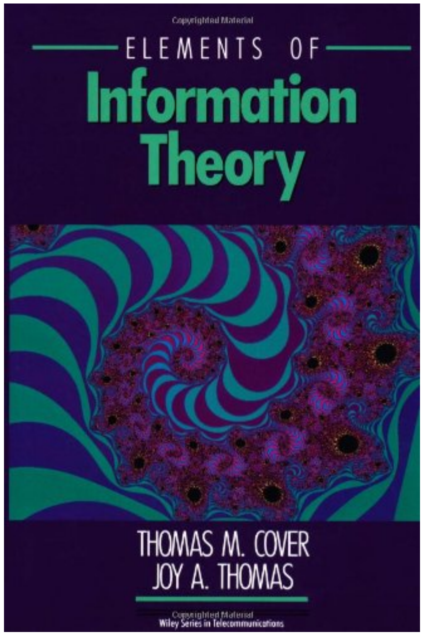
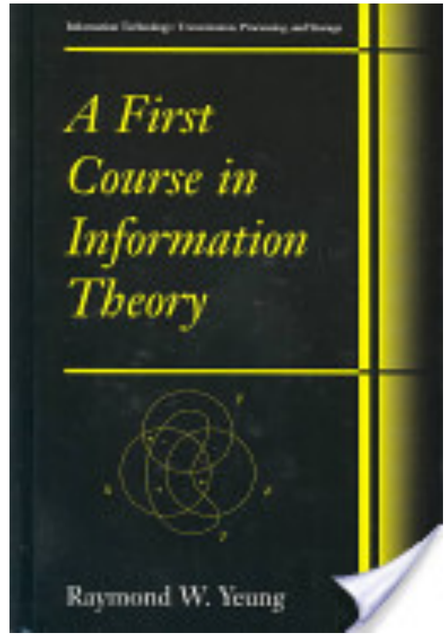
Analysis and processing of information

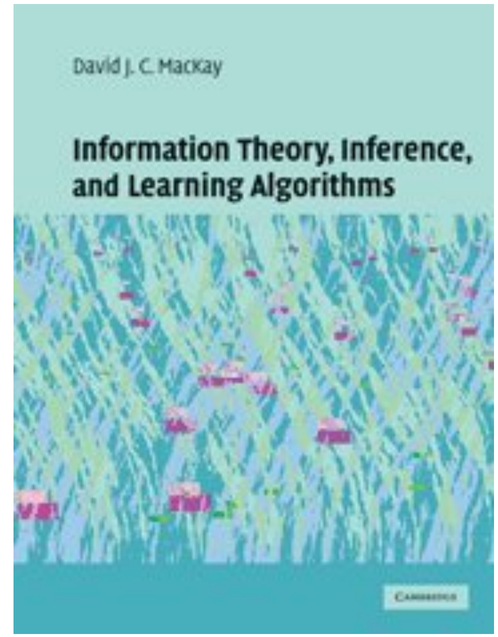
S. Voloshynovskiy

Recall of Information Theory

Recommended books







Course Outline

- Enropy, conditional entropy, joint entropy
- KLD
- Mutual information
- Particularites of IT measures for continuos random variables

Recall of probability

Chain rule for probability: joint probility $p\left(x_1, x_2, \ldots, x_n\right)$

$$p\left(x_1, x_2, \ldots, x_n\right) = p\left(x_1\right) p\left(x_2 \middle| x_1\right) p\left(x_3 \middle| x_2, x_1\right) \ldots p\left(x_n \middle| x_{n-1}, \ldots, x_2, x_1\right) =$$

$$= \prod_{i=1}^{n} p\left(x_i \middle| x_{i-1}, \ldots, x_2, x_1\right)$$

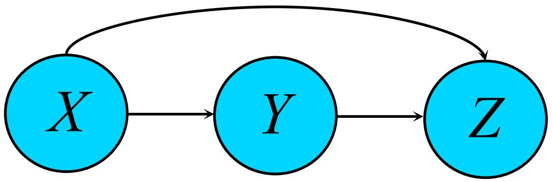
$$p\left(x_1, x_2\right) = p\left(x_1\right) p\left(x_2 \middle| x_1\right) \xrightarrow{\text{independence}} p\left(x_1, x_2\right) = p\left(x_1\right) p\left(x_2\right)$$

$$p\left(x_1, x_2, \ldots, x_n\right) = p\left(x_1\right) p\left(x_2\right) p\left(x_3\right) \ldots p\left(x_n\right) = \prod_{i=1}^{n} p\left(x_i\right)$$

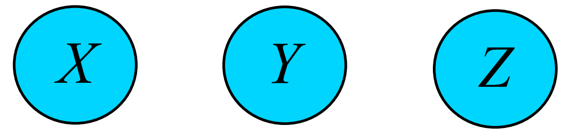
For the independent R.V.s $\{X_i\}$

$$p\left(x_1, x_2, \ldots, x_n\right) = p\left(x\right) p\left(x\right) p\left(x\right) \ldots p\left(x\right) = \prod_{i=1}^{n} p\left(x\right) = (p(x))^n$$

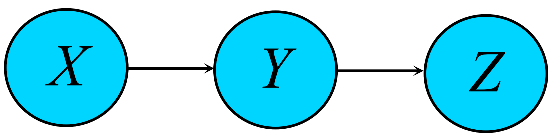
Recall of probability



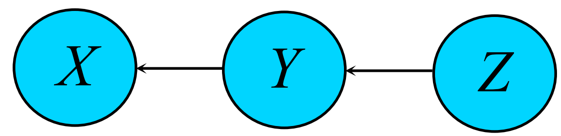
$$p(x,y,z) = p(x)\,p(y|x)\,p(z|x,y)$$



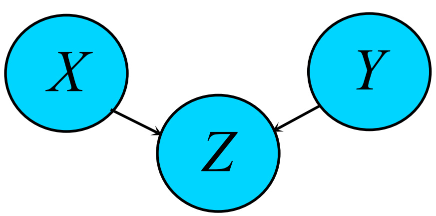
$$p(x,y,z) = p(x)\,p(y)\,p(z)$$



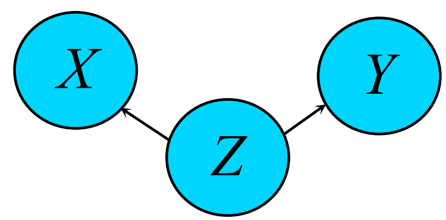
$$p(x,y,z) = p(x)\,p(y|x)\,p(z|y)$$



$$p(x,y,z) = p(z)\,p(y|z)\,p(x|y)$$



$$p(x,y,z) = p(z|x,y)\,p(x)\,p(y)$$



$$p(x,y,z) = p(z)\,p(x|z)\,p(y|z)$$

Entropy

Recall from: Elements of Information Theory

Definition (entropy): *Entropy* of discrete r.v. $X \in \mathcal{X} \quad \mathcal{X} = \left\{ x_1, x_2, \ldots, x_N \right\}$

$$\boxed{H\left(X\right) = -\sum_{x \in \mathcal{X}} p_X\left(x\right) \log_2 p_X\left(x\right) = -E_{p_X}\left[\log p_X\left(x\right)\right]} \quad \text{(bits)}$$

$$E_{p_X}\left[x\right] = \sum_{i=1}^{N} x_i p(x_i)$$

$$H\left(X\right) = -\sum_{i=1}^{n} \Pr\left\{X = x_i\right\} \log_2 \Pr\left\{X = x_i\right\}$$

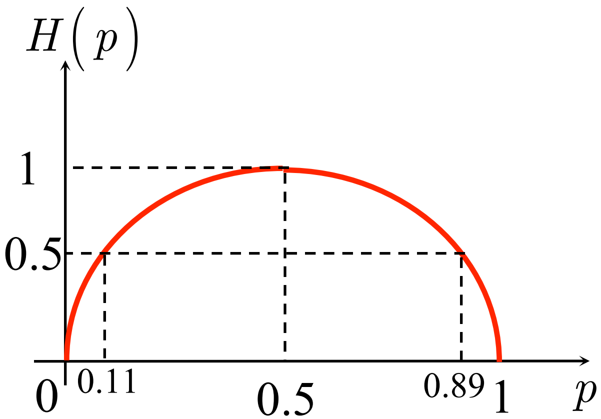
Entropie: discrete random variable

Discrete random variable: $X \in \{0,1\}$

$\Pr\{X = 0\} = p$

$\Pr\{X = 1\} = 1 - p$

$$H(X) = -p\log_2 p - (1-p)\log_2(1-p) := H(p) \text{ or } H_2(p)$$



The function $H(p)$ is:

- Symmetric wrt $p = 0.5$;
- Maximum at $p = 0.5$ $(H(p) = 1)$.

The entropy is maximal, if the symbols are equilikely.

Joint entropy

Definition (joint entropy): *Joint entropy* of two discrete random variables X and Y is defined by:

$$\boxed{H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{X,Y}(x,y)\log_2 p_{X,Y}(x,y) = -E_{p_{X,Y}}\left[\log p_{X,Y}(x,y)\right]}$$

$$\mathcal{X} = \{x_1, x_2, \dots, x_n\} \qquad \mathcal{Y} = \{y_1, y_2, \dots, y_m\}$$

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

Conditional entropy

Definition (conditional entropy): *Conditional entropy* of r.v. X given Y is defined by:

$$\boxed{H\left(X\middle|Y\right) = \sum_{y\in\mathcal{Y}} p_Y\left(y\right) H\left(X\middle|Y=y\right) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{X,Y}\left(x,y\right)\log p_{X|Y}\left(x\middle|y\right)}$$

$$\mathcal{X} = \left\{x_1, x_2, \ldots, x_n\right\} \qquad \mathcal{Y} = \left\{y_1, y_2, \ldots, y_m\right\}$$

$$H\left(X\middle|Y=y\right) = -\sum_{i=1}^{n} p_{X|Y}\left(x_i\middle|y\right)\log p_{X|Y}\left(x_i\middle|y\right)$$

$$\boxed{H\left(X\middle|Y\right) \neq H\left(Y\middle|X\right)}$$

Some properties of joint entropy

Chain rule for 2 R.V.s

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

Proof:

$$-\log p_{X,Y}\left(x,y\right)=-\log p_{Y|X}\left(y\middle|x\right)p_X\left(x\right)=-\log p_{Y|X}\left(y\middle|x\right)-\log p_X\left(x\right).$$

$$H\left(X,Y\right)=E_{p_{X,Y}}\left[-\log p_{X,Y}\left(x,y\right)\right]=E_{p_{X,Y}}\left[-\log p_X\left(x\right)\right]+E_{p_{X,Y}}\left[-\log p_{Y|X}\left(y\middle|x\right)\right]=$$
$$=H\left(X\right)+H\left(Y\middle|X\right).$$

Some properties of joint entropy

General chain rule for entropy

$$\boxed{H\left(X_1, X_2, \dots, X_n\right) = \sum_{i=1}^{n} H(X_i \middle| X_{i-1}, \dots, X_1)}$$

Proof:

=, if independent

$$H\left(X_1, X_2\right) = H\left(X_1\right) + H\left(X_2 \middle| X_1\right) \Rightarrow H\left(X_1, X_2\right) \le H\left(X_1\right) + H\left(X_2\right)$$

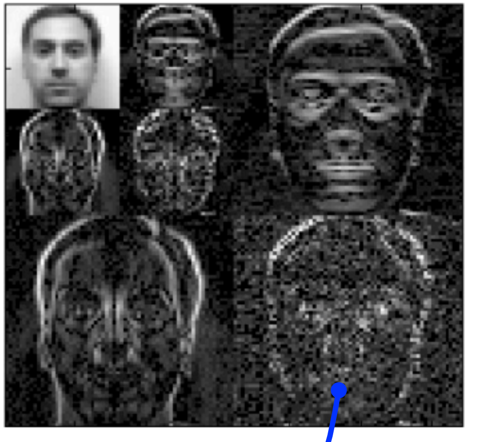
$$H\left(X_1, X_2, X_3\right) = H\left(X_1\right) + H\left(X_2, X_3 \middle| X_1\right) = H\left(X_1\right) + H\left(X_2 \middle| X_1\right) + H\left(X_3 \middle| X_2, X_1\right)$$

$$H\left(X_1, X_2, \dots, X_n\right) = H\left(X_1\right) + H\left(X_2 \middle| X_1\right) + \dots + H\left(X_n \middle| X_{n-1}, \dots, X_1\right) =$$

$$= \sum_{i=1}^{n} H(X_i \middle| X_{i-1}, \dots, X_1)$$

Conditional entropy

$$H\left(X_1, X_2, \ldots, X_n\right) = \sum_{i=1}^{n} H(X_i | \underbrace{X_{i-1}, \ldots, X_1})$$



Conditional entropy

$$\boxed{H\left(X_1, X_2, \ldots, X_n\right) \leq \sum_{i=1}^{n} H(X_i)}$$

equality for independent $\left\{X_i\right\}$

Proof:

For independent $\left\{X_i\right\}$

$$H\left(X_2 \middle| X_1\right) = H\left(X_2\right)$$

$$H\left(X_3 \middle| X_2, X_1\right) = H\left(X_3\right)$$

$$H(X_i | X_{i-1}, \ldots, X_1) = H(X_i)$$

$$H\left(X_1, X_2, \ldots, X_n\right) = H\left(X_1\right) + H\left(X_2 \middle| X_1\right) + \ldots + H\left(X_n \middle| X_{n-1}, \ldots, X_1\right) =$$

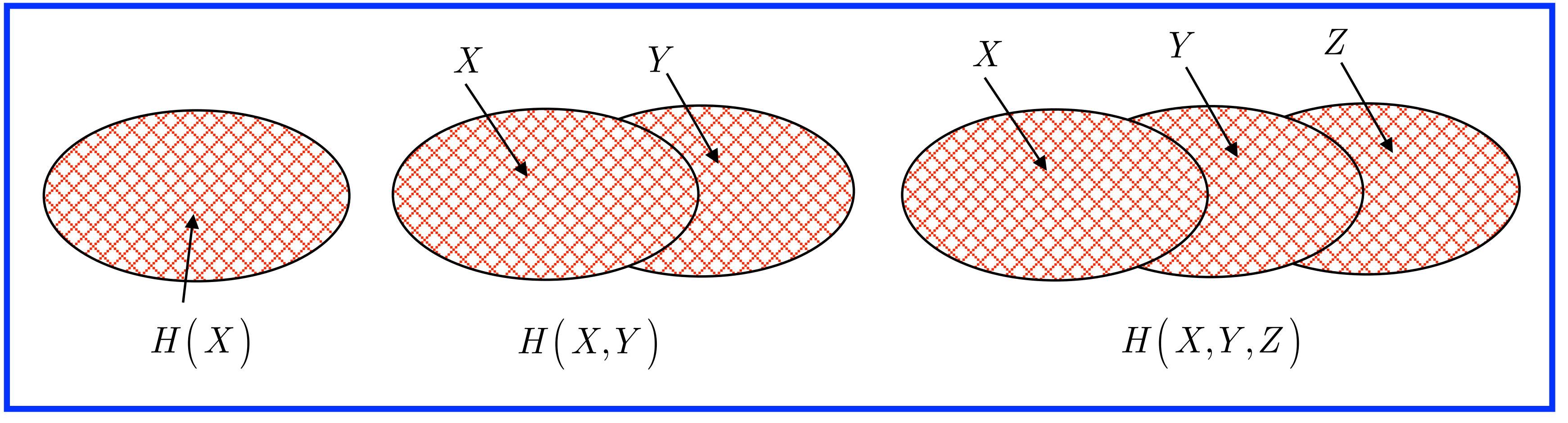
$$H\left(X_1, X_2, \ldots, X_n\right) = H\left(X_1\right) + H\left(X_2\right) + \ldots + H\left(X_n\right) = \sum_{i=1}^{n} H\left(X_i\right)$$

Properties of entropy

$$\boxed{H(X) \le H(X,Y) \le H(X,Y,Z) \le \ldots}$$

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

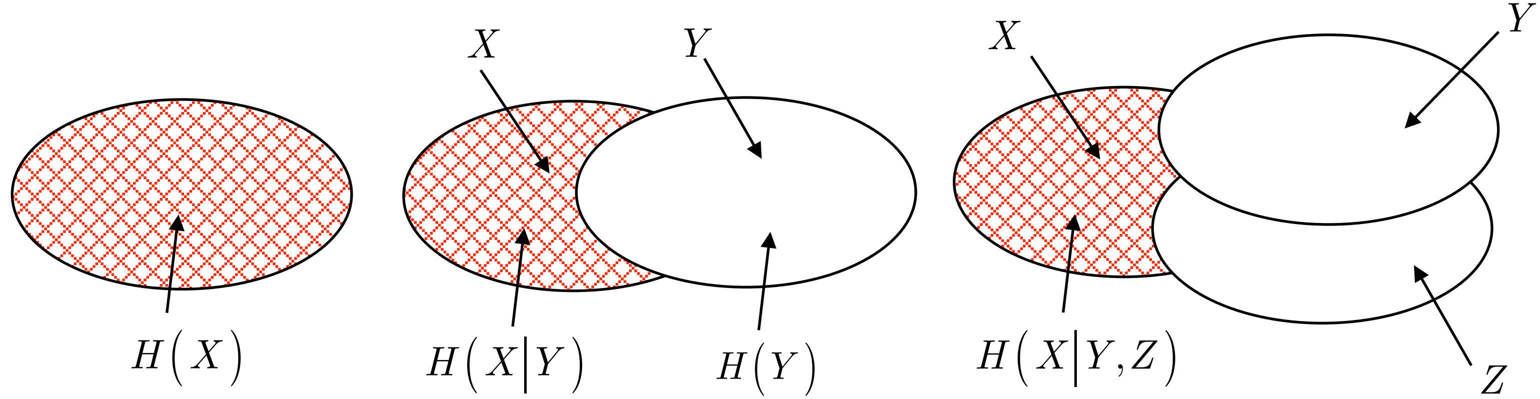
$$H(X,Y) = H(X,Y,Z) \Leftrightarrow Z = f(X,Y)$$



Properties of entropy

$$\boxed{H(X) \geq H(X|Y) \geq H(X|Y,Z) \geq \ldots}$$

$$H(X) = H(X|Y) \Leftrightarrow X \perp Y$$
$$H(X|Y) = H(X|Y,Z) \Leftrightarrow X \perp Z|Y \ldots$$



Relative entropy

Definition (relative entropy): ***Relative entropy or Kullback-Leibler distance*** **between pmfs** $p(x)$ **and** $q(x)$**:**

$$\boxed{D(p\|q) = \sum_{x\in\mathcal{X}} p(x)\log_2 \frac{p(x)}{q(x)} = E_p\left[\log_2 \frac{p(x)}{q(x)}\right]}$$

with the conventions:

$$0\log_2 \frac{0}{q(x)} = 0,\ \forall q(x)$$

$$p(x)\log_2 \frac{p(x)}{0} = +\infty,\ \forall p(x) > 0.$$

$$\boxed{D(p\|q) \neq D(q\|p)}$$

Relative entropy

Example:

$$x \in \mathcal{X} = \{0{,}1\};\ p(0) = a;\ p(1) = 1 - a;\ q(0) = b;\ q(1) = 1 - b.$$

$$D\left(p\|q\right) = \sum_{x \in \mathcal{X}} p\left(x\right) \log_2 \frac{p\left(x\right)}{q\left(x\right)} = a \log_2 \frac{a}{b} + (1 - a) \log_2 \frac{1 - a}{1 - b};$$

$$D\left(q\|p\right) = \sum_{x \in \mathcal{X}} q\left(x\right) \log_2 \frac{q\left(x\right)}{p\left(x\right)} = b \log_2 \frac{b}{a} + (1 - b) \log_2 \frac{1 - b}{1 - a}.$$

$$a = \frac{1}{4}; b = \frac{1}{8},\ D\left(p\|q\right) = \frac{1}{4} \log_2 \frac{8}{4} + \left(1 - \frac{1}{4}\right) \log_2 \left(\frac{1 - \frac{1}{4}}{1 - \frac{1}{8}} \right) = 0.0832 \text{ bit};$$

$$\boxed{D\left(p\|q\right) \neq D\left(q\|p\right)}$$

$$D\left(q\|p\right) = \frac{1}{8} \log_2 \frac{4}{8} + \left(1 - \frac{1}{8}\right) \log_2 \left(\frac{1 - \frac{1}{8}}{1 - \frac{1}{4}} \right) = 0.0696 \text{ bit}.$$

Mutual Information

Definition (mutual information): *Mutual information* between two r.v. X and Y is defined by:

$$I(X;Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_{X,Y}(x,y) \log \frac{p_{X,Y}(x,y)}{p_X(x)\, p_Y(y)}$$

$$I(X;Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_{X,Y}(x,y) \log \frac{p_{X|Y}(x|y)}{p_X(x)} \frac{\cancel{p_Y(y)}}{\cancel{p_Y(y)}}$$

$$\boxed{I(X;Y) = I(Y;X)}$$

$$\boxed{I(X;Y) \geq 0}$$

$\Rightarrow I(X;Y) = 0 \Leftrightarrow X \perp Y$ (if independent)

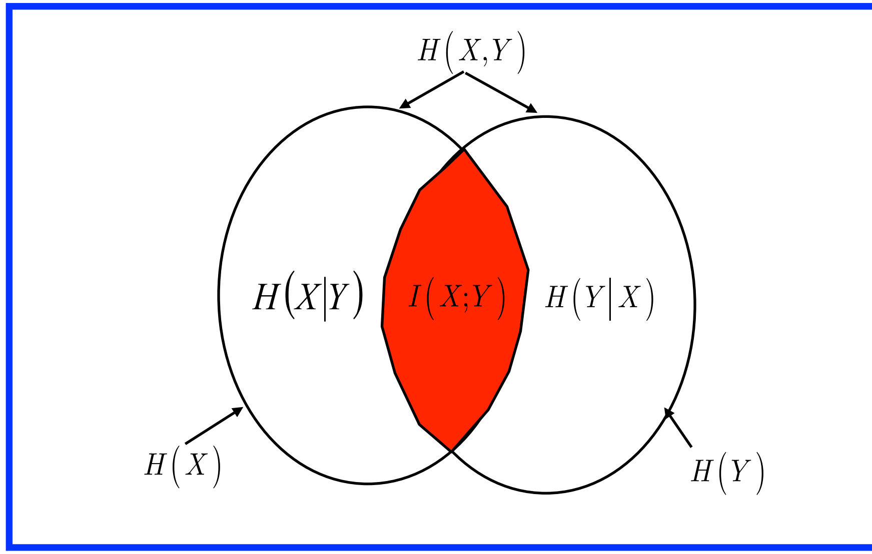
Mutual Information

Relationship between mutual information and entropy

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

$$I(X;X) = H(X)$$

$$I(X;X) = H(X) - \underbrace{H(X|X)}_{0}$$



Relationship between mutual information and KLD

$$I(X;Y) = D\big(p(x,y)\|p(x)p(y)\big) = E_{p(x,y)}\left[\log \frac{p_{X,Y}(x,y)}{p_X(x)\,p_Y(y)}\right]$$

Mutual Information

Conditional mutual information

$$\boxed{I(X;Y|Z) = H(X|Z) - H(X|Y,Z)}$$

Proof:

$$I(X;Y|Z) = E_{p(x,y,z)}\left[\log \frac{p_{X,Y|Z}(x,y|z)}{p_{X|Z}(x|z)\,p_{Y|Z}(y|z)}\right] = E_{p(x,y,z)}\left[\log \frac{p_{X|Y,Z}(x|y,z)}{p_{X|Z}(x|z)}\right]$$

$$p_{X,Y|Z}(x,y|z) = p_{Y|Z}(y|z)\,p_{X|Y,Z}(x|y,z)$$

$$= \underbrace{E_{p(x,y,z)}\left[\log p_{X|Y,Z}(x|y,z)\right]}_{-H(X|Y,Z)} - \underbrace{E_{p(x,y,z)}\left[\log p_{X|Z}(x|z)\right]}_{H(X|Z)}$$

Mutual Information

Chain rule for mutual information

$$\boxed{I\left(X_1, X_2, \ldots, X_n; Y\right) = \sum_{i=1}^{n} I(X_i; Y \mid X_{i-1}, \ldots, X_1)}$$

Proof:

$$I\left(X_1, X_2, \ldots, X_n; Y\right) = H\left(X_1, X_2, \ldots, X_n\right) - H\left(X_1, X_2, \ldots, X_n \mid Y\right)$$

$$\left\{ I\left(Z; Y\right) = H\left(Z\right) - H\left(Z \mid Y\right) \right\}$$

Chaine rule for entropy

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i \mid X_{i-1}, \ldots, X_1)$$

$$= \sum_{i=1}^{n} H(X_i \mid X_{i-1}, \ldots, X_1) - \sum_{i=1}^{n} H(X_i \mid X_{i-1}, \ldots, X_1, Y) =$$

$$= \sum_{i=1}^{n} I(X_i; Y \mid X_{i-1}, \ldots, X_1) \quad \square$$

Mutual Information

Principe on non-creation of information by processing

If $X \rightarrow Y \rightarrow Z$ form a Markov chain

$$\boxed{I(X;Y) \geq I(X;Z) \text{ and } I(Y;Z) \geq I(X;Z)}$$

$$p_{X,Y,Z}(x,y,z) = p_X(x)\, p_{Y|X}(y|x)\, p_{Z|Y}(z|y)$$

X,Y,Z form a Markov chain

If $Z = g(Y)$, one has $I(X;Y) \geq I(X;g(Y))$

Definition (« Data processing lemma»):

the amount of infromation can not be increased by any processing!

Mutual Information

Proof

Using the chain rule for the mutual information:

$$I(X;Y,Z) = I(X;Z) + I(X;Y|Z) \quad (a)$$
$$= I(X;Y) + I(X;Z|Y) \quad (b)$$

$I(X;Z|Y) = 0$ (based on Markovianity)

Thus:

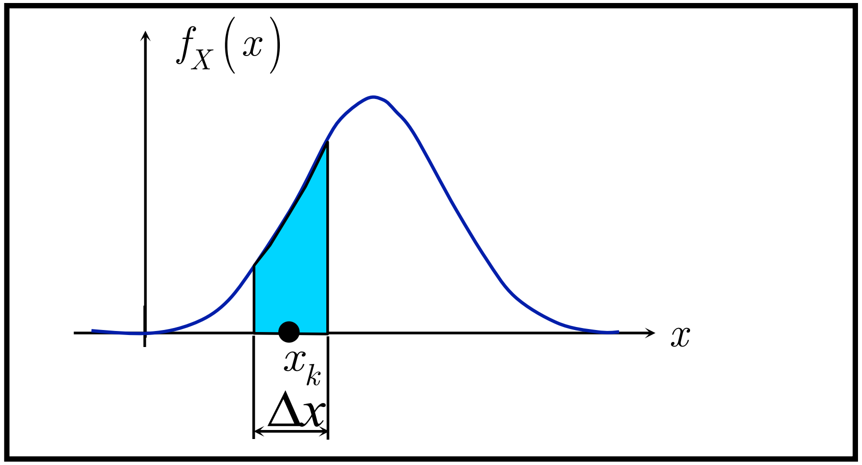
$$I(X;Y) = I(X;Z) + I(X;Y|Z) \Rightarrow \boxed{I(X;Y) \geq I(X;Z)}$$

Continuous R.Vs.: Differential entropy

Definition (entropy): *Differential entropy* of r.v. X with pdf $f_X(x)$

$$h(X) = -\int_{\mathcal{X}} f_X(x)\log_2 f_X(x)\,dx = E_{f_X}\left[-\log_2 f_X(x)\right]$$

$$P[x_k;\Delta x] = \int_{\Delta x} f_X(x)dx \cong f_X(x_k)\Delta x$$



Continuous R.Vs.: Differential entropy

$$H\left(X;\Delta x\right) = -\sum_{i=1}^{N} P\left[x_k, \Delta x\right] \log_2 P\left[x_k, \Delta x\right] = -\sum_{i=1}^{N} f_X\left(x_k\right)\Delta x \log_2 f_X\left(x_k\right)\Delta x$$

$$= -\sum_{i=1}^{N} f_X\left(x_k\right)\Delta x \log_2 f_X\left(x_k\right) - \underbrace{\sum_{i=1}^{N} f_X\left(x_k\right)\Delta x}_{=1} \underbrace{\log_2 \Delta x}_{const}$$

$$\int_{-\infty}^{+\infty} f_X\left(x\right) dx = 1$$

In the limit Δx tends to zero for large N. As a result, $\log_2 \Delta x$ tends to infinity.

$$H\left(X;\Delta x\right) = \underbrace{-\int f_X\left(x\right) \log_2 f_X\left(x\right) dx}_{h\left(X\right)} - \log_2 \Delta x.$$

Continuous R.Vs.: Differential entropy

Definition (Differential entropy of Gaussian r.v.): *differential entropy of Gaussian r.v.* $X \sim \mathcal{N}\left(0, \sigma_X^2\right)$ is:

$$\boxed{h\left(X\right) = -E_{f_X}\left[\log_2 f_X\left(x\right)\right] = \frac{1}{2}\log_2\left(2\pi e \sigma_X^2\right).}$$

Proof:

$$h\left(X\right) = -\int_{\mathcal{X}} f_X\left(x\right)\ln f_X\left(x\right)dx \quad [nants] = -\int_{-\infty}^{\infty} f_X\left(x\right)\left[-\frac{x^2}{2\sigma_X^2} - \ln\sqrt{2\pi\sigma_X^2}\right]dx =$$

$$= \int_{-\infty}^{\infty} f_X\left(x\right)\frac{x^2}{2\sigma_X^2}dx + \ln\sqrt{2\pi\sigma_X^2}\int_{-\infty}^{\infty} f_X\left(x\right)dx$$

Continuous R.Vs.: Differential entropy

$$h(X) = \int_{-\infty}^{\infty} f_X(x)\frac{x^2}{2\sigma_X^2}\,dx + \ln\sqrt{2\pi\sigma_X^2}\underbrace{\int_{-\infty}^{\infty} f_X(x)\,dx}_{=1}$$

$$Var[X] = \int_{-\infty}^{\infty} x^2 f_X(x)\,dx = \sigma_X^2$$

$$\int_{-\infty}^{\infty} f_X(x)\frac{x^2}{2\sigma_X^2}\,dx = \frac{\sigma_X^2}{2\sigma_X^2} = \frac{1}{2}$$

$$h(X) = \frac{1}{2} + \ln\sqrt{2\pi\sigma_X^2} = \frac{1}{2}\ln e + \frac{1}{2}\ln 2\pi\sigma_X^2 = \frac{1}{2}\left[\ln e + \ln 2\pi\sigma_X^2\right] = \frac{1}{2}\ln 2\pi e\sigma_X^2\ [nants]$$

$$h(X) = \frac{1}{2}\log_2 2\pi e\sigma_X^2\ [bits]$$

Differential entropy: properties

1. Translation does not change the entropy:

$$\boxed{h\left(X+a\right)=h\left(X\right).}$$

2. Impact of scaling on the differential entropy, if X is scalar r.v.:

$$\boxed{h\left(Xa\right)=h\left(X\right)+\log\left|a\right|,}$$

and if X is a random:

determinant of $\mathbf{A}$

$$\boxed{h\left(\mathbf{AX}\right)=h\left(\mathbf{X}\right)+\log\left|\det(\mathbf{A})\right|,}$$