Information and Communication Theory

Lecture 1

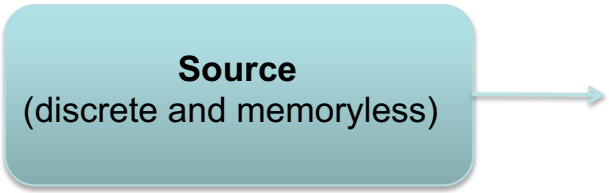
Discrete Memoryless Sources, Entropy, Mutual Information, Inequalities

Mário A. T. Figueiredo

DEEC, Instituto Superior Técnico, University of Lisbon, **Portugal**

2023

Discrete Memoryless Source

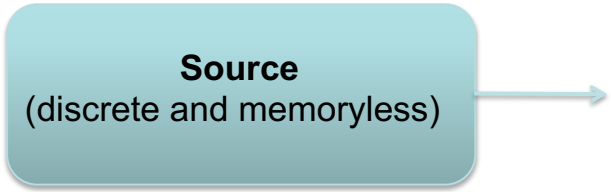


$$X \in \{1, ..., N\}$$

$$p_i = \mathbb{P}(X = i)$$

- Simplest model of a random source.
- Memoryless: every symbol is independent of past/future ones.
- Discrete: the set of possible symbols $\mathcal{X} = \{1, ..., N\}$ is discrete.
- Notation: $p_x = \mathbb{P}(X = x) = f_X(x) = p_X(x) = p(x)$
  (all commonly used)
- A DMS is just a random variable.

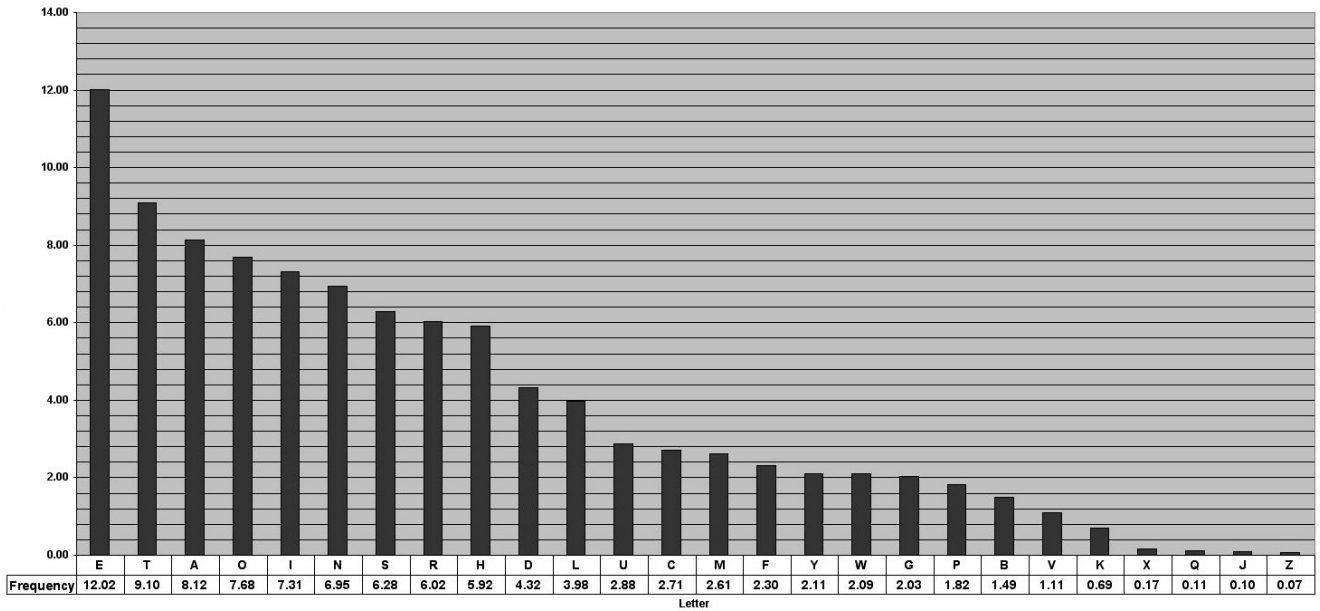
Discrete Memoryless Source



$$X \in \{1, ..., N\}$$

$$p_i = \mathbb{P}(X = i)$$

- Example: English letters, $\mathcal{X} = \{a, b, ..., z\}$



- Very crude model; sample: OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL

Measuring Uncertainty: Axioms

- Goal: find of measure of uncertainty $H(X) \geq 0$ for a DMS X
- Uncertainty only depends on $(p_1, ..., p_N)$ (why?)
- Thus, we write $H(X) = H(p_1, ..., p_N) = H(f_X(1), ..., f_X(N))$ (different notations)
- Properties/axioms that $H(p_1, ..., p_N)$ should satisfy:
  - ✓ Symmetry and continuity: symmetric continuous function of $p_1, ..., p_N$ (symmetric = invariant under argument permutations)
  - ✓ Monotonicity: $H(\frac{1}{N}, \frac{1}{N}, ..., \frac{1}{N})$ should monotonically increase with N
  - ✓ Additivity/extensivity: if X and Y are independent,

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

  - ✓ Grouping: if $p_A = p_1 + ... + p_i$ and $p_B = p_{i+1} + ... + p_N$, then

$$H(p_1, ..., p_N) = H(p_A, p_B) + p_A\, H\left(\frac{p_1}{p_A}, ..., \frac{p_i}{p_A}\right) + p_B\, H\left(\frac{p_{i+1}}{p_B}, ..., \frac{p_N}{p_B}\right)$$

Measuring Uncertainty: Examples

- Consider four DMS $X \in \{1, 2\}$, $Y \in \{1, 2, 3\}$, $Z \in \{1, 2, 3\}$, $T \in \{1, 2, 3, 4, 5\}$, with the following probability distributions

$\mathbb{P}(\cdot)$	1	2	3	4	5
X	1/2	1/2	·	·	·
Y	1/3	1/3	1/3	·	·
Z	1/2	1/4	1/4	·	·
T	1/3	1/4	1/4	1/12	1/12

- Using the axioms, can you sort the four variables by increasing order of uncertainty?

- Notice that $H(p_1, ..., p_N, 0) = H(p_1, ..., p_N)$: we can ignore zero-probability symbols.

Measuring Uncertainty: Entropy

- Shannon (1948) showed: the only function satisfying the axioms has the form

$$H(X) = H(p_1, ..., p_N) = -K \sum_{i=1}^{N} p_i \log_b p_i$$

- Due to the similarity with statistical physics, he called it entropy.
- Convention: $0 \log 0 \equiv 0$ (by continuity, since $\lim_{c \to 0^+} c \log c = 0$).
  Confirms $H(p_1, ..., p_N, 0) = H(p_1, ..., p_N)$.
- Constant $K > 0$ is arbitrary.
- The base b is arbitrary, as long as $b > 1$ (why? recall $\log_a u = \frac{\log_b u}{\log_b a}$)
- Another notation

$$H(X) = -\sum_{x \in \mathcal{X}} f_X(x) \log f_X(x) = -\mathbb{E}\big[\log f_X(X)\big]$$

Entropy: More Properties

- Since it is arbitrary, choose $K = 1$, thus $H(X) = -\mathbb{E}\big[\log_b f_X(X)\big]$
- Units: bits/symbol, for $b = 2$; nats/symbol, for $b = e$.
- Since $\log_2 e \simeq 1.443$, one nat is approximately 1.443 bits.
- Entropy equals expected surprise of outcome, $-\log f_X(x) = \log \frac{1}{f_X(x)}$

$$H(X) = \mathbb{E}\left[\log_b \frac{1}{f_X(X)}\right]$$

- Positivity:

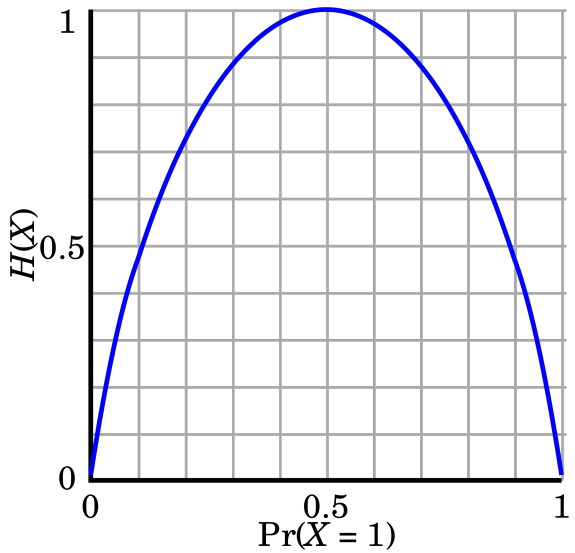
$$H(X) = -\sum_{x\in\mathcal{X}} \underbrace{\overbrace{f_X(x)}^{\geq 0}\overbrace{\log \underbrace{f_X(x)}_{\leq 1}}^{\leq 0}}_{\leq 0} \geq 0$$

- Zero entropy: $H(X) = 0$ if and only if $f_X(x) = 1$ for some x, thus $f_X(x') = 0$ for any $x' \neq x$ (prove it!)

Entropy: Binary DMS

- Binary DMS $X \in \mathcal{X} = \{0, 1\}$ (or any other set with 2 elements)
- Probabilities: $f_X(1) = \mathbb{P}[X = 1] = p$
  $f_X(0) = \mathbb{P}[X = 0] = 1 - p.$
- Entropy of a binary DMS (in bits/symbol)

$$H(X) = H(p, 1-p) = -p \log_2 p - (1-p) \log_2 (1-p)$$



Entropy: Two (or More) DMS

- Two (maybe dependent) DMS: $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$.
- Joint distribution $f_{X,Y}(x,y) = \mathbb{P}(X = x, Y = y)$
- Joint entropy:

$$H(X,Y) = -\mathbb{E}_{X,Y}\left[\log f_{X,Y}(X,Y)\right] = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} f_{X,Y}(x,y)\log f_{X,Y}(x,y)$$

- Independent variables: $X \perp\!\!\!\perp Y \Leftrightarrow f_{X,Y}(x,y) = f_X(x)f_Y(y)$,

$$\begin{aligned} H(X,Y) &= -\mathbb{E}_{X,Y}\left[\log f_X(X) + \log f_Y(Y)\right] \\ &= -\mathbb{E}_{X,Y}\left[\log f_X(X)\right] - \mathbb{E}_{X,Y}\left[\log f_Y(Y)\right] \\ &= -\mathbb{E}_{X}\left[\log f_X(X)\right] - \mathbb{E}_{Y}\left[\log f_Y(Y)\right] = H(X) + H(Y) \end{aligned}$$

- ...confirming that H satisfies the additivity axiom.

Conditional Entropies

- Conditional distribution: $f_{Y|X}(y|x) = \dfrac{f_{X,Y}(x,y)}{f_X(x)}$, for $f_X(x) > 0$.
- Naturally, $f_{Y|X}(y|x) \geq 0$ and $\sum_{y\in\mathcal{Y}} f_{Y|X}(y|x) = 1$, for any x
- Conditioned entropy (for a given x): $H(Y|X=x)$

$$H(Y|X=x) = -\mathbb{E}_Y\left[\log f_{Y|X}(Y|x)|X=x\right] = -\sum_{y\in\mathcal{Y}} f_{Y|X}(y|x)\log f_{Y|X}(y|x)$$

...uncertainty of Y, given that $X = x$.

- Conditional entropy: expectation (in X) of the conditioned entropy:

$$H(Y|X) = \sum_{x\in\mathcal{X}} f_X(x)H(Y|X=x) = -\mathbb{E}_{X,Y}[\log f_{Y|X}(Y|X)]$$

...expected uncertainty of Y, given X.

Bayes for Entropies

- From the conditional probability definition:
  $f_{X,Y}(x,y) = f_{X|Y}(x|y)\, f_Y(y)$.
- Joint entropy, $H(X,Y) = -\mathbb{E}_{X,Y}\big[\log f_{X,Y}(X,Y)\big]$,

$$\begin{aligned} H(X,Y) &= -\mathbb{E}_{X,Y}\big[\log f_{X,Y}(X,Y)\big] \\ &= -\mathbb{E}_{X,Y}\big[\log f_{X|Y}(X|Y) + \log f_Y(Y)\big] \\ &= -\mathbb{E}_{X,Y}\big[\log f_{X|Y}(X|Y)\big] - \mathbb{E}_{X,Y}\big[\log f_Y(Y)\big] \\ &= H(X|Y) + H(Y) \end{aligned}$$

- By symmetry, $H(X,Y) = H(Y|X) + H(X)$.
- Bayes for entropies:

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

- Independent variables: if $X \perp\!\!\!\perp Y$,

$H(X,Y) = H(X)+H(Y) \;\Rightarrow\; H(X|Y) = H(X)$ and $H(Y|X) = H(Y)$

Chain Rules

- Recall that $H(X,Y) = H(Y|X) + H(X)$.

- By recursion, this extends to more than two variables: X_1, X_2, ...,X_n,

$$\begin{aligned} H(X_1, X_2, ..., X_n) &= H(X_1|X_2, ..., X_n) + H(X_2, ..., X_n) \\ &= H(X_1|X_2, ..., X_n) + H(X_2|X_3..., X_n) + H(X_3, ..., X_n) \\ &\vdots \\ &= \sum_{i=1}^{n} H(X_i|X_{i+1}, ..., X_n) \end{aligned}$$

- This is called a chain rule.

Mutual Information

- Recall that $H(X,Y) = H(Y|X) + H(X) = H(X|Y) + H(Y)$.
- Consequently

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

...called mutual information

- Independent variables: if $X \perp\!\!\!\perp Y$,

$$H(X|Y) = H(X) \quad \Rightarrow \quad I(X;Y) = 0$$

- Deterministically dependent variables:

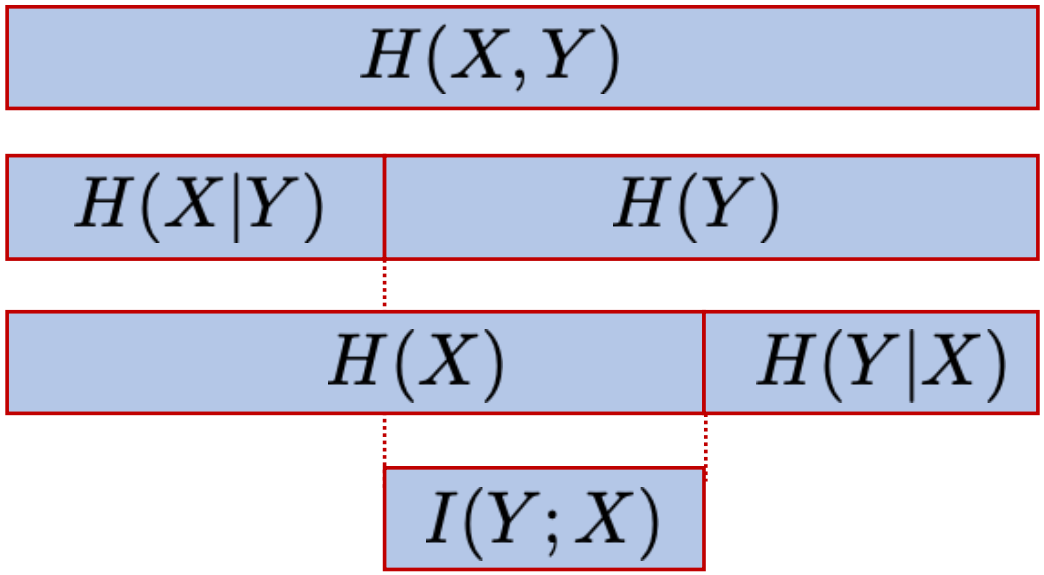
$$\begin{array}{lllll} Y = g(X) & \Rightarrow & H(Y|X) = 0 & \Rightarrow & I(X;Y) = H(Y) \\ X = g(Y) & \Rightarrow & H(X|Y) = 0 & \Rightarrow & I(X;Y) = H(X) \end{array}$$

- Upper-bound on the mutual information:

$$\left.\begin{array}{l} I(X;Y) = H(X) - H(X|Y) \leq H(X) \\ I(X;Y) = H(Y) - H(Y|X) \leq H(Y) \end{array}\right\} \Rightarrow I(X;Y) \leq \min\{H(X), H(Y)\}$$

Summary of Relationships

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



Kullback-Leibler Divergence

- Let $X, X' \in \mathcal{X}$ be two random variables on the same set $\mathcal{X}$
- The Kullback-Leibler divergence (KLD) is defined as

$$D_{\mathsf{KL}}(f_X \parallel f_{X'}) = \sum_{x \in \mathcal{X}} f_X(x) \log \frac{f_X(x)}{f_{X'}(x)}$$
$$= \mathbb{E}_X \left[\log \frac{f_X(X)}{f_{X'}(X)} \right]$$

- If $f_X(x) = f_{X'}(x)$, for all $x \in \mathcal{X}$, then

$$\log \frac{f_X(x)}{f_{X'}(x)} = 0 \quad \Rightarrow \quad D_{\mathsf{KL}}(f_X \parallel f_{X'}) = 0$$

- In general, $D_{\mathsf{KL}}(f_X \parallel f_{X'}) \neq D_{\mathsf{KL}}(f_{X'} \parallel f_X)$ (not symmetric)

Kullback-Leibler Divergence

- If, for some x, $f_{X'}(x) = 0$ and $f_X(x) > 0$, then

$$D_{\mathsf{KL}}(f_X \parallel f_{X'}) = +\infty$$

- Relationship with the mutual information:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= \mathbb{E}_X[-\log f_X(X)] + \mathbb{E}_{X,Y}[\log f_{X|Y}(X|Y)] \\ &= \mathbb{E}_{X,Y}[-\log f_X(X) + \log f_{X|Y}(X|Y)] \\ &= \mathbb{E}_{X,Y}[-\log f_X(X) + \log f_{X,Y}(X,Y) - \log f_Y(Y)] \\ &= \mathbb{E}_{X,Y}\left[\log \frac{f_{X,Y}(X,Y)}{f_X(X)\, f_Y(Y)}\right] = D_{\mathsf{KL}}(f_{X,Y} \parallel f_X\, f_Y) \end{aligned}$$

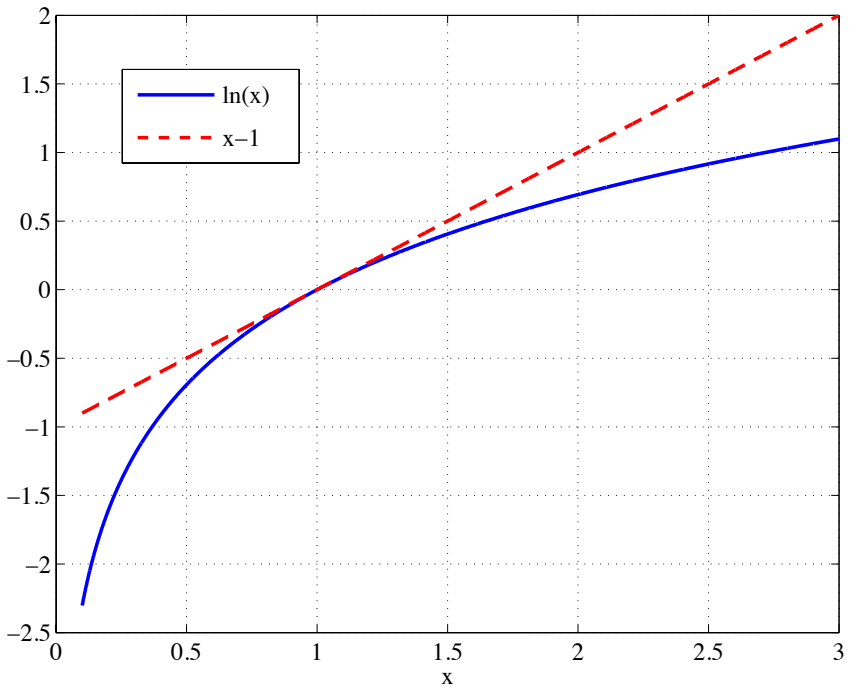
- If $X \perp\!\!\!\perp Y$, then $f_{X,Y}(x,y) = f_X(x)\, f_Y(y)$, and $I(X;Y) = 0$.

An Important Inequality

- Gibbs inequality for the natural logarithm $(\ln \equiv \log_e)$

$$\log_e x \leq x - 1$$

$$\log_e x = x - 1 \quad \Leftrightarrow \quad x = 1$$



- Other bases: $\log_b x = \dfrac{\log_e x}{\log_e b} \leq \dfrac{x-1}{\log_e b}$

Fundamental Inequality of Information Theory

- Fundamental inequality:

$$D_{\mathsf{KL}}(f_X \parallel f_{X'}) \geq 0$$
$$D_{\mathsf{KL}}(f_X \parallel f_{X'}) = 0 \quad \Leftrightarrow \quad f_X = f_{X'}$$

- Proof: let $A = \{x \in \mathcal{X} : \ f_X(x) > 0\}$;

$$\begin{aligned} -D_{\mathsf{KL}}(f_X \parallel f_{X'}) &= \sum_{x \in A} f_X(x) \log_b \frac{f_{X'}(x)}{f_X(x)} \qquad (0 \log 0 \equiv 0) \\ &\leq \frac{1}{\log_e b} \sum_{x \in A} f_X(x) \left(\frac{f_{X'}(x)}{f_X(x)} - 1 \right) \\ &= \frac{1}{\log_e b} \Bigg(\underbrace{\sum_{x \in A} f_{X'}(x)}_{\leq 1} - \underbrace{\sum_{x \in A} f_X(x)}_{=1} \Bigg) \ \leq \ 0 \end{aligned}$$

...clearly, equality requires $f_{X'}(x) = f_X(x)$, for all $x \in \mathcal{X}$

Corollaries of the Fundamental Inequality

- Non-negativity of the mutual information:

$$I(X;Y) = D_{\mathsf{KL}}(f_{X,Y} \parallel f_X\, f_Y) \geq 0$$
$$I(X;Y) = D_{\mathsf{KL}}(f_{X,Y} \parallel f_X\, f_Y) = 0 \quad \Leftrightarrow \quad f_{X,Y} = f_X\, f_Y, \text{i.e.} \quad X \perp\!\!\!\perp Y$$

- Conditioning reduces entropy; since $I(X;Y) = H(X) - H(X|Y)$,

$$H(X|Y) = H(X) - \overbrace{I(X;Y)}^{\geq 0} \leq H(X)$$
$$H(X|Y) = H(X) \quad \Leftrightarrow \quad X \perp\!\!\!\perp Y$$

- Upper-bound on joint entropy: since $H(X,Y) = H(X|Y) + H(Y)$,

$$H(X,Y) = H(X|Y) + H(Y) \leq H(X) + H(Y)$$
$$H(X,Y) = H(X) + H(Y) \quad \Leftrightarrow \quad X \perp\!\!\!\perp Y$$

More Corollaries of the Fundamental Inequality

- Upper-bound on joint entropy for several variables:

$$H(X_1, ..., X_n) = \sum_{i=1}^{n} H(X_i | X_{i+1}, ..., X_n) \leq \sum_{i=1}^{n} H(X_i)$$

$$H(X_1, ..., X_n) = \sum_{i=1}^{n} H(X_i) \Leftrightarrow \text{ all the } X_i \text{ are mutually independent}$$

- Maximum entropy: let $f_{X'}(x) = 1/N$, for $\mathcal{X} = \{1, ..., N\}$;

$$\begin{aligned} 0 \leq D_{\mathsf{KL}}(f_X, f_{X'}) &= \sum_{x \in \mathcal{X}} f_X(x) \log \frac{f_X(x)}{1/N} \\ &= \underbrace{\sum_{x \in \mathcal{X}} f_X(x) \log f_X(x)}_{=-H(X)} + \log N \underbrace{\sum_{x \in \mathcal{X}} f_X(x)}_{=1} \end{aligned}$$

...thus $H(X) \leq \log N$, with equality if and only if $f_X(x) = 1/N$.

The Data Processing Inequality

- Let X, Y, Z be such that $f_{Z|X,Y}(z|x,y) = f_{Z|Y}(z|y)$:

...conditionally on Y, X and Z and independent, $X \perp\!\!\!\perp Z \,|\, Y$.

- Since $X \perp\!\!\!\perp Z \,|\, Y$,

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

...all the information about X in the pair (Y,Z) is contained in Y.

- Also,

$$\begin{aligned} I(X;Y,Z) &= H(Y,Z) - H(Y,Z|X) \\ &= H(Y|Z) + H(Z) - H(Y|Z,X) - H(Z|X) \\ &= I(Z;X) + \underbrace{H(Y|Z) - H(Y|Z,X)}_{\geq 0} \geq I(Z;X) \end{aligned}$$

- Concluding: $I(X;Y) \geq I(X;Z)$

More on the Data Processing Inequality

- Let X, Y be a pair of random variables and $Z = g(Y)$; then

$$f_{Z|X,Y}(z|x,y) = f_{Z|Y}(z|y) \quad \Rightarrow \quad I(X;Y) \geq I(X;Z)$$

... no function of Y can have more information about X than Y itself.

- The equality case:

$$I(X;Y) = I(X;Z) \quad \Leftrightarrow \quad H(Y|Z) = H(Y|Z,X)$$

... that is, if $f_{Y|X,Z}(y|x,z) = f_{Y|Z}(y|z) \quad \Leftrightarrow \quad X \perp\!\!\!\perp Y|Z$

✓ one possibility: $Y = h(Z)$ (*e.g.*, $Z = g(Y)$, with g injective)

✓ another possibility: $X \perp\!\!\!\perp Y \quad \Rightarrow \quad I(X;Y) = I(X;Z) = 0.$

either Z knows as much as Y, or Y knows nothing.

Summary of Relationships and Inequalities

- $H(X) \geq 0$, with "$=$" $\Leftrightarrow$ $f_X(x') = 1, f_X(x) = 0$, for $x \neq x'$
- For $X \in \{1, ..., N\}$, $H(X) \leq \log N$, with "$=$" $\Leftrightarrow$ $f_X(x) = 1/N$
- $H(X, Y) = H(X|Y) + H(Y) = H(Y|X) + H(X)$
- $H(Y|X) = H(X|Y) + H(Y) - H(X)$
- $I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$
- $H(X, Y) = H(X) + H(Y) - I(X; Y)$
- $I(X; Y) \geq 0$, with "$=$" $\Leftrightarrow$ $X \perp\!\!\!\perp Y$
- $H(Y|X) \leq H(Y)$, with "$=$" $\Leftrightarrow$ $X \perp\!\!\!\perp Y$
- $H(X, Y) \leq H(X) + H(Y)$, with "$=$" $\Leftrightarrow$ $X \perp\!\!\!\perp Y$

Recommended Reading

- J. Massey, "Applied Digital Information Theory", Lectures Notes, ETH Zurich, 1980.
  `https://www.isiweb.ee.ethz.ch/archive/massey_scr/adit1.pdf`

- T. Cover and J. Thomas, "Elements of Information Theory", John Wiley & Sons, 2006 (Chapter 2).