Teoretične osnove računalništva 2
Teorija informacij

Entropy of continuous random variables

Enes Pasalic UP FAMNIT študijsko leto 20/21

Lecture overview

- Entropy of continuous random variables
  - Definition of entropy
  - AEP for contnuous random variables
  - Connection with the entropy of discrete random variables
  - Entropy of joint and conditional contin. random variables
  - Properties for entropy of contin. random variables

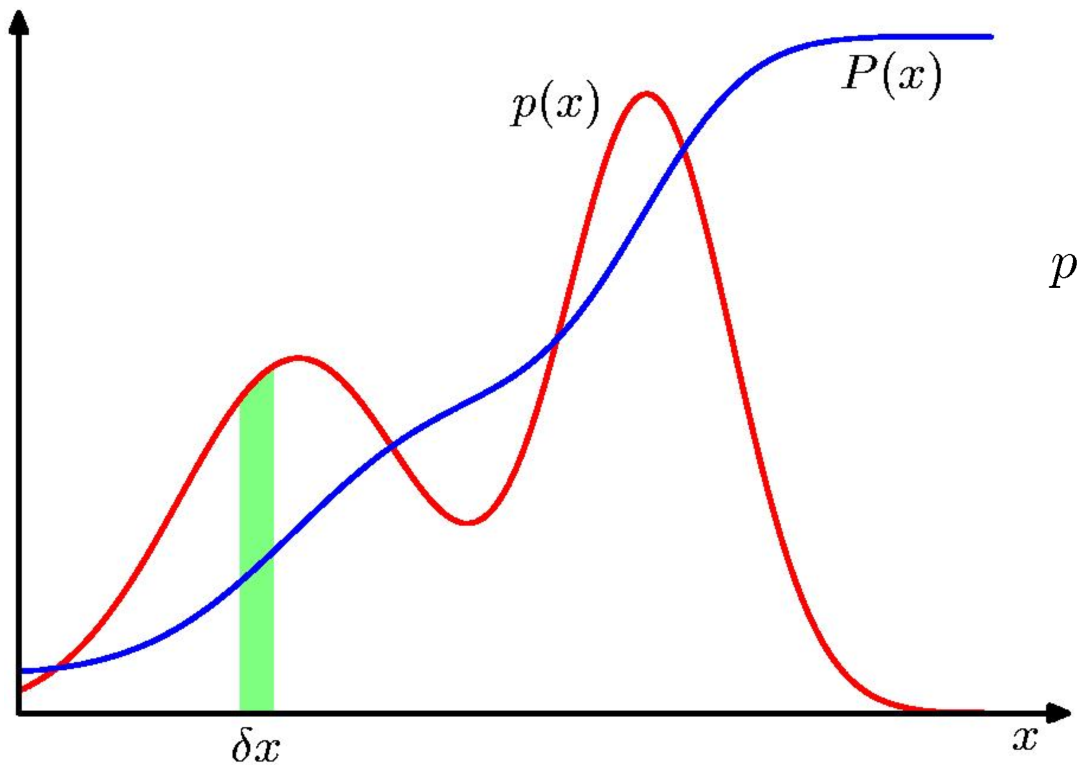
Probability mass distribution

Probability mass distribution

Let X be a random variable with the **cumulative probability function** $F(x) = Pr(X < x)$. If $F(x)$ is continuous then X is called continuous RV. With $f(x)$ we denote derivative of F, i.e. $f(x) = F'(x)$.

If we have $\int_{-\infty}^{+\infty} f(x)dx = 1$, then $f(x)$ is called probability mass distribution of X.

Probability mass distribution - example



$$p(x \in (a,b)) = \int_a^b p(x)\,\mathrm{d}x$$

$$P(z) = \int_{-\infty}^{z} p(x)\,\mathrm{d}x$$

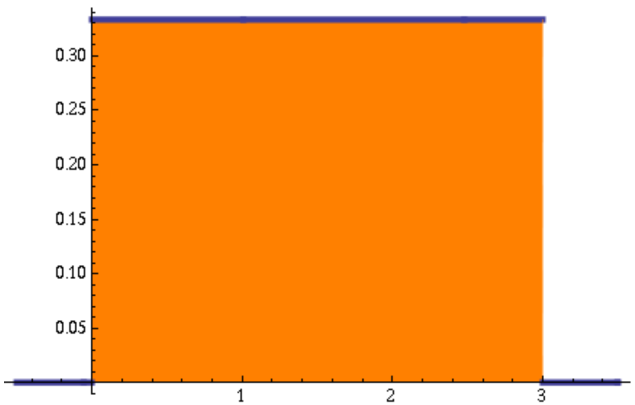
$$p(x) \geqslant 0 \qquad \int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1$$

Probability mass distribution – example II

- Uniform distribution:

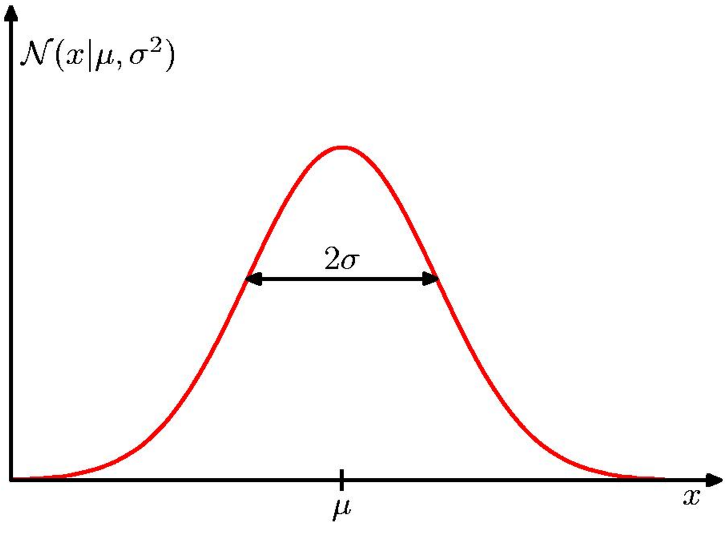
$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b, \\ 0 & \text{drugje} \end{cases}$$

f(x) for *a=0, b=3*



Normal (Gauss) distribution

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$



Diferential entropy

Entropija zvezne naključne spremenljivke

Entropija zvezne naključne spremenljivke X s porazdelitvijo gostote verjetnosti $f(x)$ je definirana kot:

$$h(X) = -\int_S f(x) \log f(x) dx,$$

kjer je S nosilec naključne spremenljivke X.

- Entropy for continuous RVs has **similar properties** as the entropy of discrete random variables, though **there are differences**.

Differential entropy of uniform distribution

- Probability mass distribution

$$f(x) = \begin{cases} \frac{1}{a} & 0 \leq x \leq a, \\ 0 & \text{drugje} \end{cases}$$

- Entropy:

$$h(X) = -\int_0^a \frac{1}{a} \log \frac{1}{a} \, dx = \log a$$

Differential entropy of normal distribution

- Pprobability mass distribution *X ~ N(0, σ)*:

$$X \sim \phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-x^2}{2\sigma^2}}$$

- Entropy:

$$\begin{aligned} h(\phi) &= -\int \phi \ln \phi \\ &= -\int \phi(x)\left[-\frac{x^2}{2\sigma^2} - \ln\sqrt{2\pi\sigma^2}\right] \\ &= \frac{EX^2}{2\sigma^2} + \frac{1}{2}\ln 2\pi\sigma^2 \\ &= \frac{1}{2} + \frac{1}{2}\ln 2\pi\sigma^2 \\ &= \frac{1}{2}\ln e + \frac{1}{2}\ln 2\pi\sigma^2 \\ &= \frac{1}{2}\ln 2\pi e\sigma^2 \quad \text{nats.} \end{aligned}$$

$$h(\phi) = \frac{1}{2}\log 2\pi e\sigma^2 \quad \text{bits.}$$

Joint probability mass distribution (optional)

Porazdelitev gostote vezane verjetnosti

Denimo, da imamo zaporedje zveznih naključnih spremenljivk $X_1, X_2, \ldots, X_n$, potem lahko definiramo porazdelitev gostote vezane verjetnosti $f(x_1, \ldots, x_n)$, tako da velja

$$Pr(X_1, \ldots, X_n \in D) = \int_D f(x_1, \ldots, x_n) dx_1 \ldots dx_n.$$

Oziroma, če je $F(x_1, \ldots, x_n) = Pr(X_1 \leq x_1, \ldots, X_n \leq x_n)$ kumulativna funkcija vezane verjetnosti, potem je porazdelitev gostote vezane verjetnosti definirana s parcialnim odvodom kumulativne funkcije

$$f(\mathbf{x}) = \left. \frac{\partial^n F}{\partial x_1 \ldots \partial x_n} \right|_{\mathbf{x}} .$$

Joint probability mass distribution : independence

English version: If $X_1, X_2, \ldots, X_n$ are mutually independent and have the same distribution $f(x)$ then

Če so $X_1, X_2, \ldots, X_n$ med seboj neodvisne in enako porazdeljene s porazdelitvijo $f(x)$, potem je

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i).$$

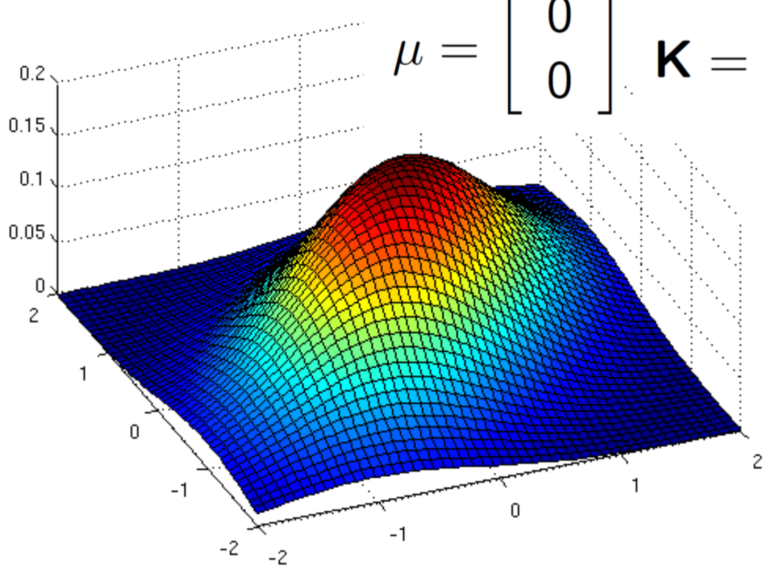
Multidimensional Gauss distribution (optional)

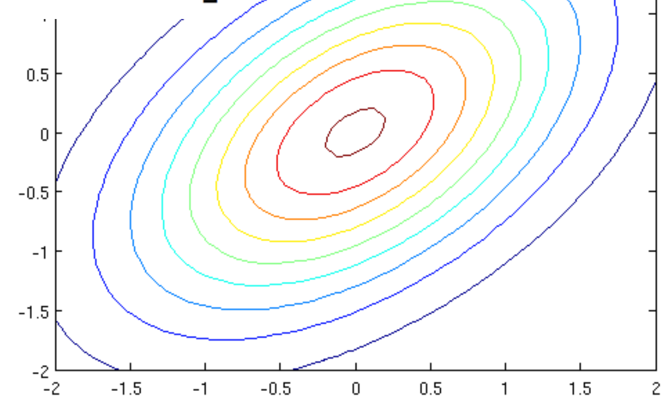
Gaussova porazdelitev naključnih spremenljivk $X_1, X_2, \ldots, X_n$ je

$$f(\mathbf{x}) = \frac{1}{(\sqrt{2\pi})^n |\mathbf{K}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \mathbf{K}^{-1}(\mathbf{x}-\mu)},$$

kjer je μ povprečni vektor in $\mathbf{K}$ kovariančna matrika,
$\mathbf{K} = \mathbb{E}[(\mathbf{x}-\mu)(\mathbf{x}-\mu)^T]$.

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \mathbf{K} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$





Joint entropy for continuous random variables

Joint entropy

Joint entropy of continuous RVs $X_1, \ldots, X_n$ with probability mass distribution $f(X_1, \ldots, X_n)$ is defined as:

$$h(X_1, \ldots, X_n) = -\int f(x) \log f(x) dx.$$

Conditional entropy

Conditional entropy

For continuous RVs X, Y with $f(X, Y)$ one can define conditional entropy $h(X|Y)$ as:

$$h(X|Y) = -\int f(x,y) \log f(x|y) dx.$$

Remark: Since $f(x|y) = \frac{f(x,y)}{f(y)}$ we get

$$h(X|Y) = h(X|Y) - H(Y).$$

Entropy of multidimensional Gauss distribution (opt.)

Entropija večdimenzionalne Gaussove porazdelitve

Entropija večdimenzionalne Gaussove porazdelitve naključnih spremenljivk $X_1, X_2, \ldots, X_n$ s povprečjem μ in kovariančno matriko $\mathbf{K}$ je

$$h(X_1, X_2, \ldots, X_n) = h(N(\mathbf{x}; \mu, \mathbf{K})) = \frac{1}{2} \log(2\pi e)^n |K| \text{ bitov},$$

kjer je $|\mathbf{K}|$ determinanta kovariančne matrike $\mathbf{K}$.

Proof

- Proof:
  - pmf $$f(\mathbf{x}) = \frac{1}{\left(\sqrt{2\pi}\right)^n |K|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T K^{-1}(\mathbf{x}-\mu)}$$
  - entropy $$h(f) = -\int f(\mathbf{x})\left[-\frac{1}{2}(\mathbf{x}-\mu)^T K^{-1}(\mathbf{x}-\mu) - \ln\left(\sqrt{2\pi}\right)^n |K|^{\frac{1}{2}}\right] d\mathbf{x}$$

$$= \frac{1}{2}E\left[\sum_{i,j}(X_i-\mu_i)\left(K^{-1}\right)_{ij}(X_j-\mu_j)\right] + \frac{1}{2}\ln(2\pi)^n|K|$$

$$= \frac{1}{2}E\left[\sum_{i,j}(X_i-\mu_i)(X_j-\mu_j)\left(K^{-1}\right)_{ij}\right] + \frac{1}{2}\ln(2\pi)^n|K|$$

$$= \frac{1}{2}\sum_{i,j}E[(X_j-\mu_j)(X_i-\mu_i)]\left(K^{-1}\right)_{ij} + \frac{1}{2}\ln(2\pi)^n|K|$$

$$= \frac{1}{2}\sum_j\sum_i K_{ji}\left(K^{-1}\right)_{ij} + \frac{1}{2}\ln(2\pi)^n|K|$$

$$= \frac{1}{2}\sum_j (KK^{-1})_{jj} + \frac{1}{2}\ln(2\pi)^n|K|$$

$$= \frac{1}{2}\sum_j I_{jj} + \frac{1}{2}\ln(2\pi)^n|K|$$

$$= \frac{n}{2} + \frac{1}{2}\ln(2\pi)^n|K|$$

$$= \frac{1}{2}\ln(2\pi e)^n|K| \qquad \text{nats}$$

$$= \frac{1}{2}\log(2\pi e)^n|K| \qquad \text{bits.} \qquad \square$$

AEP version for continuous random variables

For discrete IID RVs, typical sequences stand for the most of probability (there are few of these), we had $p(X_1, \dots, X_n) = 2^{-nH(X)}$.
For continuous RVs we have:

AEP statement

For IID RVs $X_1, \dots, X_n$ with probability mass function $f(x)$ we have:

$$-\frac{1}{n} \log f(X_1, \dots, X_n) \to E(-\log f(x)) = h(X).$$

Definition of typical „volumes“

Definicija tipične množice

Za poljuben $\epsilon > 0$ in poljuben n, definiramo *tipično množico* $A_\epsilon^{(n)}$ glede na funkcijo gostote $f(X)$ na naslednji način

$$A_\epsilon^{(n)} = \left\{ (x_1, x_2, \ldots, x_n) \in S^n : \left| -\frac{1}{n} \log f(x_1, x_2, \ldots, x_n) - h(X) \right| \leq \epsilon \right\},$$

kjer velja $f(x_1, x_2, \ldots, x_n) = \prod_{i=1}^n f(x_i)$.

Volumen množice

Volumen množice $A \subset \mathbb{R}^n$ je definiran kot

$$\mathrm{Vol}(A) = \int_A dx_1 dx_2 \ldots dx_n.$$

AEP statemment

Tipična množica $A_\epsilon^{(n)}$ ima naslednje lastnosti za dovolj velike n:

1. $Pr(A_\epsilon^{(n)}) > 1 - \epsilon$.
2. $(1-\epsilon)2^{n(h(X)-\epsilon)} \leq \mathrm{Vol}(A_\epsilon^{(n)}) \leq 2^{n(h(X)+\epsilon)}$.

Izrek AEP

Tipična množica $A_\epsilon^{(n)}$ ima najmanjši volumen med množicami z verjetnostjo $\geq 1 - \epsilon$.

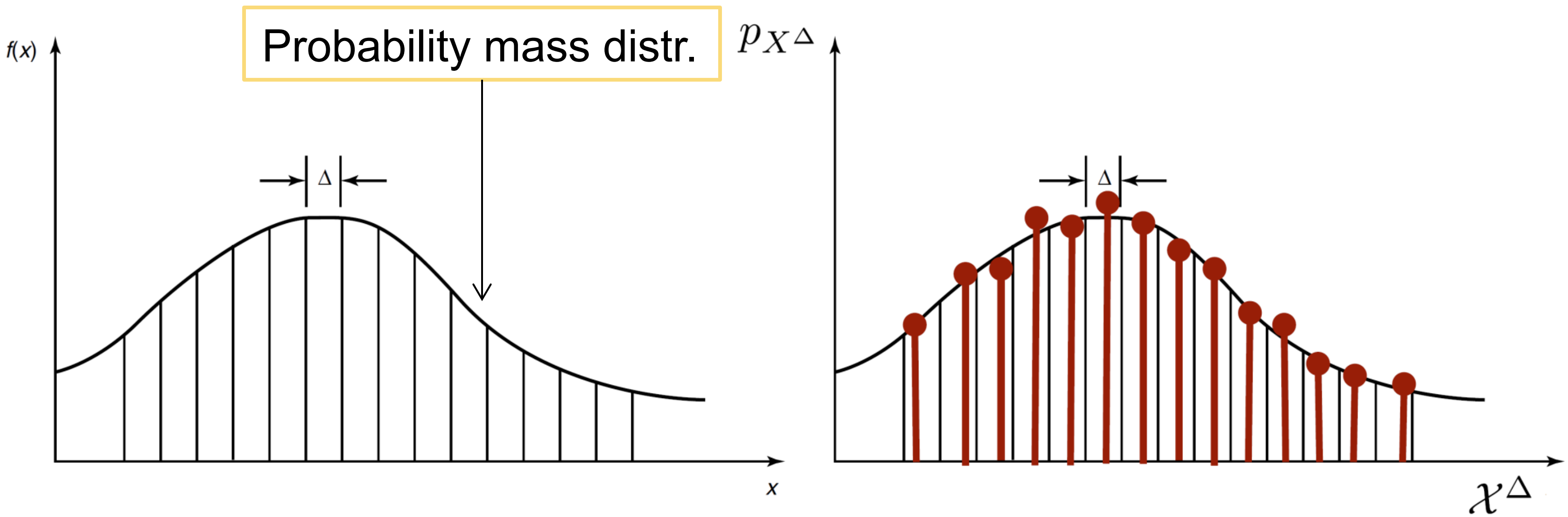
AEP statement

- AEP statement claims, that the minimal volume of typical subsets, that stands for whole probability, is approximately 2^{nh}.
  Since this is n-dimensional space, this means, that one side of this „volume“ is of length $(2^{nh})^{(1/n)} = 2^{h}$.

- Diferential entropy can be interpreted as:
  It is logarithm of volume side of the least (hyper)subspace, which has almsot all probability.

- Consequences:
  - Small entropy means, that the values of RV concentrated on a small subspace.
  - Large entropy means, that the values of RV are dispersed widely in the space.

Quantization of continuous random variables



- There exists x_i in any interval Δ, such that:

$$f(x_i)\Delta = \int_{i\Delta}^{(i+1)\Delta} f(x)\,dx$$

Connection between entropy of contin. and discrete RV

- Quantized RV: $X^{\Delta} = x_i \qquad \text{if } i\Delta \le X < (i+1)\Delta$

- Probability that $X^{\Delta} = x_i$

$$p_i = \int_{i\Delta}^{(i+1)\Delta} f(x)\,dx = f(x_i)\Delta$$

- Entropy of quantizied RV

$$\begin{aligned} H(X^{\Delta}) &= -\sum_{-\infty}^{\infty} p_i \log p_i \\ &= -\sum_{-\infty}^{\infty} f(x_i)\Delta \log(f(x_i)\Delta) \\ &= -\sum \Delta f(x_i) \log f(x_i) - \sum f(x_i)\Delta \log \Delta \\ &= -\sum \Delta f(x_i) \log f(x_i) - \log \Delta, \end{aligned}$$

Connection between entropy of contin. and discrete RV

Connection

If $f(x)$ probability mass distribution of X which can be integrated then

$$H(X^{\Delta}) + \log \Delta \to h(f) = h(X), \text{ when } \Delta \to 0.$$

This means that the entropy of quantized RV $H(X^{\Delta})$ using n bits is $h(X) + n$, notice $\Delta = 2^{-n}$.

Quantization examples

Examples

1. If X is uniformly distributed on the interval $[0, 1]$ and $\Delta = 2^{-n}$, then $h = 0$ and $H(x^{\Delta}) = n$. Thus, we need n bits to describe X with n bit precision.
2. Let $X \approx \mathcal{N}(0, \sigma^2)$ with $\sigma^2 = 100$. Then, we need $n + \frac{1}{2}\log 2\pi e\sigma$ bits to describe X with n bit precision.
3. In general, to describe X with n-bit precision we need $h(X) + n$ bits.

Relative entropy (optional)

Relativna entropija

Relativna entropija (ali Kullback-Leiblerjeva razdalja) $D(f \parallel g)$ med funkcijama gostot f in g je definirana kot

$$D(f \parallel g) = \int f \log \frac{f}{g}.$$

$D(f \parallel g)$ je končna, če je nosilec f podmnožica nosilca g. (Velja $0 \log \frac{0}{0} = 0$.)

Mutual information

Medsebojna informacija

Medsebojna informacija $I(X;Y)$ med dvema naključnima spremenljivkama s funkcijo gostote vezne verjetnosti $f(x,y)$ je definirana z

$$I(X;Y) = \int f(x,y) \log \frac{f(x,y)}{f(x)f(y)} dxdy.$$

Iz definicije sledi:

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

ali

$$I(X;Y) = D(f(x,y) \parallel f(x)f(y)).$$

Connection with discrete RVs

$$
\begin{aligned}
I(X^{\Delta}; Y^{\Delta}) &= H(X^{\Delta}) - H(X^{\Delta}|Y^{\Delta}) \\
&\approx h(X) - \log \Delta - (h(X|Y) - \log \Delta) \\
&= I(X; Y).
\end{aligned}
$$

Example (optional)

Medsebojna informacija med koreliranimi Gaussovimi naklj. spr.

Denimo, da imamo zv. naklj. spr. X in Y, ki sta normalno porazdeljeni s korelacijo ρ. Torej $(X, Y) \sim \mathcal{N}(0, K)$, kjer je

$$K = \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix}.$$

Potem je $h(X) = h(Y) = \frac{1}{2}\log(2\pi e)\sigma^2$ in
$h(X, Y) = \frac{1}{2}\log(2\pi e)^2|K| = \frac{1}{2}\log(2\pi e)^2\sigma^4(1-\rho^2)$. Sledi:

$$I(X; Y) = h(X) + h(Y) - h(X, Y) = -\frac{1}{2}\log(1-\rho^2).$$

Če je $\rho = 0$, potem sta X in Y neodvisni in je medsebojna informacija enaka 0.
Če je $\rho = \pm 1$, potem sta X in Y popolnoma korelirani in je medsebojna informacija ∞.

Properties

Izrek

$$D(f \parallel g) \geq 0.$$

Enakost velja natanko tedaj, ko je $f = g$ skoraj povsod.

Posledici

1. $I(X;Y) \geq 0$. Enakost velja natanko tedaj, ko sta X in Y neodvisni.
2. $h(X|Y) \leq h(X)$. Enakost velja natanko tedaj, ko sta X in Y neodvisni.

Properties II

Verižno pravilo pri entropiji

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

Posledica

$$h(X_1, X_2, \ldots, X_n) \leq \sum_{i=1}^{n} h(X_i).$$

Enakost velja natanko tedaj, ko so $X_1, \ldots X_n$ neodvisne.

Linear transformations

Linearne transformacije zveznih naključnih spr. in entropija

1. $h(X + c) = h(X)$. Premik naključne spremenljivke ne spremeni njeno entropijo.
2. $h(aX) = h(X) + \log|a|$.
3. $h(A\mathbf{X}) = h(\mathbf{X}) + \log|\det(A)|$.

Maksimal entropy

- **Entropy of multidimensional Gauss distribution is largest among ALL entropies of joint RVs having equal covariance metrix.**

Entropija naključnih spremenljivk z enako kovarianco

Naj bo $\mathbf{X} \in \mathbb{R}^n$ vektor naključnih spremenljivk s povprečjem $\mathbf{0}$ in kovarianco $K = E\mathbf{X}\mathbf{X}^T$ $(K_{ij} = EX_iX_j, 1 \leq i,j \leq n)$. Potem velja $h(\mathbf{X}) \leq \frac{1}{2}\log(2\pi e)^n|K|$, enakost velja natanko tedaj, ko je $\mathbf{X} \sim N(\mathbf{0}, K)$.