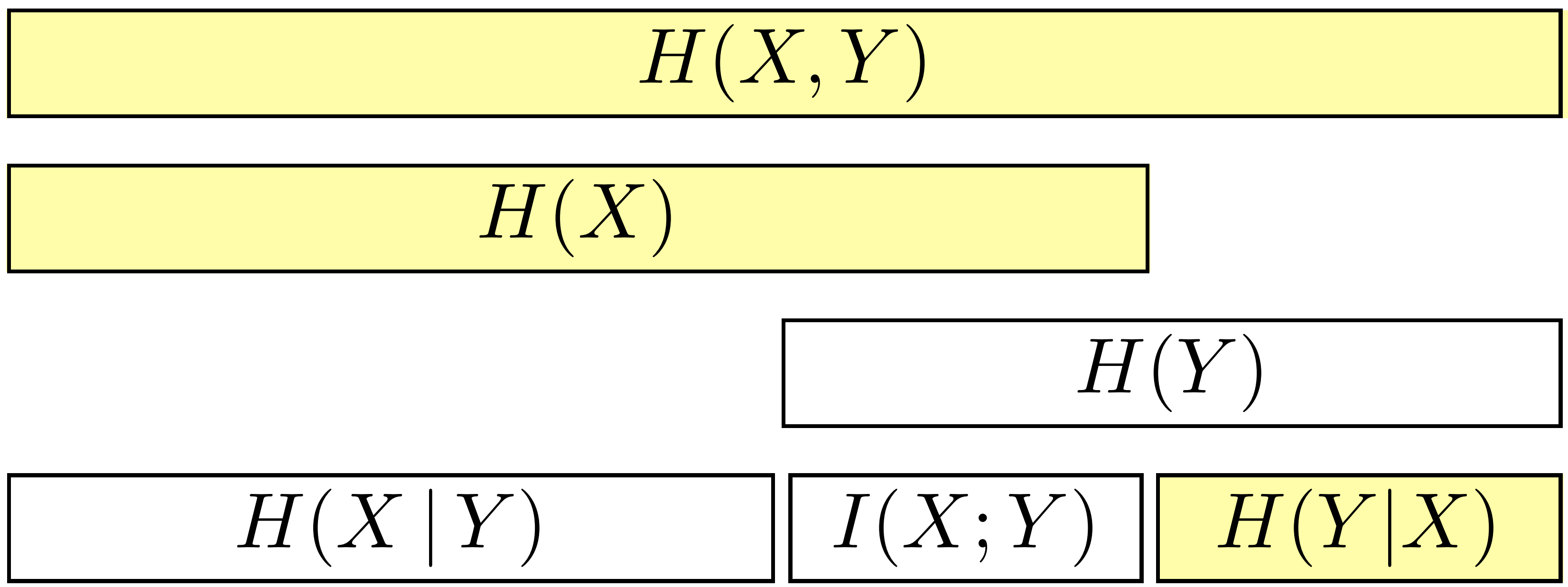


$$H(X,Y) = H(X) + H(Y|X)$$

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

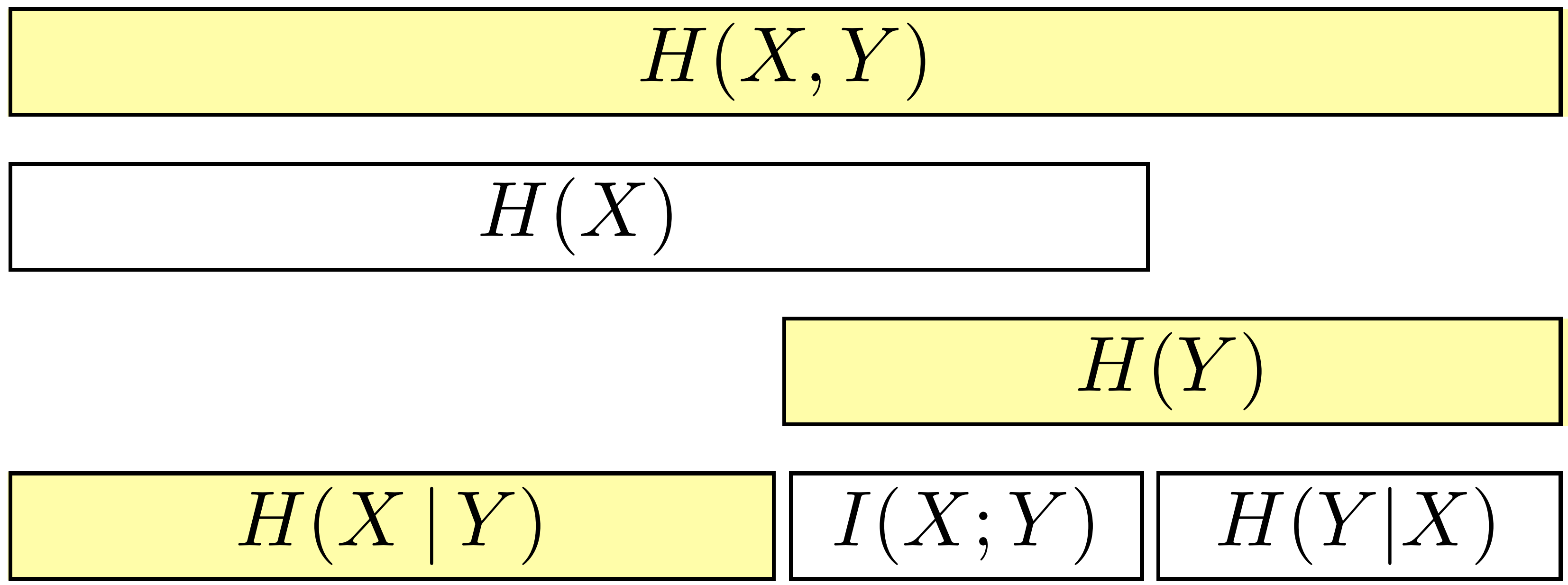
$$I(X;Y) = H(X) - H(X|Y)$$

Relationship between Mutual information and Entropies



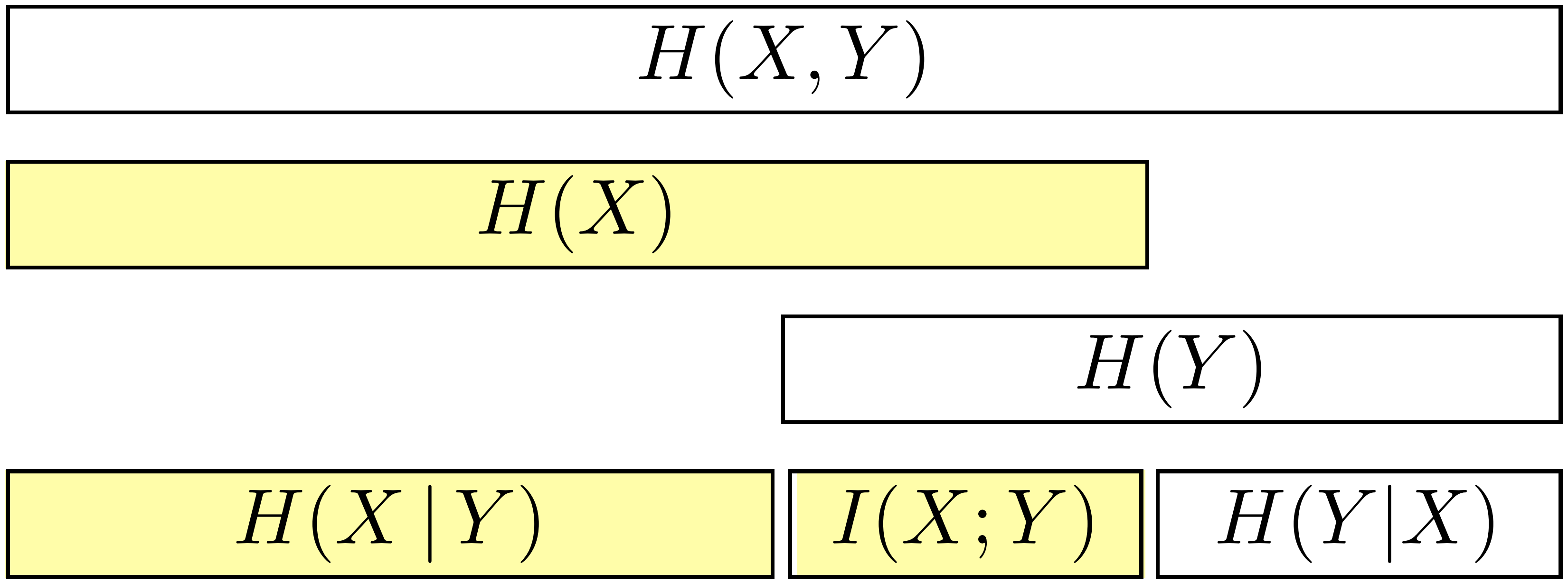
$$H(X,Y) = H(X) + H(Y \mid X)$$

Relationship between Mutual information and Entropies



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

Relationship between Mutual information and Entropies

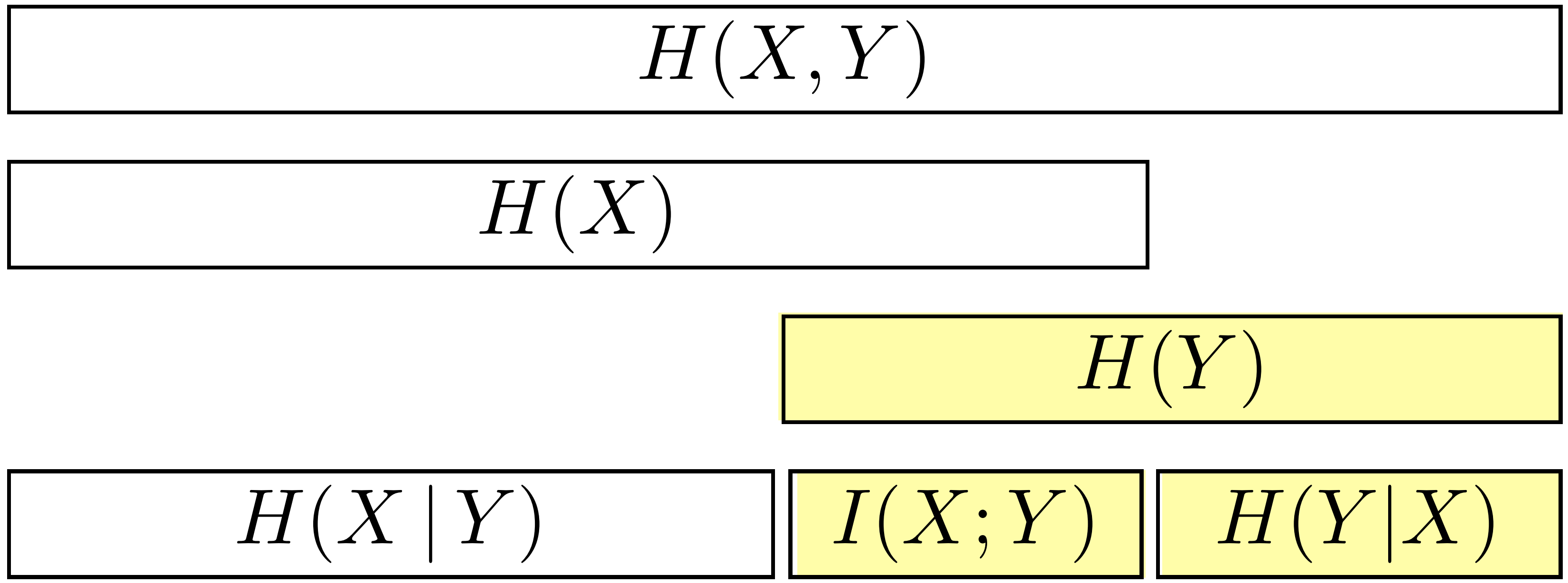


$$I(X;Y) = H(X) - H(X|Y)$$

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

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

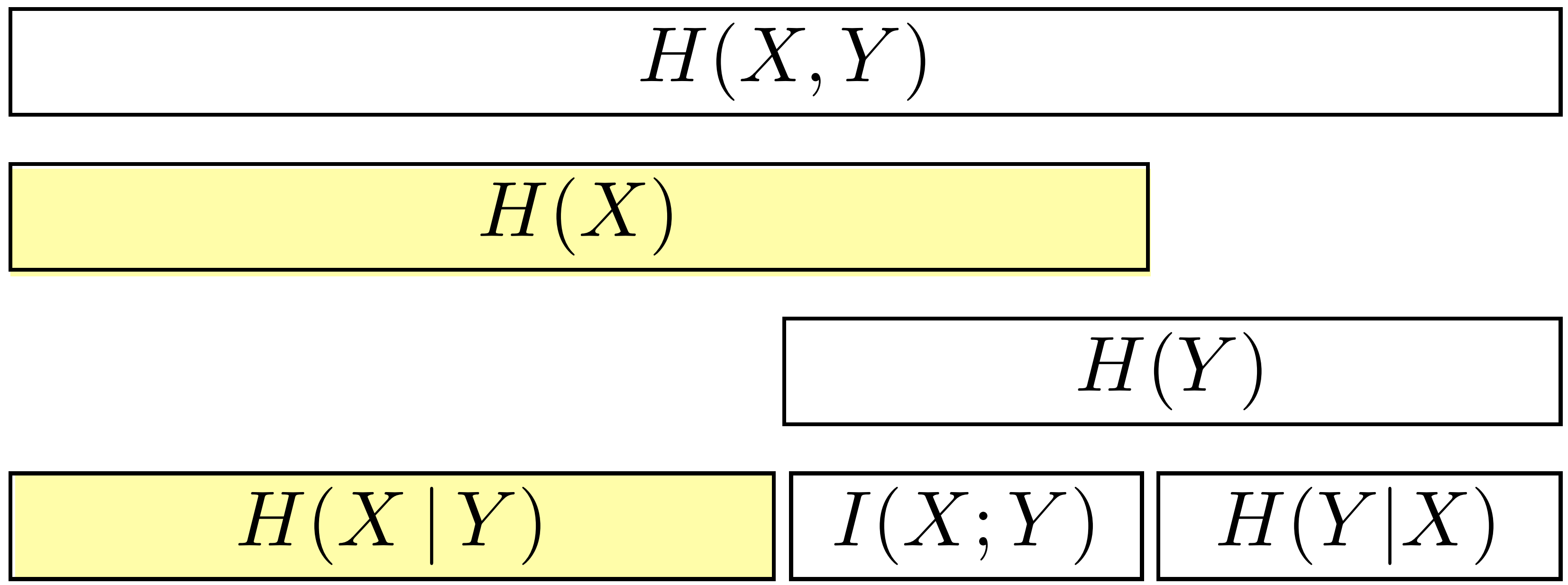
Relationship between Mutual information and Entropies



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

$$H(Y) = I(X;Y) + H(Y \mid X)$$

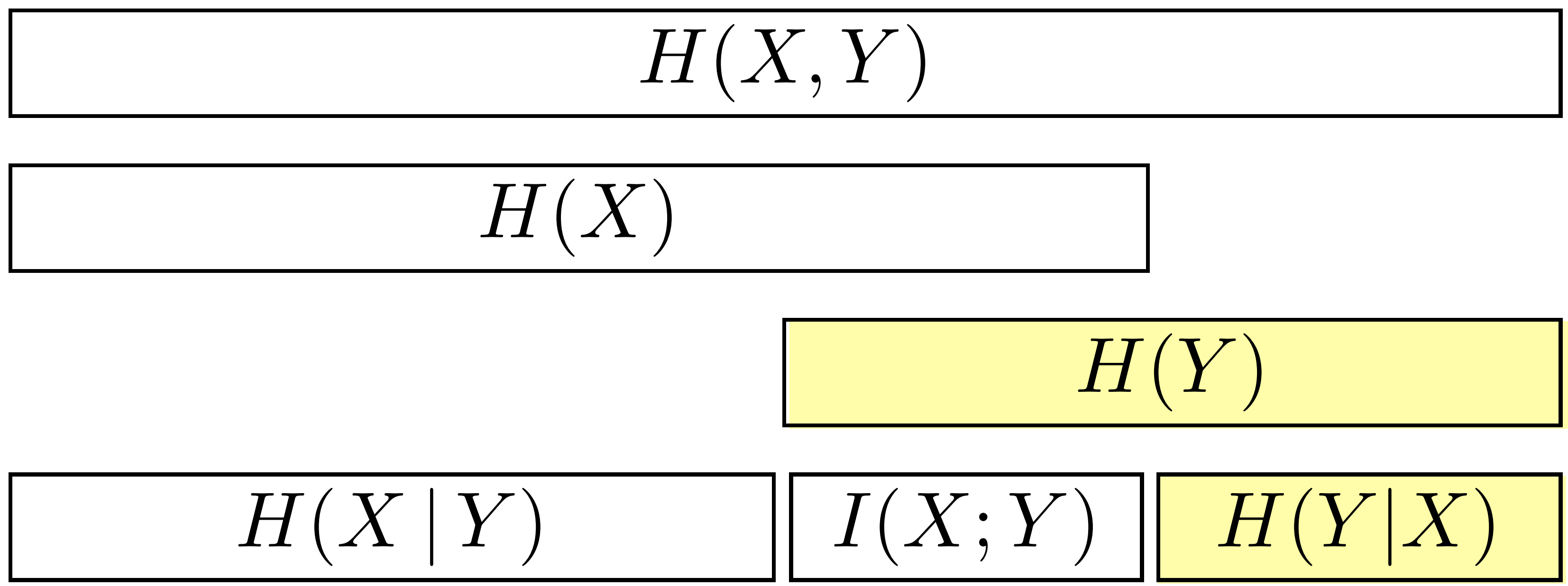
Relationship between Mutual information and Entropies



- This image also suggest that

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

Relationship between Mutual information and Entropies



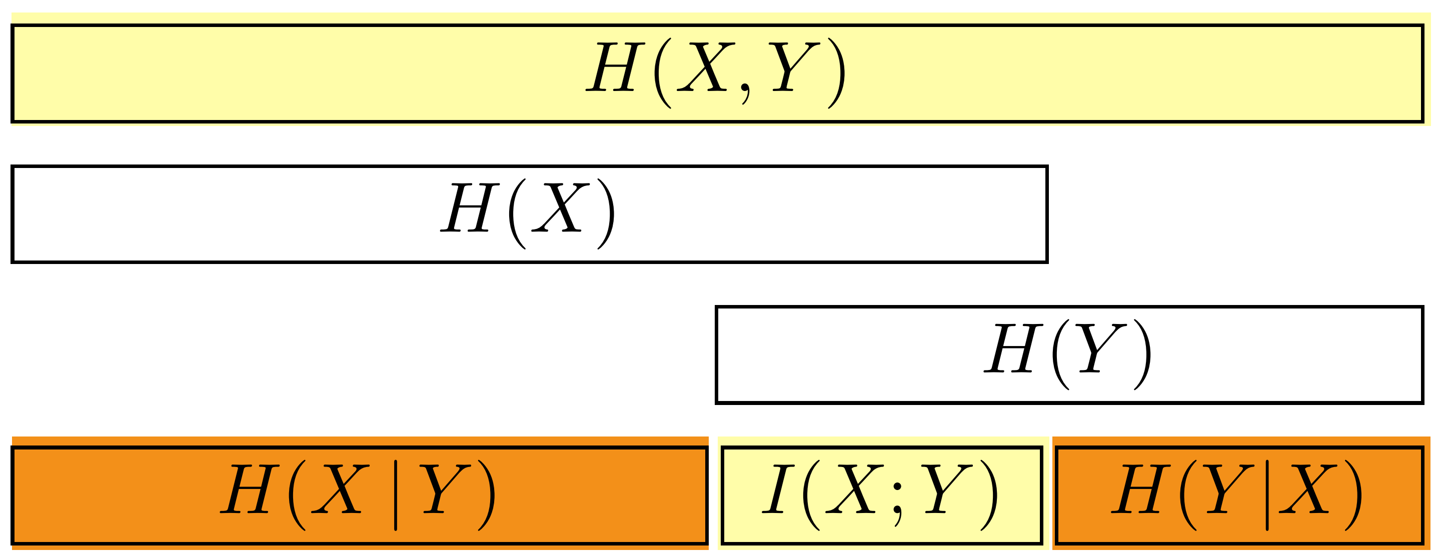
- This image also suggest that

$$H(Y) \geq H(Y \mid X)$$

Entropy Distance

- The '**entropy distance**' between two random variables can be defined to be the difference between their joint entropy and their mutual information:

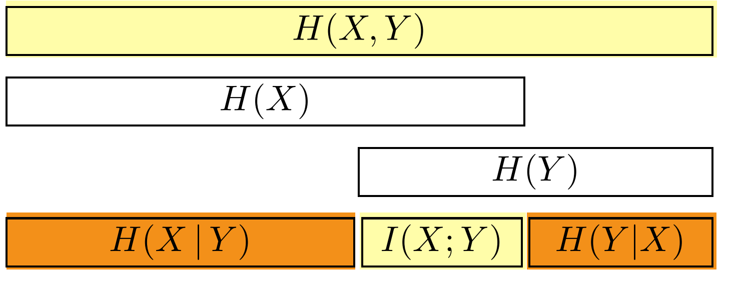
$$D_H(X,Y) = H(X,Y) - I(X;Y)$$



Entropy Distance

- The '**entropy distance**' between two random variables can be defined to be the difference between their joint entropy and their mutual information:

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



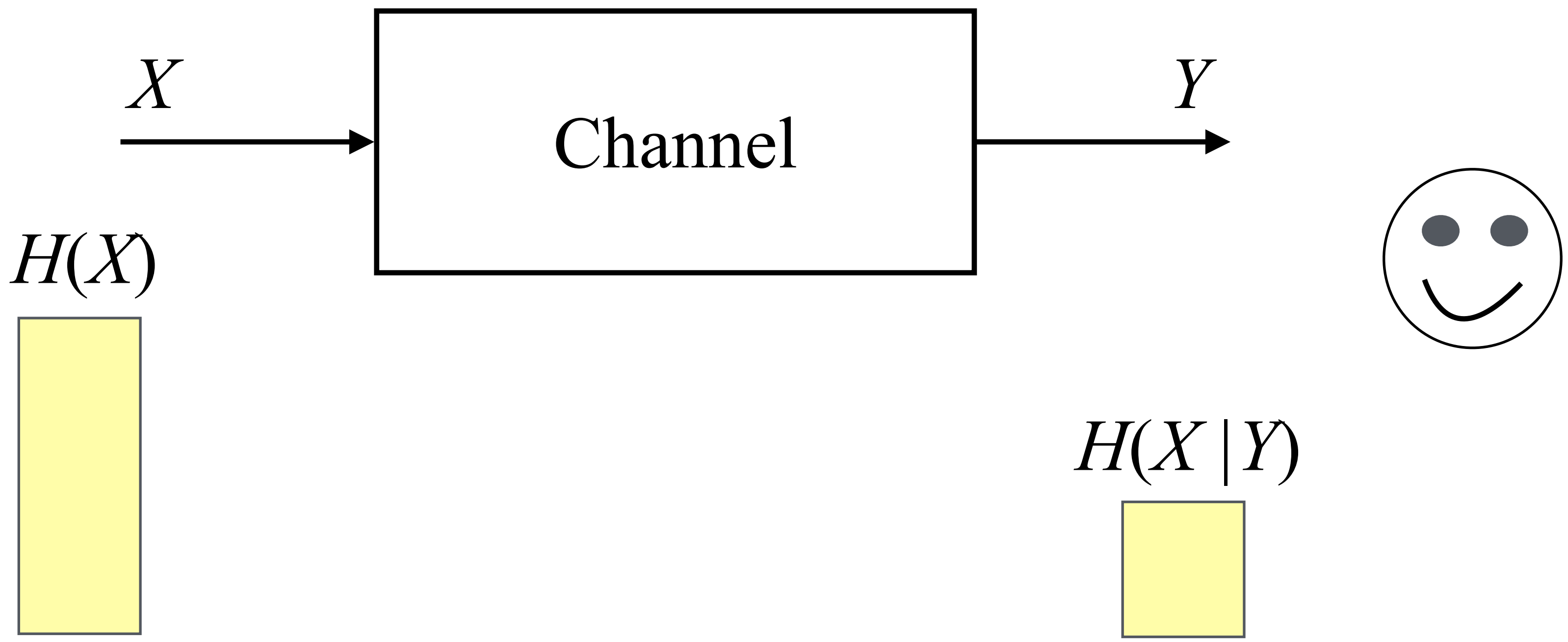
- Satisfies the following properties:
  - $D_H(X,Y) \geq 0$
  - $D_H(X,X) = 0$
  - $D_H(X,Y) = D_H(Y,X)$
  - $D_H(X,Z) \leq D_H(X,Y) + D_H(Y,Z)$

Axioms for a distance

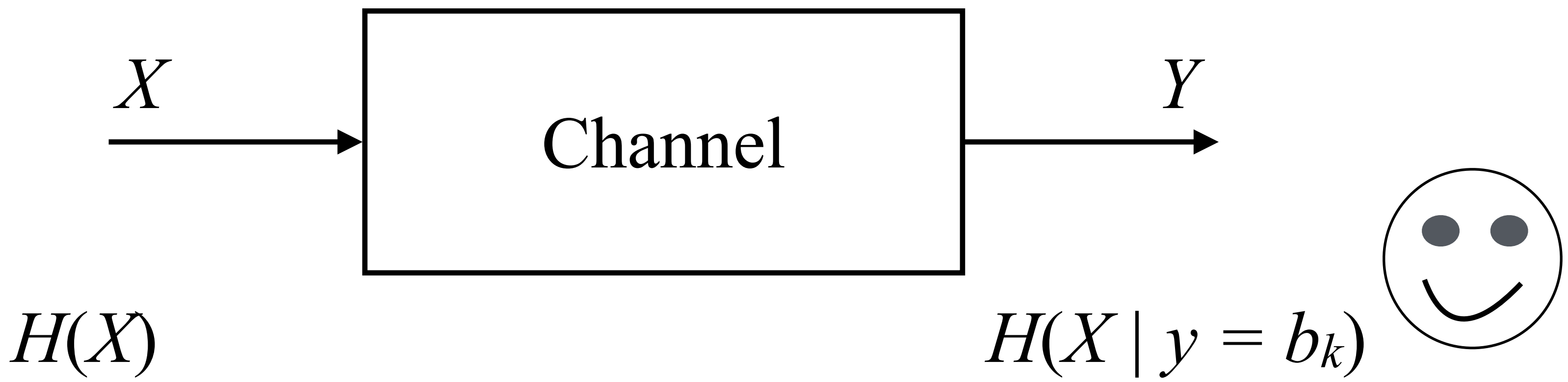
Mutual Information, Entropy and a Channel

Mutual Information, Entropy and a Channel

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



Mutual Information, Entropy and a Channel

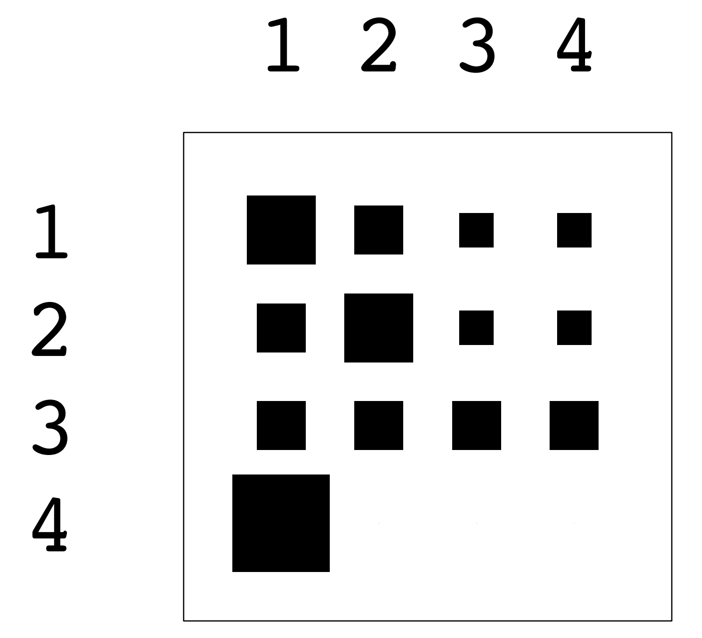


Notice that $\boldsymbol{H(X \mid y = b_k)}$ **may be larger or smaller than** $\boldsymbol{H(X)}$

An Example

- A joint ensemble XY has the following joint distribution

$P(x,y)$		$x=1$	$x=2$	$x=3$	$x=4$
y	1	$1/8$	$1/16$	$1/32$	$1/32$
	2	$1/16$	$1/8$	$1/32$	$1/32$
	3	$1/16$	$1/16$	$1/16$	$1/16$
	4	$1/4$	0	0	0



- Calculate $H(X)$, $H(Y)$
- Calculate $H(X \mid y)$ for all values of y,
- $H(X \mid Y)$ and $H(Y \mid X)$
- $I(X; Y)$

An Example

- Compute the marginal probabilities

$P(x,y)$		x 1	2	3	4	$P(y)$
y	1	$1/8$	$1/16$	$1/32$	$1/32$	$1/4$
	2	$1/16$	$1/8$	$1/32$	$1/32$	$1/4$
	3	$1/16$	$1/16$	$1/16$	$1/16$	$1/4$
	4	$1/4$	0	0	0	$1/4$
$P(x)$		$1/2$	$1/4$	$1/8$	$1/8$	

- Compute the Joint Entropy

$$H(X,Y) = \sum_{xy \in A_x A_y} P(x,y) \log \frac{1}{P(x,y)}$$

$$H(X,Y) = 27/8 bits = 3.375 bits$$

- The marginal entropies

$$H(X) = 7/4 bits = 1.75 bits$$

$$H(Y) = 2 bits$$

An Example

- We can compute the **conditional distribution of *x*** for **each value of *y***, and the entropy of each of those conditional distributions

$P(x,y)$		x 1	2	3	4	$P(y)$
	1	$1/8$	$1/16$	$1/32$	$1/32$	$1/4$
y	2	$1/16$	$1/8$	$1/32$	$1/32$	$1/4$
	3	$1/16$	$1/16$	$1/16$	$1/16$	$1/4$
	4	$1/4$	0	0	0	$1/4$
$P(x)$		$1/2$	$1/4$	$1/8$	$1/8$	

$$P(x \mid y) = P(x,y) / P(y)$$

$P(x \mid y)$		x 1	2	3	4
	1	$1/2$	$1/4$	$1/8$	$1/8$
y	2	$1/4$	$1/2$	$1/8$	$1/8$
	3	$1/4$	$1/4$	$1/4$	$1/4$
	4	1	0	0	0

An Example

- We can compute the conditional distribution of x for each value of y, and the entropy of each of those conditional distributions

$P(x,y)$		x 1	2	3	4	$P(y)$
	1	$1/8$	$1/16$	$1/32$	$1/32$	$1/4$
y	2	$1/16$	$1/8$	$1/32$	$1/32$	$1/4$
	3	$1/16$	$1/16$	$1/16$	$1/16$	$1/4$
	4	$1/4$	0	0	0	$1/4$
$P(x)$		$1/2$	$1/4$	$1/8$	$1/8$	

$$P(x \mid y) = P(x,y) / P(y)$$

$$H(X \mid y) = \sum_{x \in A_x} P(x \mid y) \log \frac{1}{P(x \mid y)}$$

$P(x \mid y)$		x 1	2	3	4
	1	$1/2$	$1/4$	$1/8$	$1/8$
y	2	$1/4$	$1/2$	$1/8$	$1/8$
	3	$1/4$	$1/4$	$1/4$	$1/4$
	4	1	0	0	0

An Example

- We can compute the conditional distribution of x for each value of y, and the entropy of each of those conditional distributions

$P(x,y)$		x = 1	2	3	4	$P(y)$
y	1	$1/8$	$1/16$	$1/32$	$1/32$	$1/4$
	2	$1/16$	$1/8$	$1/32$	$1/32$	$1/4$
	3	$1/16$	$1/16$	$1/16$	$1/16$	$1/4$
	4	$1/4$	0	0	0	$1/4$
$P(x)$		$1/2$	$1/4$	$1/8$	$1/8$	

$$P(x \mid y) = P(x,y) / P(y)$$

$$H(X \mid y) = \sum_{x \in A_x} P(x \mid y) \log \frac{1}{P(x \mid y)}$$

$P(x \mid y)$		x = 1	2	3	4	$H(X \mid y)$/bits
y	1	$1/2$	$1/4$	$1/8$	$1/8$	$7/4$
	2	$1/4$	$1/2$	$1/8$	$1/8$	$7/4$
	3	$1/4$	$1/4$	$1/4$	$1/4$	2
	4	1	0	0	0	0

$$H(X \mid Y) = \sum_{y \in A_Y} P(y) H(X \mid y)$$

An Example

- We can compute the conditional distribution of x for each value of y, and the entropy of each of those conditional distributions

$P(x,y)$		x				$P(y)$
		1	2	3	4	
	1	1/8	1/16	1/32	1/32	1/4
y	2	1/16	1/8	1/32	1/32	1/4
	3	1/16	1/16	1/16	1/16	1/4
	4	1/4	0	0	0	1/4
$P(x)$		1/2	1/4	1/8	1/8	

$$P(x\,|\,y) = P(x,y)\,/\,P(y)$$

$$H(X\,|\,y) = \sum_{x\in A_x} P(x\,|\,y)\log\frac{1}{P(x\,|\,y)}$$

$P(x\,|\,y)$		x				$H(X\,|\,y)$/bits
		1	2	3	4	
	1	1/2	1/4	1/8	1/8	7/4
y	2	1/4	1/2	1/8	1/8	7/4
	3	1/4	1/4	1/4	1/4	2
	4	1	0	0	0	0
						$H(X\,|\,Y) = 11/8$

$$H(X\,|\,Y) = \sum_{y\in A_Y} P(y)H(X\,|\,y)$$

$$H(X\,|\,Y) = 1.375\,bits$$

An Example

- Lets compare $H(X)$ with $H(X \mid Y)$ and with each $H(X \mid y)$

$P(x,y)$		x 1	2	3	4	$P(y)$
	1	1/8	1/16	1/32	1/32	1/4
y	2	1/16	1/8	1/32	1/32	1/4
	3	1/16	1/16	1/16	1/16	1/4
	4	1/4	0	0	0	1/4
$P(x)$		1/2	1/4	1/8	1/8	

$$H(X) = 7/4 bits = 1.75 bits$$

$$\boxed{H(X \mid Y) \le H(X)}$$

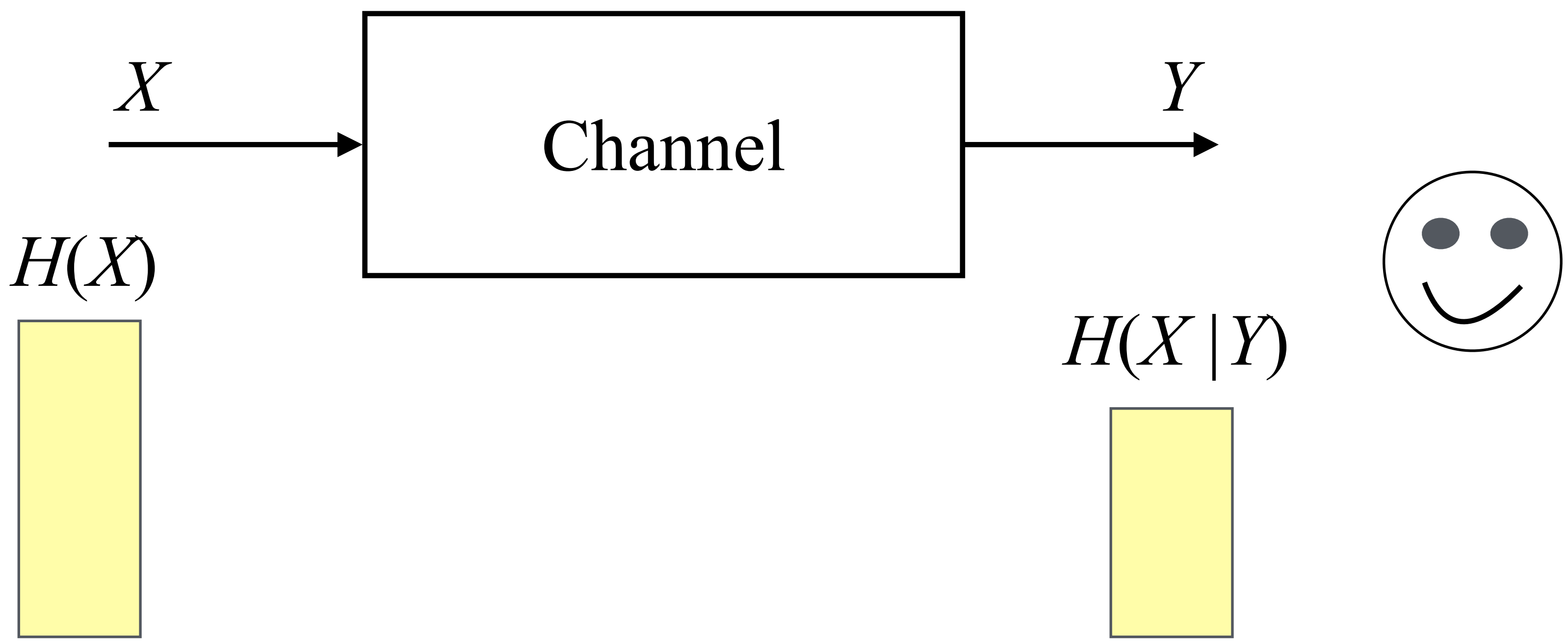
$P(x \mid y)$		x 1	2	3	4	$H(X \mid y)$/bits	
	1	1/2	1/4	1/8	1/8	7/4	Equal to $H(X)$
y	2	1/4	1/2	1/8	1/8	7/4	
	3	1/4	1/4	1/4	1/4	2	Larger than $H(X)$
	4	1	0	0	0	0	Smaller than $H(X)$

$$H(X \mid Y) = 11/8$$

$$H(X \mid Y) = 1.375 bits$$

Mutual Information, Entropy and a Channel

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



- For some y
  - $H(X \mid y) > H(X)$ - Some y increase the uncertainty about X
  - $H(X \mid y) < H(X)$ - Some y reduce the uncertainty about X
  - $H(X \mid y) = H(X)$ - Some y do not change the uncertainty about X

Few Demos

Show that $H(X, Y) = H(X) + H(Y / X)$

- The chain rule for entropy follows from the decomposition of a joint probability

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

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

$$H(X,Y) = \sum_{xy} P(x)P(y|x) \left(\log \frac{1}{P(x)} + \log \frac{1}{P(y|x)} \right)$$

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

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

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

Show that the Mutual Information is symmetric

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

$$= \sum_x P(x) \log \frac{1}{P(x)} - \sum_{xy} P(x,y) \log \frac{1}{P(x \mid y)}$$

$$= \sum_x P(x) \log \frac{1}{P(x)} \sum_y P(y \mid x) - \sum_{xy} P(x,y) \log \frac{1}{P(x \mid y)}$$

$$= \sum_{xy} P(x,y) \log \frac{1}{P(x)} - \sum_{xy} P(x,y) \log \frac{1}{P(x \mid y)}$$

$$= \sum_{xy} P(x,y) \log \frac{P(x \mid y)}{P(x)}$$

$$= \sum_{xy} P(x,y) \log \frac{P(x,y)}{P(x)P(y)}$$

$$\boxed{I(X;Y) = \sum_{xy} P(x,y) \log \frac{P(x,y)}{P(x)P(y)}}$$

Show that the Mutual Information is symmetric

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

$$\boxed{I(X;Y) = \sum_{xy} P(x,y) \log \frac{P(x,y)}{P(x)P(y)}}$$

- This expression is symmetric in x and y so

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

Show that the Mutual Information is non negative

$$\boxed{D_{KL}(P \| Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}}$$

$$\boxed{D_{KL}(P \| Q) \geq 0}$$

$$I(X;Y) = \sum_{xy} P(x,y) \log \frac{P(x,y)}{P(x)P(y)}$$

$$I(X;Y) = D_{KL}(P(x,y) \| P(x)P(y))$$

- The mutual information is a relative entropy between the distribution $P(x,y)$ and the distribution $P(x).P(y)$, and so due to the Gibbs' inequality $I(X;Y) \geq 0$
- The equality only if $P(x,y) = P(x)P(y)$, that is, if X and Y are independent

Further Reading and Summary

Q&A

Further Reading

- **Recommend Readings**
  - Information Theory, Inference, and Learning Algorithms from David MacKay, 2015, pages 138 - 144.
- **Supplemental readings:**

What you should know

- The meaning and the definition of:
  - $H(X, Y)$
  - $H(X \mid y)$ and $H(X \mid Y)$
  - $I(X; Y)$
- The main relations between $H(X)$, $H(Y)$, $H(X, Y)$, $H(X \mid Y)$, $H(Y \mid X)$, $I(X.Y)$
- How to interpret them in terms of a communication channel
- The main properties of them
- How to express some in terms of relative entropies

Further Reading and Summary

Q&A